

Domain-Specific Visual Representation Learning Using Natural Language Supervision

Master's Thesis

Master Artificial Intelligence – Utrecht University

13-01-2022



**Utrecht
University**

Author: Alexander S.Y. Kern – 5711088 – a.s.y.kern@students.uu.nl

Words: 13472

Daily Supervisor: Dr. Raimon Pruim

First Supervisor: Dr. Ruud Hortensius

Second Reader: Dr. Roy Hessels

Abstract

The military intelligence domain is one of many fields investigating deep learning methods to automate various processes, especially for the task of recognizing specific entities in large sets of images. Current state-of-the-art methods cannot be easily applied in the military domain since they require large sets of labelled images, which are challenging to acquire for the domain-specific classes. Recently, research has investigated the possibility of learning visual features with natural language supervision by using image captioning as a pre-training task for visual backbones. This study investigates the possibility of pre-training with domain-specific image-captions to learn domain-specific visual features. We pre-train convolutional neural networks from scratch, using a military-specific image-caption dataset (Janes Captions) collected for this study. The effect of different image captioning pre-training tasks on the learning of the visual features was also evaluated. Although these models did not outperform the current state-of-the-art methods, they outperformed models pre-trained on similar amounts of generic image-captions. Ultimately, natural language supervision for pre-training visual models is a promising concept that, if applied correctly, could solve the problems of current state-of-the-art methods, especially for application in specific domains.

Acknowledgements

The thesis before you is the final product of my Master's in Artificial Intelligence. Nine months of intensive working from home sessions, in the midst of a pandemic, have resulted in this behemoth. This project would not have been possible without the Intelligent Imaging department at TNO. I especially want to thank Dr. Raimon Pruim for his never-ending stream of feedback and guidance throughout the entire project. I also want to thank Utrecht University and Dr. Ruud Hortensius for helping me to pursue my interests during my Master's.

I want to thank Paul, for never shutting down another explanation of a random AI topic, which I barely even comprehended myself most of the time. I want to thank my parents for letting me (almost) endlessly discover all my different interests during my time as a student while always keeping an eye on me. And finally, thank you Romy, for always being there for me. You were compassionate when I was being too hard for myself and firm when I was being too lenient. Without you, this project would have been too complex, probably months overdue, and turned me into a night owl.

Table of Contents

Abstract.....	i
Acknowledgements	ii
1 Introduction	1
1.1 Automation of vision tasks in the military intelligence domain.....	1
1.2 Pre-training for vision tasks	1
1.3 Pre-training for vision-language tasks.....	4
1.4 Natural language supervision for visual representation learning	7
1.5 Current work	11
2 Methodology.....	13
2.1 Experiments	13
2.2 Pre-training datasets	14
2.3 Pre-training tasks.....	14
2.4 Architecture.....	15
2.5 Pre-training & Evaluation	17
2.6 Benchmarks.....	18
2.7 Qualitative analysis	19
2.8 Additional analysis.....	20
3 Results	22
3.1 Experiment 1	22
3.2 Experiment 2	23
3.3 Benchmarks.....	23
3.4 Qualitative results	24
3.5 Additional analysis.....	27
4 Discussion.....	31
4.1 Vision-language pre-training datasets and tasks	31
4.2 Visual modality in multimodal transformers.....	33
4.3 Conclusion	35
References	36

1 Introduction

1.1 Automation of vision tasks in the military intelligence domain

The military intelligence domain is one of the many fields investigating deep learning methods to automate various (intel) processes. Analysts in the intelligence domain must combine information from many different data sources into a single intel product. The analysed data may consist of various modalities, varying from textual to auditive and visual information. An example of a task in this domain that is viable for automation through deep learning methods is classifying specific entities or objects in large sets of images and videos.

With the invention of Convolutional Neural Networks (CNN), deep learning has recently achieved impressive results regarding computer vision tasks, such as image classification and object detection [39]. However, these methods cannot be easily applied across all domains. Most successful deep learning methods require the computer vision model to be trained on a large set of labelled images specific to the domain the model will be applied in. For example, if one would want a model that can classify cars into the correct brand or type, one would need to have a large set of images from different cars and annotate each image with the correct label. Creating large, annotated datasets is difficult and time-consuming, even in generic domains. These problems get even more severe when applying them in more specific domains, such as military intelligence. For example, a system that can detect certain types of tanks in images requires a large training set of correctly labelled images from tanks. On top of the fact that images of tanks are available only in relatively small numbers, expert knowledge is also required to annotate the images accurately, which increases the collection cost of these datasets. For that reason, much research is focused on developing methods that can achieve similar performances on vision tasks with a smaller amount of (labelled) images. Examples of such methods are self-supervised and multimodal pre-training. Such methods aim first to train a model on a large out-of-domain dataset to learn how to extract meaningful visual features from images (i.e., pre-training). Subsequently, the model is typically transferred to another domain or another downstream vision task [17,29,60], for which now less labelled data is required to further finetune the model on the specific downstream task.

1.2 Pre-training for vision tasks

1.2.1 Supervised pre-training

One of the first pre-training approaches to be explored was supervised pre-training, also known as transfer learning. This method, visualized in figure 1, pre-trains a network on a large-scale labelled dataset and then transfers the pre-trained network to downstream vision tasks, only needing finetuning with a limited set of samples [12]. This approach has been successfully applied across various vision tasks in different domains. Pre-training on ImageNet, which contains 1.28M images labelled with 1,000 different classes, achieved state-of-the-art (SOTA) results on multiple different vision tasks. However, this approach has some limitations. Large sets of accurately annotated images need to be available. Several of such large-scale datasets do exist, such as Visual Genome, Open Images & ImageNet [38,41,61]. Moreover, although these datasets contain more than 10 million different images, the class taxonomy represents a very generic domain. These datasets would benefit from extending their class taxonomy with more specific classes rather than extending the dataset with more and more

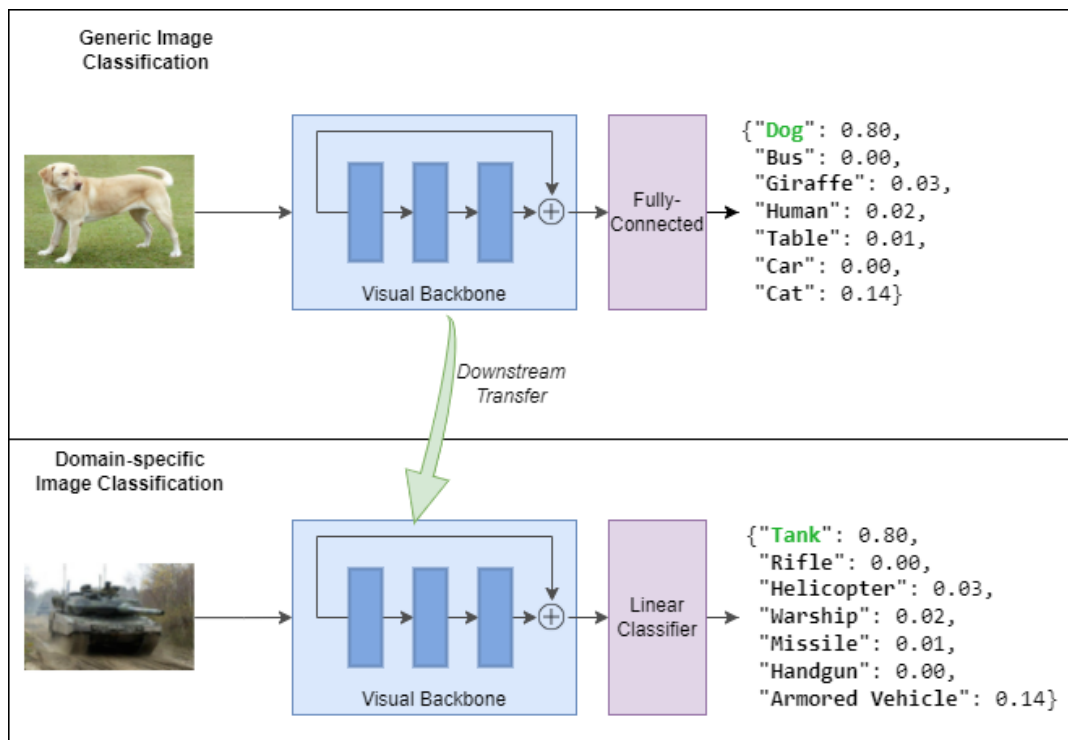


Figure 1: **Supervised Pre-training:** Supervised Pre-training pre-trains a visual backbone on a large scale image classification task, such as ImageNet. After pre-training the visual backbone is then transferred downstream for finetuning on a domain-specific vision tasks that, due to pre-training requires less training samples.

(similar) images. A fine-grained set of classes is, however, difficult to manage. It requires expert knowledge to annotate the collected images correctly. Also, collecting enough images to represent each class equally proves a problem [57,68].

1.2.2 Weakly supervised pre-training

One of the solutions investigated in the literature is to use datasets that have less accurate annotations for pre-training. These methods implement supervised pre-training on large sets of images collected from the internet known to have "noisy" labels. For example, studies investigated pre-training on the JFT300M dataset, an internal Google dataset of 300 million images for training image classification [37,64,70]. These images are labelled using an algorithm that uses a complex mixture between raw web signals, connections between webpages and user feedback, which resulted in a set of 300m "noisy" labelled images. Researchers also investigated an even more extensive set of 3.5 billion images from Instagram for pre-training, using the hashtags as "noisy" labels [53,71]. Using large sets of "noisy" labelled images is also described as the *quantity over quality* approach. This approach reflects the recent trend in natural language processing (NLP), where sizeable uncurated corpora are collected from the internet to pre-train models. Although this approach requires less annotation effort, data collection can still be time extensive depending on the acquisition method.

Moreover, since the labels are weaker, more images are required to reach a performance comparable to models trained with strong-labelled datasets, increasing the required training time. Mahajan and colleagues [53] showed that for the weakly-supervised approach to achieve similar performance, the dataset needed to contain ten times

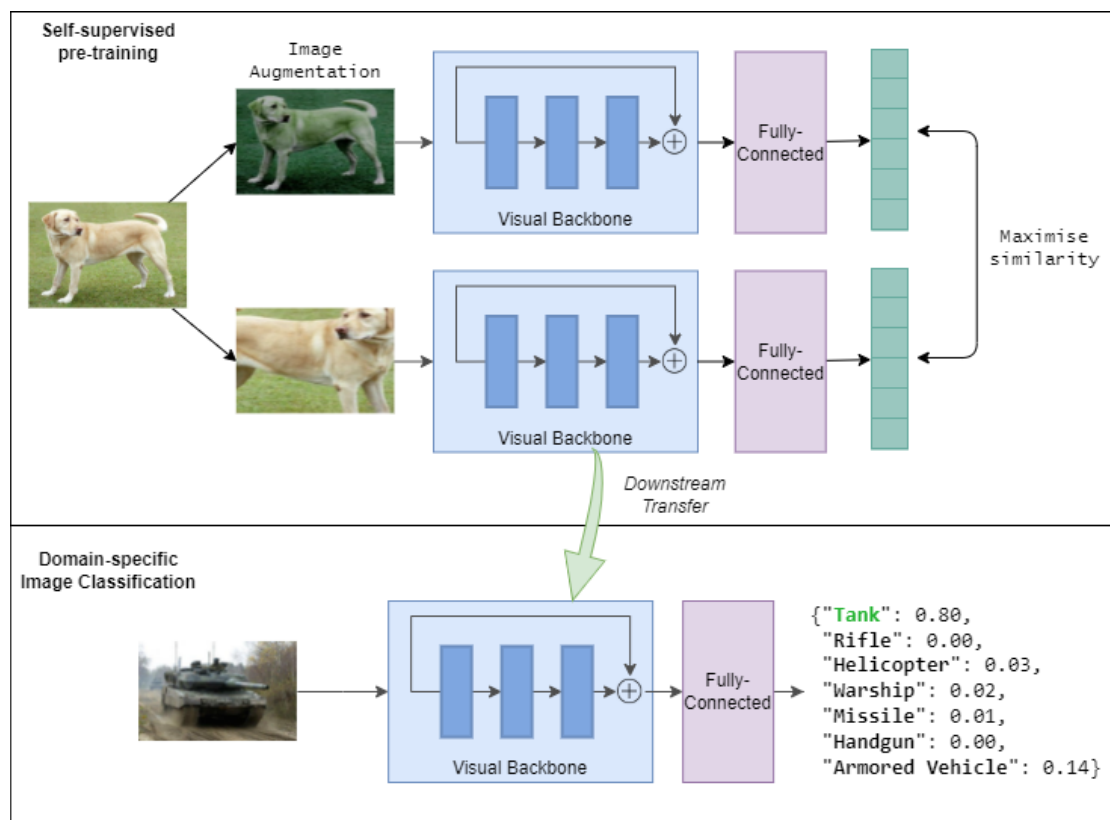


Figure 2: Self-Supervised pre-training: This method trains a network pre-trains a network on "pretext tasks", which are task that do not require annotated samples. The figure visualizes the simCLR self-contrastive learning framework, in which the model is tasked with learning to maximise the similarity between two representations of a single, but twice random augmented, image. The self-supervised pre-training approach allows for training on much more samples, since it doesn't require any annotated dataset. Resulting in increased performance on many difficult vision tasks that require a lot of samples.

the number of images compared to a standard supervised approach. Because of this inefficiency, researchers started investigating alternative approaches such as self-supervised pre-training, which can operate on large sets of unlabeled data in a supervised manner.

1.2.3 Self-supervised pre-training

Self-supervised pre-training is similar to supervised pre-training. It trains a network on sets of images to extract visual features using so-called "pretext tasks" and is then transferred to downstream vision tasks. Pretext tasks concern tasks where unlabelled data samples are automatically labelled using information or patterns within the sample itself. This approach has been shown to exceed supervised pre-training on various vision datasets [7,8]. The main benefit of self-supervised pre-training is that it does not require annotated images. Accordingly, this approach can quickly scale to larger datasets, having already been scaled up to hundreds of millions or even billions of images [28].

Early studies on self-supervised pre-training primarily focused on pretext tasks that involved a label that was artificially computed from the unlabeled images, like context prediction [16], rotation prediction [23] and colourisation [76]. These studies showed that learning a model to extract useful visual features was possible given large networks, enough samples, and long training times. However, these discrete pretext tasks limited the model in learning generalisable visual features.

Therefore, more recent work focuses on contrastive pre-training, resulting in more generalisable models [55,66]. Contrastive pre-training involves encouraging similarity between image features obtained from the same image after random transformation and dissimilarity between image features obtained from different images after random transformation. This contrastive approach has been extended to pretext tasks like context prediction [30,55], mutual information maximization [2,66,72], predicting masked regions [67] and clustering [5,45,79]. SimCLR is one of the methods that implemented a Contrastive Learning framework, obtaining SOTA on various vision tasks [7]. This framework is visualised in figure 2.

Self-supervised approaches have resulted in considerable progress in the field of computer vision. However, to learn useful visual representations with this approach, a much larger network and dataset are required for notable performance. While scaling up the datasets with more unlabeled images would result in even more increased performance, it would also result in the approach becoming even more computationally heavy due to the (necessary) increased size of the network and training time. Due to the inefficient use of the available data, researchers have returned to supervised pre-training on a visual-language task.

1.3 Pre-training for vision-language tasks

As mentioned above, military intelligence is a typical example of a domain in which not only computer vision tasks but also natural language processing (NLP) is relevant. With the rise of so-called "transformers" architectures, these two domains have come closer. Transformers architectures implemented the notion of "self-attention", a construct that takes global dependencies (i.e., the interaction between words in a sentence) into account when processing a sentence. Self-attention helped overcome a common problem in training NLP models related to processing long sentences, thereby heavily improving the performance of language tasks. A prime example is the BERT model, which implements transformers and shows SOTA performance on several natural language tasks [14]

Researchers interested in multimodality reasoned that these architectures could be altered to allow interaction between signals from different modalities, allowing the transformer to combine two single modal representations into one multimodal representation. A "standard" single modal transformer and its component are visualized in figure 3, and a multimodal transformer is visualized in figure 4.

These multimodal transformers have encouraged research concerning so-called "vision-language tasks". Examples of such tasks are image-text-retrieval [19,25], visual question answering [1,26,78], visual grounding [18,78] or image captioning [11,54] (see figure 5). Good performance on these vision-language tasks requires the model to jointly reason over images (vision) and text (language).

With the rise of the internet, and social media platforms like Instagram, Flickr and Wikipedia, large datasets of image-caption pairs have become more accessible. Since the BERT models have shown how large corpus can be used for pre-training on transformer architecture, researchers have investigated a similar approach for pre-training multimodal transformers, with such sets of image-captions pairs [10,33,34,36,46,49,50,52,69]. On a high level, these proposed models are very similar. A pre-trained visual backbone is used to extract the visual features from the image. Then, a pre-trained language model was implemented to extract the textual features from the text. These two features are then combined in a multimodal transformer architecture that decodes the two representations into one multimodal representation. Inspired by BERT [14], these models are pre-trained on image-caption pairs in a self-supervised manner. For example, the ViLBERT model [49] implemented two self-supervised pre-training tasks. The first task concerned masked multimodal learning, where the models were asked to predict the semantic of masked regions in the image or masked tokens in the caption sequence. The other task

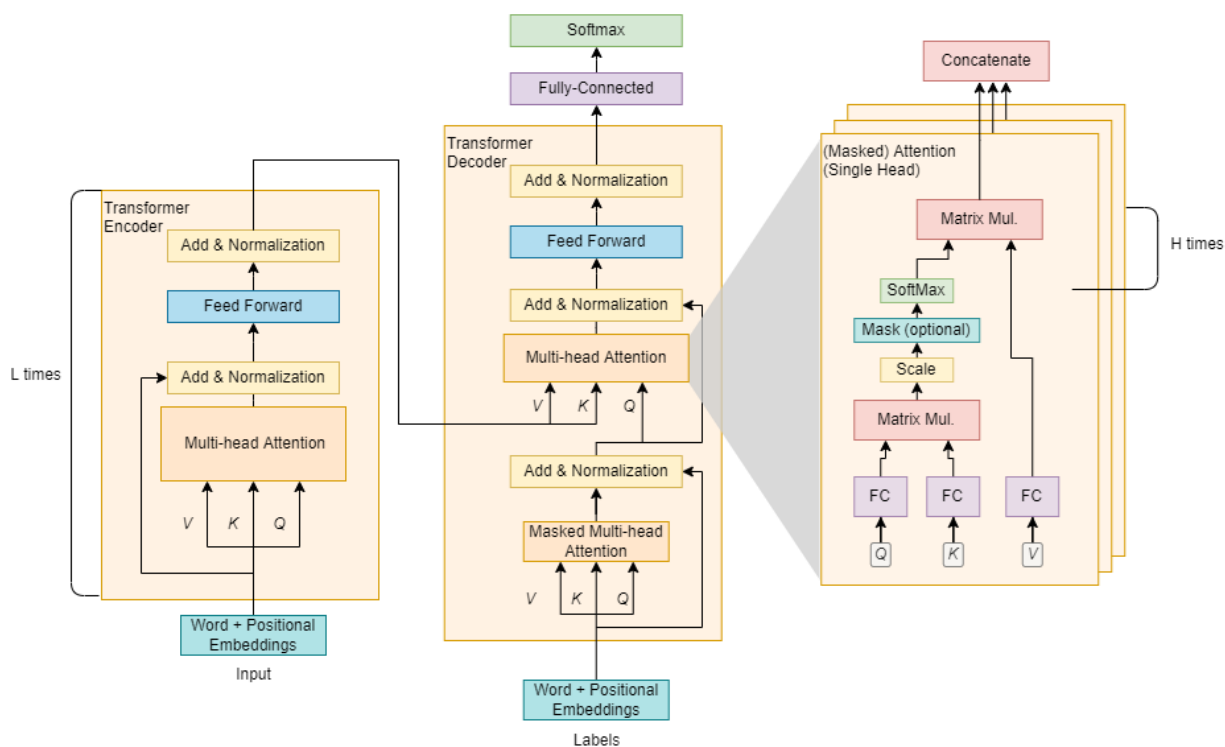


Figure 3: Transformer Encoder and Multi-head Attention component: The transformer architecture consists of two components, the transformer encoder and transformer decoder component, which are used in combination or on their own depending on the task. The main reason for the transformer’s success is the notion of attention, implemented in the multi-head attention component. By taking a combination of word and positional embeddings, attention focusses on one word and learns which other parts of the sentence it needs to *attend* to. It does this by performing matrix multiplications over the Query (Q), Key (K) and Value (V) vectors, and learning the optimal weights for each’s respective fully connected layer. The multi-head attention component, consists of multiple Attention Heads, defined by the hyperparameter H . The multi-head attention module can perform self-attention, in which the vectors originate from the same sequence, or normal attention, in which the vectors V & K originate from a different sequence than the Q vector.

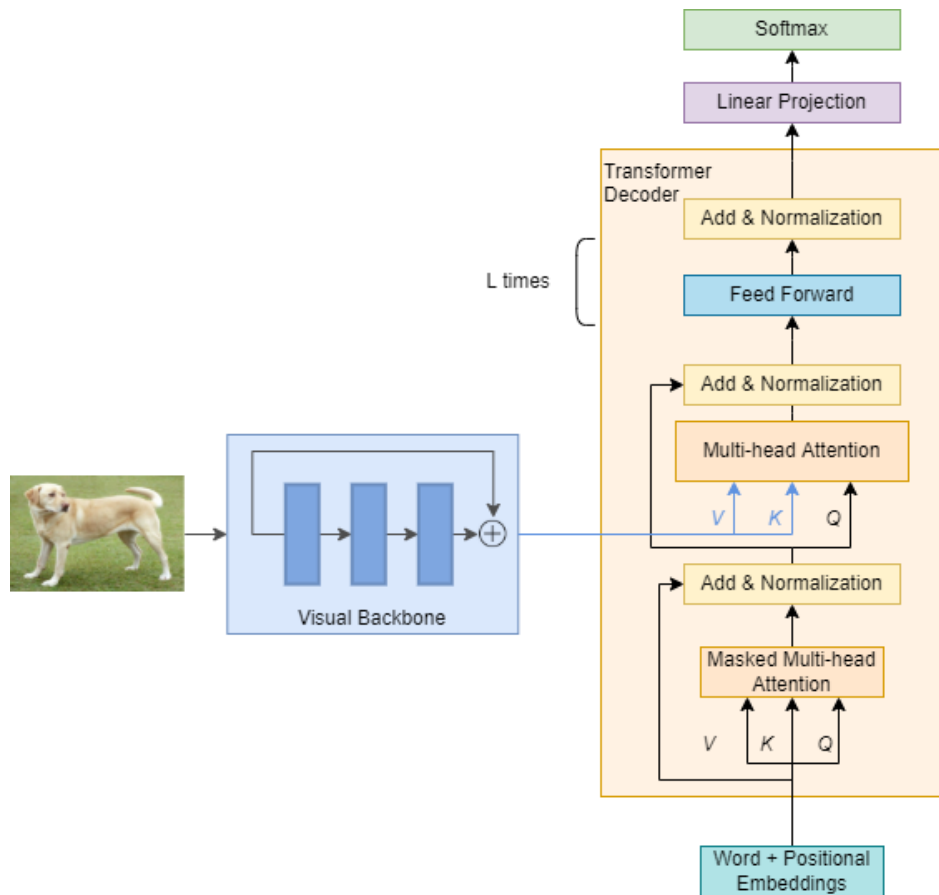


Figure 4: Multimodal Transformer: The multimodal transformer only slightly differs from the more "standard" transformer. The textual and visual features are combined using the multi-head attention component. The value (V) and key (K) vectors represent the visual modality, while the textual features are inputted as the Query (Q) vector. This small adaption allowed for an increased integration of both modalities, resulting in state-of-the-art results on various Vision-Language tasks.

concerned multimodal alignment prediction, where the models needed to predict if the inputted image and text correspond. The pre-trained model is then transferred and finetuned to specific downstream vision-language tasks.

Most of these multimodal architectures for vision-language modelling opted to use a pre-trained visual backbone. The goal of most of these studies was to achieve high performance on various vision-language tasks. Using a pre-trained visual backbone allowed the models to start learning the multimodal features without paying too much attention to whether or not the visual component extracts valuable features. However, when the visual backbone is not frozen (i.e., the parameters will be optimized/adapted during training), the parameters of the backbone will change due to the transformer component propagating a learning signal backwards through the backbone. This raises the question of whether the visual backbone still extracts valuable features relevant to vision tasks such as image classification after pre-training on a multimodal task or whether the visual backbone now only extracts features relevant in a multimodal representation. This notion sparked a new research question: Is it possible to learn visual features from scratch (i.e., without any pre-training) using natural language supervision and a multimodal transformer architecture? If so, this would offer new opportunities for training vision models.

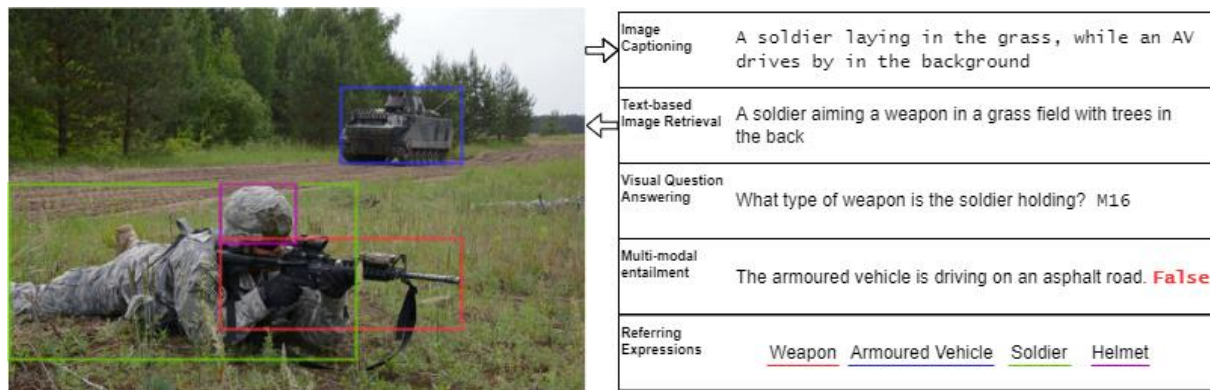


Figure 5: Examples of vision-Language Tasks in the Military Intelligence domain: Vision-Language tasks involve tasks that require joint reasoning over both the vision and language modality. Multiple VL tasks have been designed. Image captioning involves describing an image using natural language. Text-based image retrieval concerns retrieving an image from database using a natural language query. Visual question answering requires to answer a question given an image. Multi-modal entailment involves checking the relationship between a statement in natural language and an image. Referring Expressions regards detecting objects referred to natural language.

1.4 Natural language supervision for visual representation learning

Some studies have investigated learning visual representation using images and their corresponding captions. Before transformer architecture became popular, using natural language as a supervised learning signal for visual representation learning was already investigated. One study investigated auxiliary prediction tasks to help learn a correct visual representation [59]. The auxiliary problem involved predicting whether or not a particular content word is in the associated caption. Two other studies investigated how captions could be used extract labels. The first study labelled images using the associated caption's n-grams and used these labels for training a CNN [43]. The second study extracted topic models from Wikipedia pages and trained a CNN to predict the topic representations for the associated images [24].

Recently, papers have shared a similar, straightforward approach [13,62,77]. They pre-train a multimodal transformer architecture on an image captioning task on a large set of image-caption pairs, like the COCO dataset. The visual backbone, which was pre-trained from scratch, is then transferred to downstream visual tasks to be evaluated against other methods of visual representation learning. This approach is visualized in figure 6.

The notion behind learning visual representation through natural language supervision is that these captions can provide semantically denser learning signals compared to learning signals used in self-supervised learning methods and standard single-modal supervised methods. Namely, captions tend to be more descriptive than class labels. They typically do not only define the entities in the image but also mention the attributes of the entities and the relationship between them. An additional benefit is that, since the internet is full of images with additional information like tags and captions generated by users, it is relatively easy to harvest these datasets. Weakly supervised learning approaches have already tried to exploit this benefit. These approaches aim to learn visual representations on large sets of images with low-quality labels, such as associated tags and other meta-data. The most common approach in current research is learning visual representations with high-quality labels, the associated natural language captions. The previously mentioned studies also showed that the image-caption datasets need to contain far fewer images to still result in competitive visual representation. For example, one

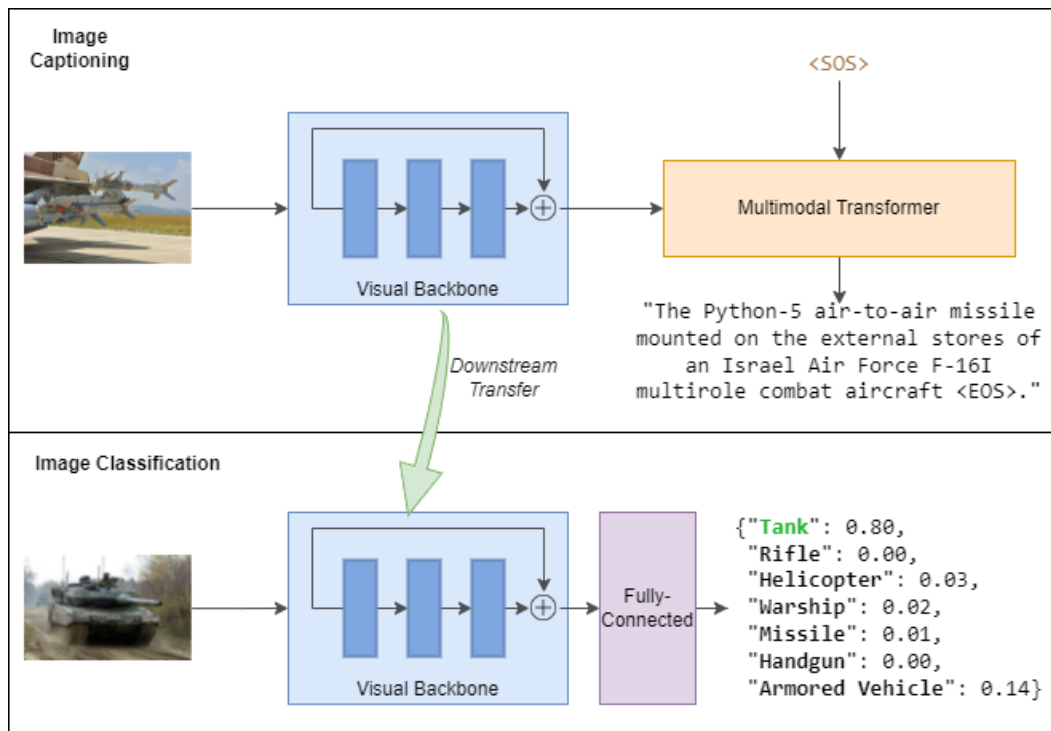


Figure 6: Natural language pre-training of visual features: We investigated the possibility of using natural language supervision for domain-specific visual features. By pre-training on image captioning using a domain-specific image-caption dataset, we aimed to learn domain-specific visual features. We evaluated the learned features on a domain-specific image classification task.

study showed that a ResNet-50 (a SOTA image classification network) pre-trained on an image captioning dataset needed only 10% of the images in ImageNet to achieve similar results as a ResNet-50 pre-trained on the entire ImageNet dataset [13].

The argument for natural language supervision for visual representation learning is threefold: (1) the captions provide a much more semantically dense learning signal than class labels, (2) with the internet, it is has become very easy to collect large image-caption datasets that require little annotation effort and (3) studies have shown that these pre-training methods can achieve a notable performance using fewer images. Therefore, this might be a promising alternative for pre-training vision models.

1.4.1 Vision-language pre-training tasks

Various studies have proposed different pre-training tasks to implement visual-representation learning through image-captions. One study investigated two different pre-training tasks involving image-tag prediction and image-conditioned masked language modelling (ICMLM) [62], visualized in the middle section of figure 7. The success of pre-training in natural-language processing inspired this second task. It involved the prediction of masked tokens (i.e., word representations) in the sequence given the visual information. The authors argued that these tasks directed the model to learn visual representations that localize semantic concepts in the captions. Another study investigated caption generation as a pre-training task, which involved the model generating the natural language caption for a given image [13], visualized in the upper section in figure 7. They argue that this directs the model to distil denser semantic information from the image since it needs to predict all the tokens in the caption

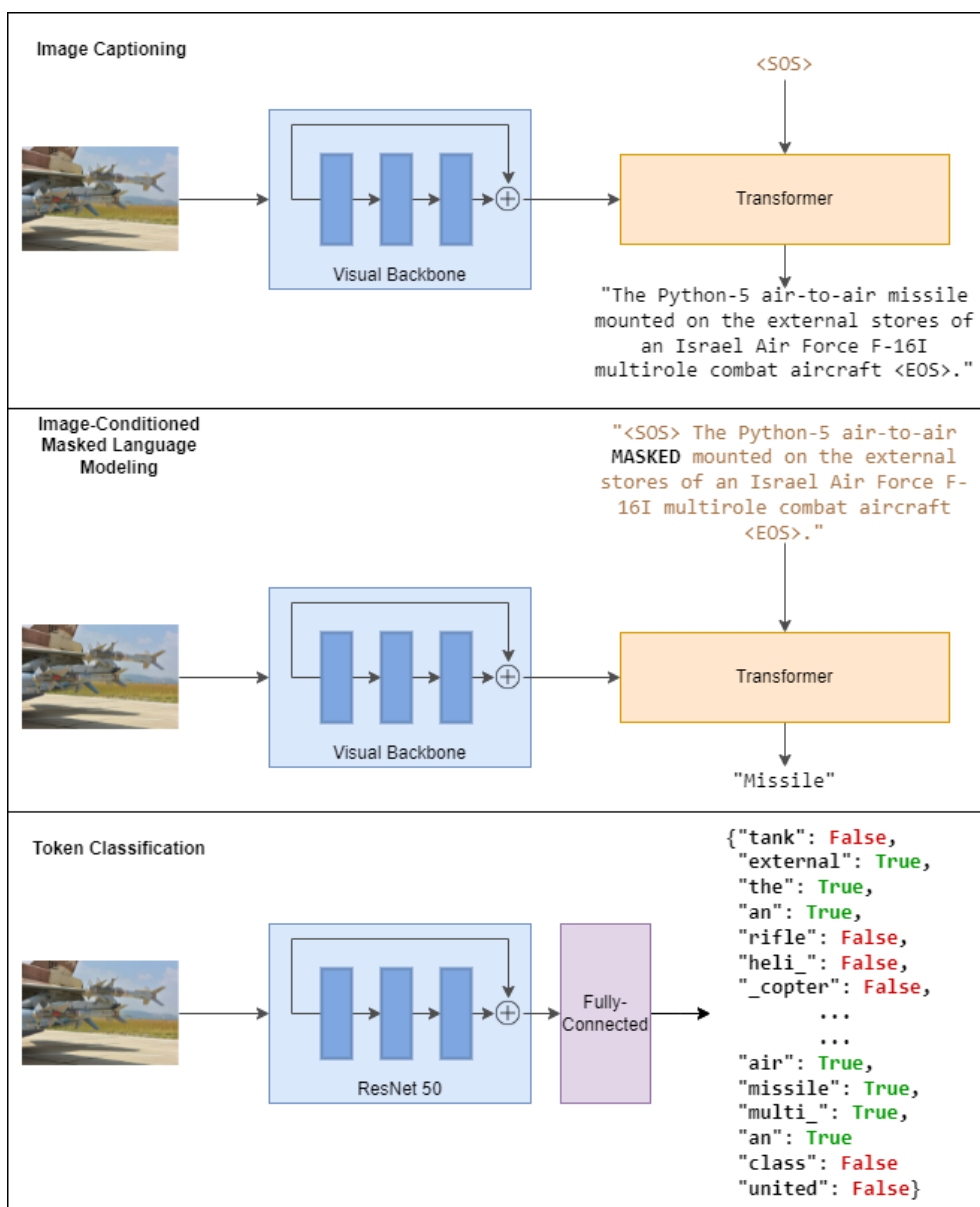


Figure 7: Image captioning pre-training tasks: This study will experiment on the use of three different vision-language tasks, each concerning some sort of image captioning tasks. The first task involves image captioning, in which the model needs to generate a caption given an image. The second task used, is that of Image-conditioned Masked Language Modelling (ICMLM), in which the model needs to predict a masked word, given a caption and an image. The third task concerns Token-Classification, in which the model needs to predict which words are in the corresponding caption, given the set of words in all the captions.

instead of a subset of the tokens. They also explain that masked-language modelling tasks tend to converge a lot slower, which results in decreased computation efficiency for these types of pre-training tasks. Researchers have also investigated multimodal contrastive pre-training tasks to learn visual representations [54,75]. Similar to the previously discussed single-domain contrastive learning, multimodal contrastive learning stimulates the model to have similar representations for corresponding images and captions and dissimilar representations for different images and captions.

Another task relevant to image captioning is that of caption-based image retrieval. Caption-based image retrieval entails finding the best image described by the given caption and a set of images. This task has been investigated by various studies, implementing different approaches such as global [20,22] and local [42] matching of images and text and, more recently, the previously mentioned vision-language models pre-trained on the large scale datasets [10,49]. Since the task of caption-based image retrieval requires the model to reason jointly over visual and textual representations, it could be that this task could also be helpful as a pre-training task for learning visual representations.

Studies concerning vision-language tasks have also used pre-training on extensive collections of image-caption pairs to learn multimodal representations. As stated in the previous section, these studies have primarily opted for masked multimodal modelling and multimodal alignment prediction [10,33,34,36,46,49,50,52,69]. However, other studies also explored pre-training tasks such as image question answering tasks [44,65], contrastive pre-training [51] or caption and tag prediction [69]. Some studies also proposed slight variations of masked multimodal modelling, such as masking with continuous rather than random words [47]. In this study, the researchers opted for another variation of multimodal alignment prediction by explicitly implementing hard negative image-caption pairs. These pre-training tasks are helpful for multimodal representation learning. It could be that these pre-training tasks could also help a model learn visual representations.

Lastly, token classification is another task that researchers have investigated, visualized in the lower section in figure 7. This task requires the model to predict the set of tokens present in the given image. In other words, the model needs to perform multi-label classification over the set of all the tokens present in the vocabulary. Since this task completely ignores the syntactical structure of the captions and therefore does not require a multimodal transformer, strictly speaking, this is a single modal task. Moreover, it is more in line with previously described weakly supervised tasks since the images are annotated with weakly/noisier labels.

1.4.2 *Image-caption datasets*

Language-guided supervision for visual representation learning often requires a dataset that contains image-caption pairs. Most studies have opted for using the Common object in Context (COCO) Captions dataset [9], containing 330k images with five reference sentences per image. A different human annotator generated each sentence/caption and was asked to describe the scene in natural language extensively. This resulted in strongly aligned image-captions pairs, i.e. each caption gives an accurate, dense and relevant description of what is going on in the image. Another dataset that contains strongly aligned image-captions pairs and has also been used for visual representation learning is the flickr30k dataset, which contains 31k images with five reference sentences per image [73].

	Alignment			
	Strong			Weak
	Image	Caption	Image	Caption
Generic Domain		A dog holding a pink frisbee in his mouth.		The remarkable growth of Austin, fueled by a technology boom, has long been shadowed by a rise in homelessness.
Military Intelligence Domain		An air-to-air missile mounted on the external stores of an F-16I combat aircraft.		The Tempest future fighter is likely being earmarked as a host platform for 'smart' missile systems.

Figure 8: Image Captions: Studies into image captioning have primarily opted the use of the COCO dataset, which is a dataset that mostly contains samples that represent the generic domain. And because of the annotation process, this datasets only contains strongly-aligned image-captions, i.e. all the captions contain a fairly accurate and relevant description of the image. This study will opt for a more specialized dataset, containing samples that mostly represent the Military Intelligence domain. Since these images are sourced from the internet, the dataset contains both weakly and more strongly aligned images.

The internet is full of images provided with (textual) information by human users. Moreover, due to progress in processing and cleaning methods, it has become possible to create larger image-caption datasets that contain image-caption pairs crawled from the internet [6,63]. Since these pairs have been harvested from the internet, these datasets mainly contain weakly aligned image-caption pairs. As these datasets are very new, not much work has been done on such weakly aligned image-caption data. However, very recent work on these new datasets has already indicated that it is also possible to learn useful visual representation using weakly aligned image-caption pairs [35]. Accordingly, since it is easier for many domains to collect weakly aligned image-caption pairs than to collect and manually annotate datasets, pre-training on weakly aligned image-caption pairs could be valuable for learning visual representations as a pre-training phase for developing computer vision models. These differences are illustrated in figure 8.

1.5 Current work

As shown above, previous studies into language-guided visual representation learning have implemented various pre-training tasks and datasets. These studies aimed to achieve an adequate performance on generic downstream vision tasks. These studies used generic image-caption pre-training datasets without justifying the choice for this dataset relatively to the targeted downstream vision task and dataset. Moreover, these studies implemented different pre-training tasks without elaborating on these tasks' (dis)advantages. Therefore, it is unclear which pre-training procedure is optimal for learning visual features using natural language supervision.

The military intelligence domain is known for the lack of large, annotated data and its highly specific class taxonomy, resulting in problems with the automation of vision tasks using deep neural networks. Because of the earlier described arguments, natural language supervision could provide a solution. In this study, we want to investigate the possibility to learn visual features relevant to the military intelligence domain by pre-training a visual model using natural language supervision and a domain-specific dataset and investigating the effect of different pre-training tasks.

The main research question is as follows:

1. **Can we use vision-language pre-training to learn useful features for a domain-specific vision task?**

Which this study breaks down into the following two research questions:

1. **How does the domain of the pre-training dataset affect the learned visual features?**
2. **How do different vision-language pre-training tasks affect the learned visual features?**

To investigate if the visual model can extract visual features that are useful for a vision task in the military intelligence domain, we will evaluate the visual model's capability on an image classification task. The effect of different design choices, such as the different pre-training datasets and tasks, can then be compared based on the visual backbone's performance on this task. Image classification is chosen as the evaluation task for two reasons. Firstly, it is the most straightforward vision task from a technical perspective. It only requires the model to label the image and does not require more complicated operations, such as localizing and labelling multiple entities. At this stage, the performance on more difficult vision tasks is less relevant for evaluating the learned visual features. Secondly, as stated before, properly annotated datasets for the military domain are rare, and the few that exist regard image classification datasets. Therefore, the visual features will be evaluated on image classification, using the images labelled with military entities present in the ImageNet dataset. Finally, the different design choices will also be evaluated against the current go-to methods, such as supervised and self-supervised pre-training.

2 Methodology

This study investigates different approaches to learn domain-specific visual representations by pre-training a visual backbone on vision-Language tasks. Two different pre-training datasets and three different vision-language pre-training tasks related to image captioning will be investigated. Besides evaluating against pre-training on a generic image-caption dataset, the performance will also be evaluated against two single-modal benchmarks, namely supervised and self-supervised pre-training. The following sections will first explain the designed experiments, followed by the details of the datasets and tasks, the model's architecture, the protocol for pre-training and evaluation and the benchmarks.

2.1 Experiments

Two experiments will be performed to investigate the possibility of natural language guidance for domain-specific representation learning. The overall goal, design, and expectations of each experiment is discussed below. The overview of the two experiments is shown in table 1.

2.1.1 Experiment 1: Pre-training datasets

The first experiment of this study aims to investigate the effect of pre-training a visual backbone using domain-specific weakly aligned versus generic strongly aligned image captions pairs on the performance on a domain-specific downstream image classification task. Therefore, this experiment will only vary over datasets and focus on a single pre-training task: image captioning. As a generic image-captions dataset, we use a commonly used publicly available dataset, and as a domain-specific dataset, we use a web-scraped image-caption dataset relevant to the military intelligence domain. We will use both the full and a reduced version of the generic image-captions dataset to ensure a fair comparison. The upcoming section, "pre-training datasets", will describe the datasets in more depth. We expect an increased performance on the domain-specific image classification task when pre-trained on domain-specific image-caption tasks since the domain of the pre-training dataset better represents the domain of the downstream task.

2.1.2 Experiment 2: Pre-training tasks

The goal of the second experiment is to investigate the effect of different pre-training tasks on the performance of the downstream vision task. Besides pre-training on image captioning as done in experiment 1, models will also be pre-trained on two different image captioning tasks: image-conditioned masked language modelling and token classification. All three pre-training tasks will be evaluated on the downstream image classification task. We expect that image-conditioned masked language modelling can increase the performance on the downstream task compared to the image captioning task. We also expect that token classification has an inferior performance

Table 1: Overview of Experiments & Benchmarks

Pre-training Task	Pre-training Dataset	Experiment
Image Captioning	COCO Captions	1
Image Captioning	COCO Captions (20%)	1
Image Captioning	Janes Captions	1
IC-MLM	Janes Captions	2
Token-Classification	Janes Captions	2
Supervised	ImageNet	Benchmark
Self-Supervised	ImageNet	Benchmark

compared to the other two pre-training tasks. We expect the model to benefit from learning to predict the sequential structure of the captions and not only the semantic information.

2.2 Pre-training datasets

Both experiments will use two different image-caption datasets: (i) the COCO captions dataset and (ii) Janes Captions, a Military Intelligence dataset scraped from the web for this study. Figure 8 illustrates the main differences between these datasets, namely the image-caption alignment and domain. The datasets are further explained below.

2.2.1 COCO Captions

As described in the introduction, the COCO captions datasets, which stands for Common Objects in Context, is a collection of 118K images with five annotations per image. Each caption was generated by a different individual, asked to describe the scene in natural language extensively. This resulted in strongly aligned image-captions pairs, i.e. each caption gives an accurate and relevant description of what is going on in the image.

It is known that the performance of neural networks tends to increase with an increased number of training samples. Since the COCO captions dataset contains about five times more image-caption pairs than the Janes Captions, a difference in performance could also be explained by this size difference. Therefore, we will also implement a reduced dataset containing only 20% of images in the original COCO Captions dataset, with only one annotation per image.

2.2.2 Janes Captions

For this study, an image-caption dataset is collected from the internet. The image-caption pairs concerned images and captions specific to the Military Intelligence domain, achieved by scraping image-captions from the Janes magazines. Janes magazines provide an authoritative news source and impartial, independent insight across all essential defence and security subject areas. We have access to seven magazines, which mainly concern general defence news, navy news and missiles & rockets. Since this dataset is not explicitly annotated for model development purposes, i.e. the captions are not necessarily dense and descriptive, we will regard this as a dataset containing weakly aligned image-caption pairs. The scraping will result in about 25k image-text pairs.

2.3 Pre-training tasks

This study investigates the use of three different image-caption pre-training tasks: (i) Image captioning, (ii) Image-Conditioned Masked Language Modelling (ICMLM) and (iii) Token-classification (See figure 7). These three tasks will be further explained below.

2.3.1 Image Captioning

Image captioning concerns the task of generating a natural language sequence that describes a given image, similar to the study of Desai and Johnson since they experimented with caption generation [13]. This pre-training task stimulates the model to learn specific semantics from the captions. However, it is quite computational demanding and emphasizes the sequential structure in natural language since the models need to predict an entire sequence in grammatically correct natural language.

2.3.2 Image-Conditioned Masked Language Modelling (ICMLM)

ICMLM concerns the task of predicting masked tokens in a caption given the image [62]. In other words, the models need to predict the masked tokens in a caption given the visual information. This pre-training task allows us the creation of multiple training samples from one image-caption pair while still stimulating the model to learn specific semantics from the caption. However, it has been shown that ICMLM, similar to the single domain task of Masked Language Modeling, requires many training samples and a lot of training time to achieve adequate performance. ICMLM is more computationally heavy than the other pre-training tasks [62].

2.3.3 Token Classification

Token classification concerns the tasks of predicting the set of tokens present in the given image. The model needs to perform multi-label classification tasks over the set of all the tokens present in the vocabulary. This task completely ignores the syntactical structure of the captions and only focuses on the semantics, i.e. which words are present, allowing us to investigate if learning the structure present in natural language is helpful for learning visual representations.

2.4 Architecture

The main architecture used in this study resembles the architecture first proposed by Desai and Johnson [13]. Similarly, it will consist of two main components: a visual backbone employing a ResNet-50 model, which processes the image, and a textual head containing a multimodal transformer that combines the visual output of the visual backbone with textual features. The implemented architecture, and its different components, are further explained below.

2.4.1 Visual component

The visual component of the architecture is responsible for the computation of visual features from raw image pixels. We, therefore, use a convolutional network that can take in the pixel values of an image and output a grid of image features while keeping in mind the spatial properties of the pictures. We will opt for the commonly-used ResNet-50 architecture as the visual backbone to facilitate comparison with our benchmarks since they implemented the same ResNet-50 architecture. The ResNet (Residual Neural Network) architectures are convolutional neural networks that implement residual connections between each layer to allow for deeper neural networks without the problem of vanishing gradients (visualized in figure 9). The ResNet architecture achieved state-of-the-art results in many vision tasks such as image classification, object detection and face recognition[29].

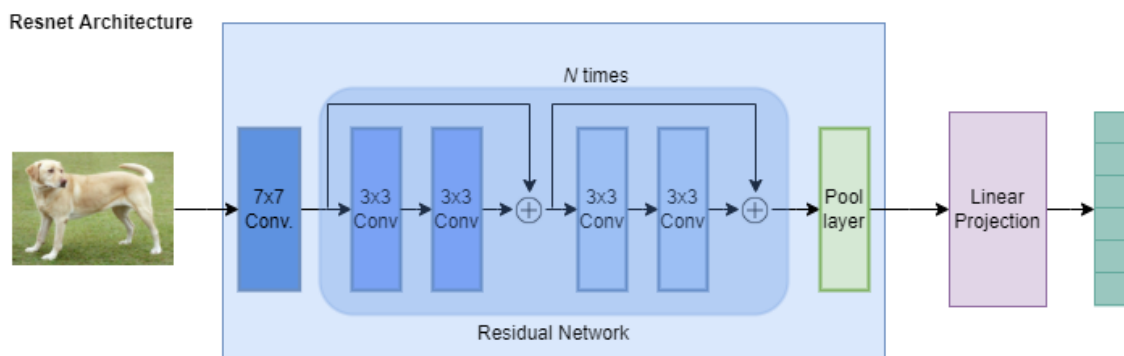


Figure 9: **Residual Network:** ResNets implement residual connections that jump over some layers. Jumping over the layers, effectively simplifies the network in the early stages, since the skipped layers are paid less attention to at the beginning of training. This helps to avoid the problem of vanishing gradient and therefore allows for much deeper layers.

The visual component takes in a 224 x 224 image and produces, after the pooling layer, a 7 x 7 grid of 2048-dimensional visual features corresponding to the input dimensions of the textual head. During pre-training, the shape of the visual features is reduced using a fully connected layer to be passed to the textual component to fuse with the textual features and perform caption prediction. We train one final fully connected layer to the visual component during the evaluation to evaluate the extracted features on image classification, further described below.

2.4.2 *Text tokenization*

To enable the model to handle the textual input, it needs to have a vocabulary, which is done by tokenising the text. Tokenization is a preprocessing step fundamental to most machine learning models that involve language. It involves splitting text into smaller units called tokens, which can consist of words and subwords (smaller word segments). After creating our set of unique tokens, we can use this vocabulary to turn any piece of text into a sequence of discrete elements. The model can then input the embedding for each token at every point in the sequence. It can also use the set of tokens to turn the output of probabilities into a sequence of tokens, generating a caption.

Like other studies, we use SentencePiece [40] to tokenize our captions, i.e. create a vocabulary of tokens from the words present in our captions. SentencePiece is an unsupervised text tokenizer that aims to have the most reflective set of tokens, given the desired token vocabulary size and a set of words, in this case, all the words present in our captions. Compared to other basic tokenization methods (such as splitting on whitespace), Sentence Piece makes fewer linguistic assumptions and exploits subword information, resulting in fewer out-of-vocab tokens, which is very important because of the highly domain-specific vocabulary present in our captions. Before tokenization, our captions are lowercased, and accents are stripped. The size of our token vocabulary is set to 10K and includes the SOS (Start-of-Sentence) and EOS (End-of-Sentence) tokens, as well as the UNK (Out-of-vocab) token.

2.4.3 *Textual component*

The textual component of the architecture is responsible for combining the visual and textual features to predict captions for the inputted image. These components provide the learning signal to the visual backbone. In line with recent advances in language and vision-language modelling, we will opt to use multimodal transformers, which can implement multiheaded self-attention to fuse the visual and textual features (figure 6).

We will use two identical transformer models for image captioning to predict the captions in forward and backward directions, respectively, also known as bi-captioning. The task of Image-conditioned Masked Language Modelling will only consist of a single transformer model, predicting the masked token in a given caption and given image. We will implement a simple fully-connected layer for token classification, which performs multilabel classification on the entire set of tokens in the vocabulary.

The transformer receives both the image features from the visual backbone and a caption, a sequence of 30 tokens describing the image. Before the sequence is inputted, each token is converted to a vector via learned tokens embeddings and positional embeddings. The first layer in the multimodal transformer performs masked multiheaded self-attention over each token vector. For each of the target tokens in a given sequence, the consecutive tokens are masked such that the computation depends solely on the preceding ones. This is introduced

to maintain the casual nature of the sequence. Together with the image vectors, the token vectors are combined using multiheaded attention. After this, a fully connected network is applied to each vector. An overview of the multimodal transformer component in our architecture is given in figure 10. After the multimodal transformer, a fully connected layer is applied to each vector to get probability over all the tokens in the vocabulary at each spot in the sequence.

The pre-training tasks are visualized in figure 7, and the detailed visualization of the proposed architecture, with image captioning as a pre-training task, can be found in figure 10.

2.5 Pre-training & Evaluation

The architecture described above will be pre-trained on three different pre-training tasks. This pre-training will follow the same protocol for each task. After pre-training, we will evaluate the visual backbone's performance on an image classification task using an evaluation protocol. The protocols used for pre-training and evaluation are further described below.

2.5.1 Pre-training

The pre-training protocol follows the protocol implemented by Desai & Johnson. To save computation time, we only pre-train the architecture on our dataset, the Janes dataset and load in a visual backbone pre-trained on the COCO dataset, distributed by Desai & Johnson[13]. To compare the performances of these visual backbones, we will follow the same pre-training protocol to ensure that any differences in performance are not due to underlying differences.

At the start of pre-training, the visual backbone is initialized randomly. During the pre-training and evaluation phases, we apply standard image augmentation. We first randomly crop the image to between 20% and 100% of its size, apply jitter to the brightness, contrast, saturation and hue, and normalize the colour using the mean

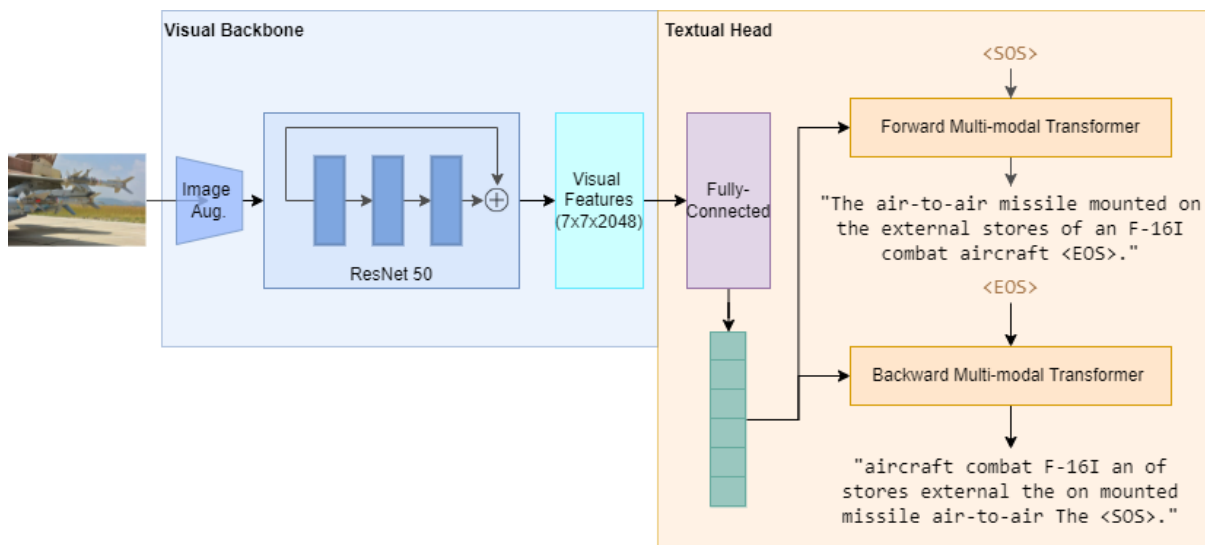


Figure 10: Proposed Architecture: Our architecture consists of a visual backbone, a ResNet-50, and a textual head, a multimodal transformer that combines the visual and textual features. In this visualization, image captioning is implemented as pre-training task. The visual backbone extracts the visual features from an image and the textual head combines these image features, after linear projection, with the textual features using multimodal transformers. This figures implements two transformers in order to perform bi-captioning. This entire architecture is pre-trained end-to-end, after which the visual backbone is transferred to the image classification task for evaluation.

ImageNet RGB values. We finally randomly flip images horizontally, interchanging the words ‘left’ and ‘right’ during pre-training.

Optimization during pre-training implements Stochastic Gradient Descent (SGD) with a momentum of 0.9 and weight decay of 10^{-4} . Pre-training is performed across 2 GPUs with normalization per GPU, with a batch size of 256 images (128 per GPU). Pre-training runs for 200k iterations or roughly 1080 epochs. The learning rate follows a linear warmup protocol, increasing linearly from 0 to the desired value in the first 4000 iterations and then decreasing to 0 using a cosine decay schedule. We used a max learning rate of 2×10^{-1} for the visual backbone and a max learning rate for the textual component of 10^{-3} . All models, datasets and pre-training tasks were implemented using PyTorch [58].

2.5.2 Evaluation

The visual backbone’s capability to learn domain-specific visual representations will be evaluated using a commonly-used vision task, namely, image classification. We have opted for image classification because of the absence of military datasets for more complicated vision tasks. Besides, since little is known about visual backbones pre-trained using natural language, the model’s capabilities concerning more difficult vision tasks are less relevant at this stage.

Since we are trying to learn visual features that could be relevant for application in the military domain, we have opted to evaluate on a subset of the ImageNet image classification dataset. We have handpicked 21 classes relevant to the military domain present in the ImageNet taxonomy and only selected the images labelled with one of these classes for our evaluation task. This resulted in a training dataset of 27,190 images, out of 1,28M (2,1%) images, and a test set of 1050 images, out of 50K (2,1%) images. The handpicked classes are *Aircraft Carrier*, *Airship*, *Amphibian*, *Assault Rifle*, *Bulletproof Vest*, *Cannon*, *Gasmask*, *Half-Track*, *Holster*, *Jeep*, *Military Uniform*, *Missile*, *Parachute*, *Police Van*, *Projectile*, *Revolver*, *Rifle*, *Submarine*, *Tank*, *Warplane & Wreck*.

We will follow the widely used linear evaluation protocol for the evaluation, where a fully connected layer is trained on top of the frozen base network using the train split of the dataset. The achieved accuracy on the validation split is then used as a proxy for the quality of the visual representations learned by the visual backbone. Besides allowing for a more direct evaluation of the model’s ability to extract visual representations, this protocol saves computation costs. We will perform the same augmentations as in the pre-training protocol described in the previous section during training and testing.

The linear evaluation protocol starts by adding a fully connected layer applied to the 2048-dimensional global average pooled features extracted from the last layer of the visual backbone. Training is then performed using the train split of our ImageNet subset dataset with a batch size of 256 for 100 epochs, using just a single GPU. Like pre-training, we implement SGD for optimization with a momentum of 0.9 and weight decay of 0. We set the initial learning rate to 0.3 and decay to zero using a cosine schedule.

2.6 Benchmarks

Besides evaluating different pre-training tasks and datasets, we also evaluate the methods against other recent pre-training methods that have successfully learned visual representations.

2.6.1 *ImageNet supervised*

The first benchmark that we will use is straightforward pre-training on ImageNet. We will use a pre-trained ResNet-50 pre-trained on ImageNet classification[29]. Since their introduction, Resnet models pre-trained on the entire ImageNet dataset have been the go-to for downstream vision tasks. In this study, we have opted for a ResNet-50 since this is also the architecture used for our visual backbone. Therefore, any (noteworthy) differences in performance will be due to the differences in pre-training methods and datasets. The linear evaluation protocol will be applied when evaluating the representations learned by this benchmark to compare the representations of this benchmark against the representations learned by our models. Figure 1 visualizes the supervised-pre-training method.

2.6.2 *SimCLR self-supervised*

The other benchmark used in this study is that of contrastive self-supervised pre-training. Specifically, a ResNet-50 pre-trained using the contrastive learning framework that recently has achieved adequate performances on downstream vision tasks, also known as SimCLR [7] (figure 2). Once again, the ResNet-50 architecture is chosen to ensure that any differences in performance are not due to differences in architectures. Similar to the supervised model, we will apply the linear evaluation protocol to compare these representations against the representations learned by our models.

2.7 **Qualitative analysis**

Besides using accuracy as a proxy to measure the quality of the visual features learned by the visual backbone, we also perform some qualitative analysis to further investigate the results and differences between our architecture, pre-training tasks and datasets. We will construct the confusion matrices on the downstream task, visualize the learned visual features using dimensionality reduction and investigate predicted captions on the pre-training and downstream dataset images.

2.7.1 *Confusion matrices*

First, the achieved performance on the domain-specific task is further investigated by inspecting the confusion matrices. The confusion matrices will allow us to investigate the classification mistakes made by the different pre-trained visual backbones.

2.7.2 *Dimensionality reduction*

Next, we will also perform dimensionality reduction on the visual features, reducing the visual features to a 2-dimensional representation, allowing us to compare the features between different pre-trained models visually. Since the visual backbones output 2048-dimensional visual features, these cannot be visualized in a two or three-dimensional graph. Therefore, we need to apply a dimensionality reduction technique, which projects the high-dimensional (2048) features to a low (2) dimensional embedding space while keeping overall structure and distances between vectors similar. Using Uniform Manifold Approximation and Projection (UMAP), we will project the 2048-dimensional features to a 2-dimensional space.

2.7.3 *Image Captioning*

Finally, we will also address the image captioning performance of our pre-trained models by investigating captioned examples. Although our study is interested in learning domain-specific visual features through image captioning and we have not looked detailed into the image captioning task itself. We will compare some predicted

captions on the Janes captions dataset by a model pre-trained on the domain-specific dataset and a model pre-trained on the generic captions dataset. We will also review some predicted captions on the downstream validation split images. Although these images are not annotated with captions, it is interesting to observe what captions the models predict and if they contain any reference to the correct label.

2.8 Additional analysis

Some additional analyses will be conducted to identify further the effect of different design choices on the learning of the visual features and the resultant performance on the downstream vision task. These investigations further reveal the inner dynamics of the architecture, but it also points towards interesting future works. These additional analyses cover the (1) model complexity, (2) pre-training parameters and (3) dataset cleaning.

2.8.1 Textual head complexity

The first analysis involves changing the complexity of the model, more specifically, the textual component. Besides the number of training samples, it is known that neural networks also increase in performance when their complexity, i.e. the number of changeable parameters, increase in size. Therefore, varying the complexity of the textual head is investigated, specifically increasing the number of attention heads and layers in the multimodal transformer component. Altering these two hyperparameters allows increasing or decreasing the number of the parameters in the textual head. We expect that increasing the complexity of the textual head could increase the performance of the visual backbone on the image classification task, even though the textual head is not directly involved with the classification task. A more complex textual head has more computation power to integrate the visual and textual features and create a better supervisory signal for the visual backbone.

2.8.2 Pre-training iterations

The second analysis concerns an investigation into the pre-training performance and relation to the performance on the image classification task. Our architecture is trained on image captioning tasks for 200k iterations during pre-training, increasing its performance throughout these iterations. This analysis investigates the relationship between the performance on the pre-training task and the performance on the classification task by comparing the performances during the pre-training phase. We expect that increased performance on image captioning precedes an increase in the performance on the classification task.

2.8.3 Captions pre-processing

Finally, the effect of pre-processing our captions on the image classification task is investigated. Since the captions in our dataset represent the military intelligence dataset and are collected from a highly specialized source, the captions are characterized by a large and specific vocabulary. For instance, certain weapons are often described by their exact type, instead of a simple denominator such as “weapon”. Also, since our dataset is collected from the internet, the captions show a weaker alignment to the images. As an example, captions often speak of locations in which an image is taken (such as a country’s name) which is information that is not directly derivable from the image itself. By cleaning the captions, we aim to decrease the number of unique words in the captions, resulting in a less specific vocabulary and (slightly) increasing the alignment between the images. We expect that pre-training on the cleaned captions would increase the performance on the image classification task. The pre-processing of the Janes captions consists of two steps (1) cleaning the captions and (2) transforming specific named entities to more generic named entities. These two steps decrease the number of unique words in the captions from 40K to 32K.

The cleaning involves removing the punctuation marks from the captions, lower-casing all the characters and removing the accents, similar to the cleaning process applied to the COCO captions. Since the captions will be collected from the internet, some captions also contain extra whitespace, which we remove to make them more similar to the COCO captions. Similarly, we strip all the text between parentheses since this is not present in the COCO captions. These cleaning steps are adapted from the COCO captions and the Conceptual Captions datasets pre-processing steps [9,63].

The transforming of named entities consists of two steps. First, we use the Wikidata knowledge base to recognize specific named entities in the dataset and transform them to their more generic hypernym. Using this method, we can, for example, transform the named entity ‘HMS Prince of Wales’ to its hypernym ‘Aircraft Carrier’ and the named entity ‘M1969 Mountain Gun’ to its hypernym ‘Cannon’. We also use a pre-trained named entity recognition model to detect the named entities of the type location, date and quantity and transform these named entities to their types. For example, this allows us to transform the named entity ‘Pakistan’ into ‘Location’, ‘April 12’ into ‘Date’ and ‘Fifty’ into ‘Quantity’.

3 Results

In this study, two experiments were executed to better understand the learning of domain-specific visual features through natural language supervision. Below, the first experiment's results concerning a domain-specific dataset will be discussed. Following this, the second experiment, focusing on different pre-training tasks, will be discussed. Besides these two main experiments, qualitative analyses were conducted to further investigate the inner workings, which are also presented below. Finally, some additional results will be discussed. Table 2 gives an overview of all the results presented below.

Table 2: **Overview of the results**

Pre-training Task	Pre-training Dataset	Accuracy (%)
Image Captioning	COCO Captions	58,9
Image Captioning	COCO Captions (20%)	71,6
Image Captioning	Janes Captions	55,0
IC-MLM	Janes Captions	55,2
Token-Classification	Janes Captions	52,6
Supervised	ImageNet	80,5
Self-Supervised	ImageNet	76,0

3.1 Experiment 1

The first experiment involved using a domain-specific image-caption dataset (the Janes Captions dataset) to learn more domain-specific visual features. This was compared against using a generic captions dataset (the COCO dataset) to test our hypothesis that domain-specific captions increase the specificity of the learned visual features.

The results of the experiment are shown in figure 11. The visual features learned using pre-training on the Janes dataset are outperformed by the visual features learned using the COCO captions dataset. The COCO dataset was

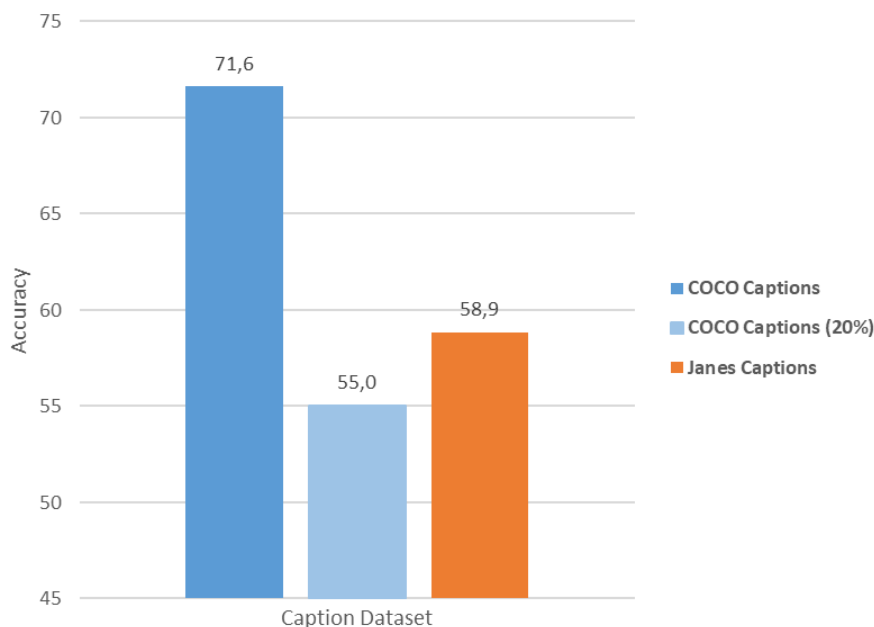


Figure 11: **Generic Image-Captions versus 20% Generic Image-Captions versus Domain-Specific Captions:** Pre-training on the larger generic dataset resulted in an increased performance on the domain-specific Image Classification task (71,6%). When we pre-train using a comparable sample-size, pre-training on the domain-specific dataset resulted in a better performance (58,9%) compared to pre-training on the scaled down generic dataset (55,0%).

also reduced by 80%. This allowed for a fairer comparison to the Janes dataset, independent of the dataset size. A comparison between pre-training with the 20% COCO dataset and the Janes dataset reveals that pre-training on domain-specific captions outperforms pre-training on generic captions with comparable dataset sizes.

3.2 Experiment 2

The second experiment focused on the effect of using different pre-training tasks on the learned visual features. This experiment compared three different tasks, namely (1) bi-directional image captioning, (2) Image-conditioned Masked Language Modelling and (3) token classification, to test our hypothesis. We expected that tasks that require the model to predict the sequential structure of the caption (1&2) outperform models that only need to predict the semantic information (3).

The initial results of this experiment are shown in figure 12. The visual features learned by pre-training on either image captioning or Image-conditioned MLM outperform the visual features learned by pre-training on token classification.

3.3 Benchmarks

Besides comparing different design choices, the best performing pre-training method (bi-directional image captioning) was also evaluated on two pre-training methods that have successfully learned visual features for downstream vision tasks. We compared our method against (1) supervised pre-training on the ImageNet (IN) dataset and (2) unsupervised pre-training using the SimCLR framework.

The results of this comparison are shown in figure 13. Learning visual features through image captioning is outperformed by visual features learned through both supervised and unsupervised pre-training.

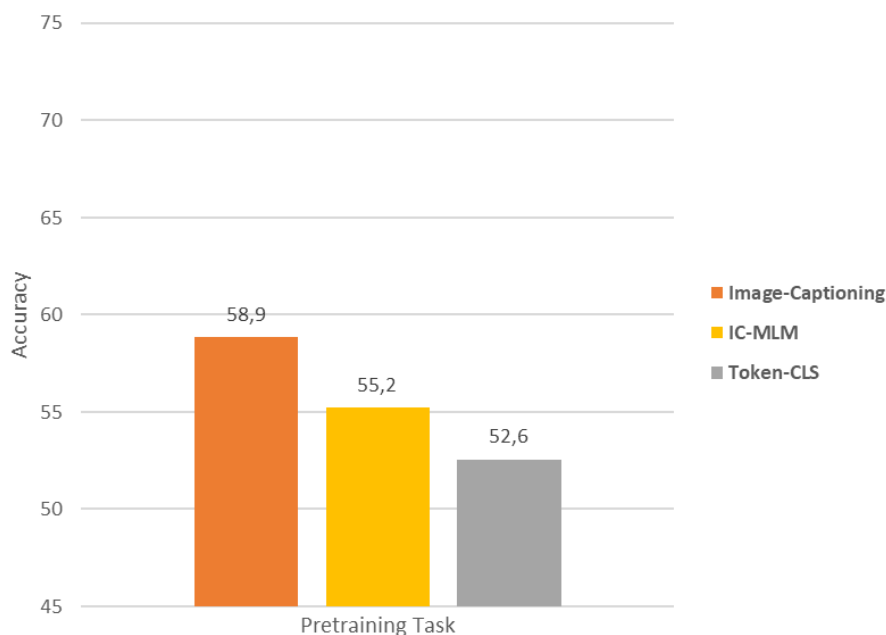


Figure 12: Image captioning versus Image Conditioned Masked Language Modelling versus Token-Classification: Pre-training on Image captioning (58,9%) and Image-Conditioned Masked Language Modelling (55,2%) leads to an increased performance on the domain-specific image classification task, when compared to Token-Classification (52,6%).

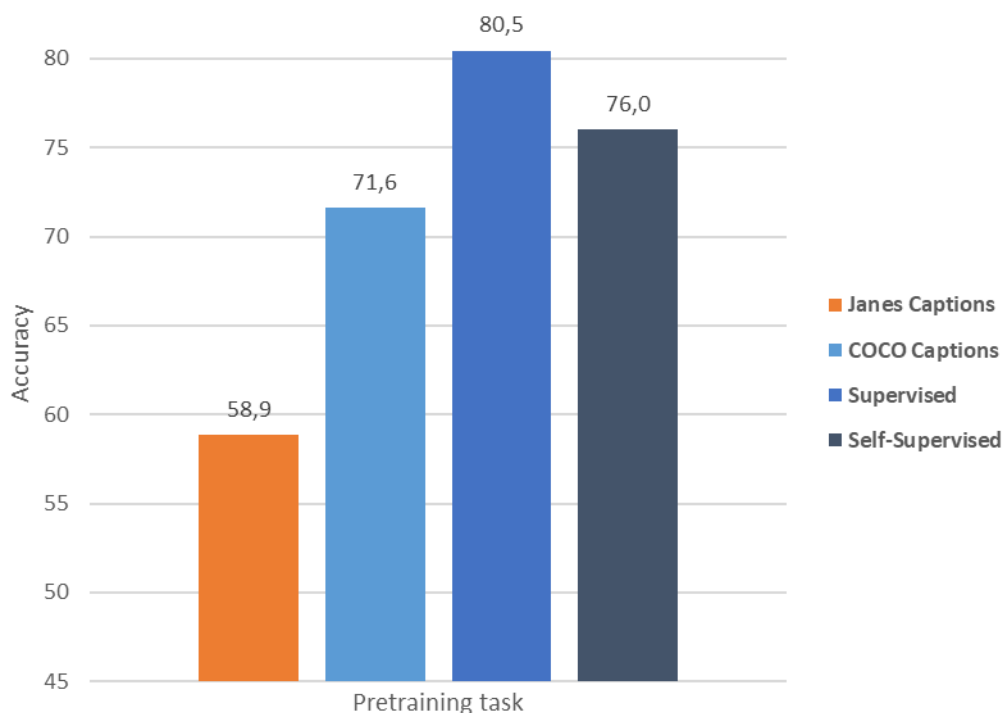


Figure 13: Domain-Specific Image captioning versus Generic Image captioning versus ImageNet Supervised versus Unsupervised: Image captioning on domain-specific image-captions (58,9%) and Image captioning on Generic image-captions (71,6%) are outperformed by the two benchmarks. Unsupervised pre-training (76,0%) is outperformed by Supervised pre-training, which used the ImageNet dataset.

3.4 Qualitative results

Besides using accuracy as a proxy to measure the quality of the visual features learned by the visual backbone, we also performed a qualitative analysis. Confusion matrices were constructed on the downstream task, the learned visual features were visualized using dimensionality reduction, and predicted captions were investigated on both the pre-training and downstream dataset images.

3.4.1 Confusion matrices

Figure 14 compares the performance of the Janes Captions pre-trained model against the model pre-trained on the reduced (20%) COCO Captions and the two benchmarks. One can note that for the natural language-guided models, ‘Missile’ and ‘Projectile’ classification is still very confusing. This confusion is also present, to a smaller extent, in both the supervised and unsupervised pre-trained models. All four models do not experience difficulties classifying ‘Police Wagon’, ‘Parachute’ and ‘Aircraft Carrier’. However, the benchmarks can often classify ‘Cannon’ and ‘Gasmask’ correctly, while the natural language guided models seem to fail slightly more often. Part of the performance difference between pre-training on a domain-specific and generic dataset could be explained by the increased accuracy of the ‘Airship’ class.

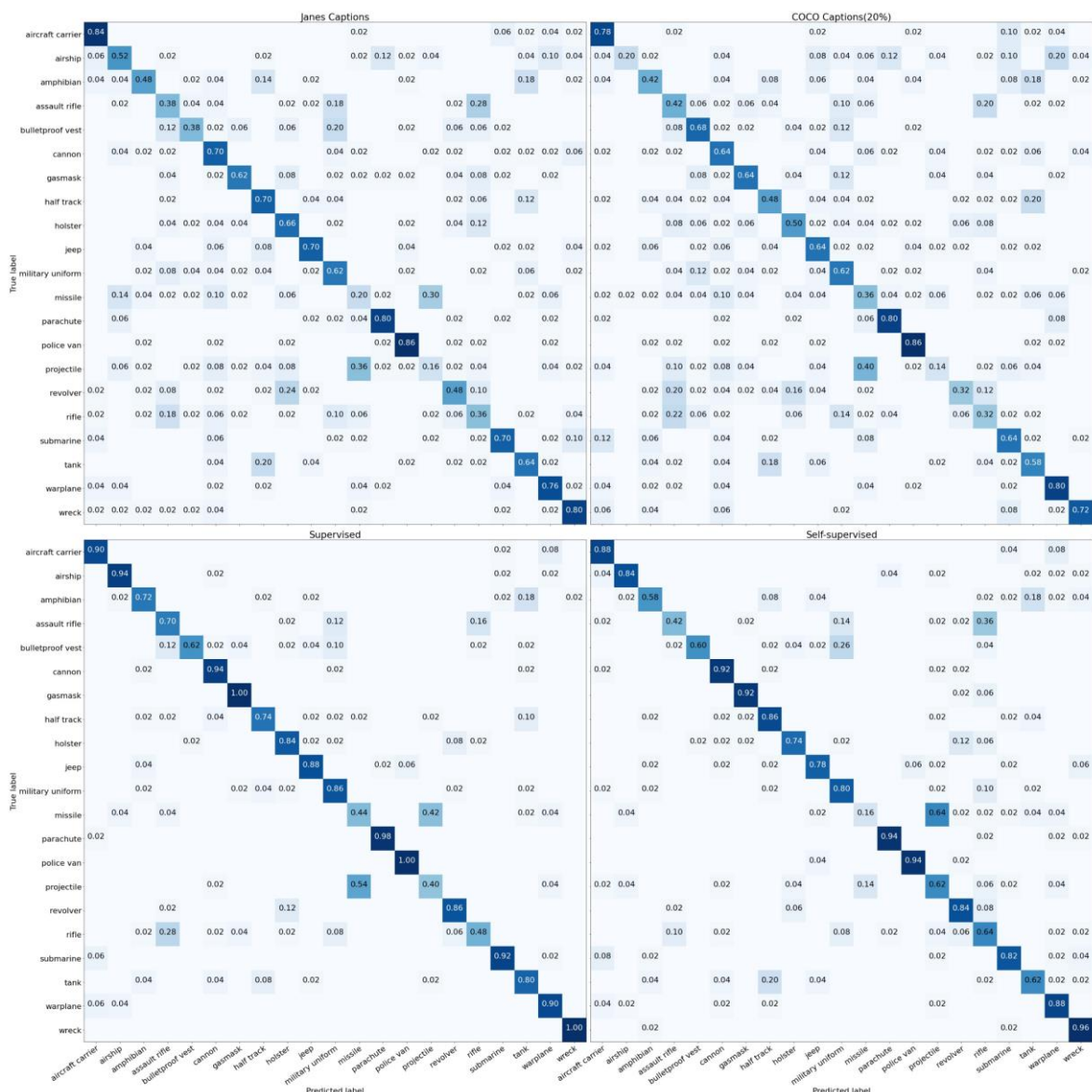


Figure 14: Confusion Matrices on the downstream domain-specific Image Classification task: The confusion matrices of four different pre-training methods have been constructed to further investigate the difference between the methods. In the figure above, the x-axis represents the predicted labels, whereas the y-axis represents the true labels. The numbers in each box represent how many times (in percentage) a certain class was predicted for a true class. If a certain class was never predicted for that class, the box remains empty. As can be observed, the benchmarks, shown in the bottom row, show much less confusion when compared to the natural language guided methods.

3.4.2 Dimensionality reduction

The following analysis concerned visually inspecting the learned visual features. Applying dimensionality reduction allows us to inspect the 2048-dimensional vectors in a 2-dimensional space.

Figure 15 visualizes the visual representations of the models pre-trained on the Janes Captions and the reduced (20%) COCO captions and the two benchmarks. Only the supervised benchmark shows an improved clustering of the different classes in the feature space, most likely because the supervised benchmark is pre-trained on the ImageNet dataset and therefore trained on the military classes in our domain-specific classes. It is also interesting to note that the unsupervised method does not have this clear divide but still shows some structure in the shape of



Figure 15: Dimensionality reduction on the visual features: The figure above visualizes dimensionality reduction on the visual features extracted for the downstream image classification task, for four different pre-training methods. The plots are scaled, such that the length of the x-axis is the same as the length of the y-axis. The colors represent the 21 different classes in the downstream task. As can be observed, the benchmarks, shown in the bottom row, show more structure in their visual representations. The supervised method shows the best clustering of the different classes.

the features, whereas the methods pre-trained on image captioning do not show any structure. This could explain the difference in performance between the benchmarks and natural language-guided models.

3.4.3 Image captioning

Although this study is interested in learning domain-specific visual features through image captioning, the image captioning task itself has not yet been discussed in great detail. Figure 16 shows some predicted captions of images from the Janes validation split by the model pre-trained on the Janes captions dataset and the model pre-trained on both 100% and 20% of the COCO captions. Captions predicted by the Janes model contain very specific vocabulary. In contrast, the COCO captions models use a more generic vocabulary. However, the captions generated by the Janes caption model are much more incoherent considering the syntax; i.e. the captions are not grammatically correct. The COCO captions seem to be able to describe on a high level what is happening in the four images containing either a ship or an airplane without using specific or explicit language. However, both COCO captions models seem to struggle with the image that contains a helicopter, whereas the Janes captions refer to the helicopter by its correct name (AH-1z).





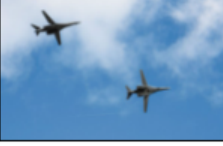
Image	Predicted Captions			True Caption
	Janes Captions	COCO Captions	COCO Captions (20%)	
	"the ah-1z has been deployed into the apache ah-1y."	"a close up of a motorcycle with a helmet on it."	"a man riding a skateboard on the side of a road."	'In two major features IDR investigates technologies adopted to keep helicopters on the frontline of asymmetric battles.'
	a mock-up of lockheed martin lockheed martin's harrier. the usaf is looking for up to carry the usaf's impression.	"a plane on the tarmac with people standing around."	"a fighter airplane sitting on top of a runway."	'The first aircraft to carry the MP-RTIP active array radar will be the Northrop Grumman Global Hawk. Flight tests will start in 2007.'
	"the anzac-class frigate hms albion during an exercise 'joint winter 2004'."	"a full view of several boats in the water."	"a man riding a skateboard on the side of a road."	The front line of the fight against piracy. ships from Combined Task Force 150 replenish at sea.
	"the royal netherlands navy's new zealand navy type 45 destroyer hms albion."	"a boat on the water with a boat in the water."	"a man standing on top of a paddle boat in the water."	"One victim of the Sea Dart missile system during the Falklands War was a British Army Gazelle helicopter."
	"northrop grumman has conducted the first flight of a unmanned aerial vehicle."	"a plane flying in the sky with a sky background"	"a plane flying through the sky in the sky."	'The B-1B Lancer bomber is experiencing a public renaissance after years of performing critical, but lower-profile missions'

Figure 16. Predicted Captions on the Janes Captions validation split: The models pre-trained on image captioning were used to generate natural language captions on five images of the Janes Captions dataset. This allows observation of the model’s entire architecture capacity to caption an image correctly. Although not the main target of the study, it is interesting to observe that the Janes Captions model uses a much more specific vocabulary, when compared to the COCO Captions models.

We also show some predicted captions from the military subset validation split. Although these images were not annotated with captions, it is interesting to observe what captions the models predict and if these contained any reference to the correct label. These can be observed in figure 17. Once again, one can observe that the Janes captions model uses a much more specific vocabulary compared to the COCO captions models. The Janes captions model also seems to understand better which entity is represented in the image, referring to “fighter aircraft” in the “Warplane” image and “aircraft carrier” in the “Aircraft Carrier” image. The “Assault Rifle” image and “Tank” image have captions that are not far off, using “Machine Gun” and “Light Assault Vehicle” instead. On the other hand, the COCO captions seem to struggle more with these captions, referring to the “Assault Rifle” as “a chair” and the “Aircraft Carrier” as a “Stop sign”.

3.5 Additional analysis

After the initial results, additional analyses were conducted to identify further the effect of different design choices on the learning of the visual features and the resulting performance on the downstream vision task. The results of these additional analyses are further discussed below.

3.5.1 Textual head complexity

The first additional analysis involved changing the complexity of the textual component of the model. This was investigated by altering the number of attentions heads and layers in the multimodal transformer. As shown in figure 18, decreasing the number of attention heads resulted in lower performance on the downstream task while






Image	Predicted Captions			True Label
	Janes Captions	COCO Captions	COCO Captions (20%)	
	"the first example of the italian air force's next-generation fighter aircraft."	"a plane flying in the air with a mountain in the background."	"a plane flying over the air on top of an airport."	Warplane
	"lockheed martin's latest version of the italian army's new mobile surveillance platform."	"a clock tower in the middle of a park."	"a man standing next to a red fire hydrant."	Missile
	"the modular weapon system is equipped with a 12.7 mm machine gun."	a chair and a chair in front of a window."	"a stove top oven sitting on top of a stove."	Assault Rifle
	"an lav 8x8 light combat vehicle equipped with a roof-mounted remote weapon station."	"a truck parked on the side of the road."	"a boat in urban area with people on it."	Tank
	"the french navy's nimitz-class aircraft carrier uss harry s truman (pictured) with the royal navy's nimitz-class nuclear-powered aircraft."	"a boat floating on top of a body of water."	"a stop sign on the beach in the ocean."	Aircraft Carrier

Figure 17: Predicted Captions on the Military Image Classification tasks validation split: The models pre-trained on image captioning were also used to generate natural language captions on five images of downstream military image classification task. Although these images do not have any annotation in the form of a natural language caption, they do contain a class annotation. Performing caption generation on these images allows us to compare the model’s capabilities, to observe if they can recognize the object in the image correctly and if they mention the object’s class. It is interesting to note that the Janes Captions models can mention at least two classes correctly.

increasing the number of attention heads or the number of layers slightly increased the performance. These results could indicate that difference in performance between pre-training tasks, as shown in experiment 2, could also be due to a difference in textual component since image captioning uses a more complex textual head than ICMLM.

3.5.2 Pre-training iterations

The second analysis investigated the relationship between the performance on the downstream task and the pre-training task and the effect of the pre-training iterations. We observed that the model quickly overfitted on the image captioning task during pre-training. However, after pre-training (and overfitting), the visual backbone could still perform well on the downstream task. Accordingly, we investigated how the performance of the backbone on the downstream task changed as the performance on the pre-training task changed during pre-training.

Figure 19 shows how the downstream performance of the visual backbone changes during pre-training. At around ~30k iterations, the model starts to overfit as the training loss decreases to zero while the validation loss increases. The overfitting on the image captioning task does not seem to harm the downstream performance of the visual backbone.

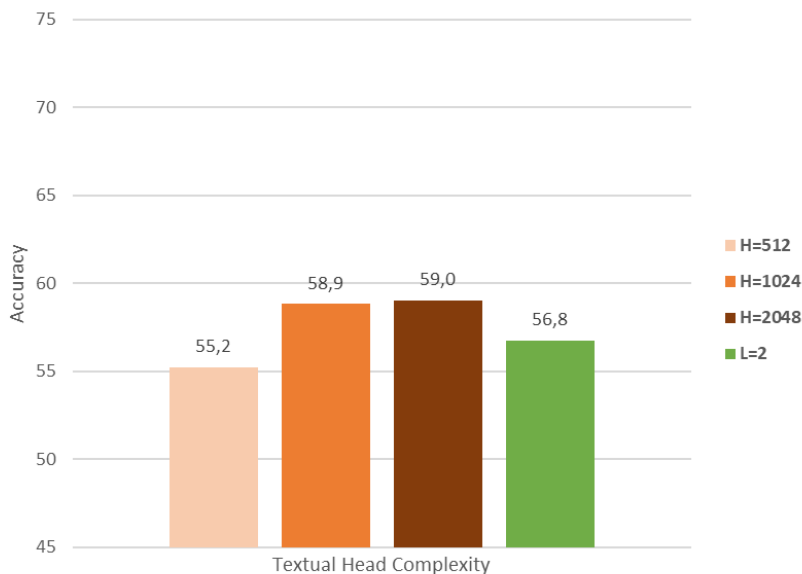


Figure 18. **Decreased Attention Heads (H) versus Increased Attention Heads versus Increased Layers (L):** Pre-training using a more complex textual head, implemented by increasing the amount of Attention Heads (H=1024 & H=2048) and Layers (L=2), increases the performance on the domain-specific Image Classification task, compared to pre-training on a less complex textual head (H=512).

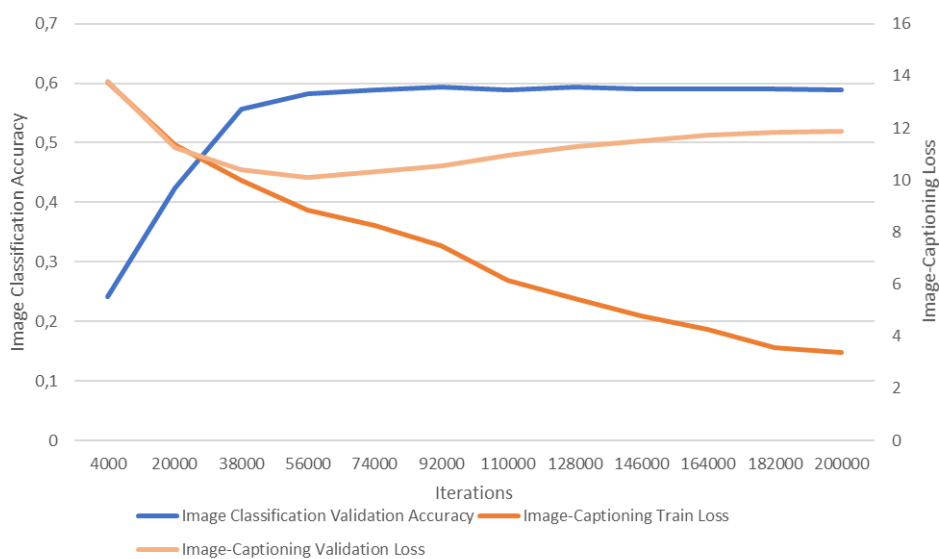


Figure 19: **Image captioning loss during pre-training versus Image Classification accuracy during evaluation:** During the pre-training phase, the model starts to overfit on the pre-training task (Image captioning), as observed through the loss function, after ~30K iterations. Around the same number of iterations, the performance of the visual backbone on the downstream image classification task stops increasing, plateauing around ~59% accuracy on the validation split. Although the architecture overfits on the pre-training task, this overfitting does not seem to hurt the visual backbone’s performance on the downstream task.

3.5.3 Captions cleaning

The final additional analysis regards pre-processing the captions to reduce the number of unique words in the captions from 40K to 32K. In figure 20, one can observe the effect of this cleaning process on ten captions. The effect of pre-training on these processed captions on the performance on the image classification task was

investigated. The results of this experiment are shown in Figure 21. The figure shows that the pre-processing of the captions negatively impacted the performance of the visual backbone on the downstream task.






Image	Original Captions	Cleaned Captions
	"Northrop Grumman's X-47B is a leading contender to be the first unmanned combat aerial vehicle (UCAV) to make it into service. This month's cover feature surveys the current crop of UCAVs in development around the world, in an attempt to find out how close the era of the UCAV is."	"grumman s is a leading contender to be the combat vehicle ucaV to make it into service this date ' cover feature surveys the crop of in development around the world in an attempt to find out how close the era of the ucaV is"
	"An SEP company demonstrator in 8x8 configuration fitted with mechanical drive system and Lemur RCWS armed with a .50 calibre M2 machine gun."	"an company demonstrator in 8x8 configuration fitted with drive system and lemur rcws armed with a quantity m2 machine gun"
	"Training over the last decade has focused mainly on supporting infantry operations, but in line with the drawdown from Afghanistan and Iraq is now looking once again at formation manoeuvres."	"training over date has focused mainly on supporting infantry operations but in line with the drawdown from location and location is now looking once again at formation manoeuvres"
	"USS Freedom is currently forward deployed to Singapore with a modified SUW mission package embarked."	"Warship is currently forward deployed to location with a suw mission package embarked"
	"The future HMS Prince of Wales puts to sea for the first time from Rosyth, Scotland, on 19 September."	"the future aircraft carrier puts to sea for the time from location on date"

Figure 20: **Captions processing:** To reduce the amount of unique words in the captions, we performed some pre-processing of the captions. We have visualized the effect of this cleaning on 5 captions. As can be seen, some captions have become more generic and got a better alignment with the image. Especially transforming named entities to their hypernym has a very notable effect on the captions.

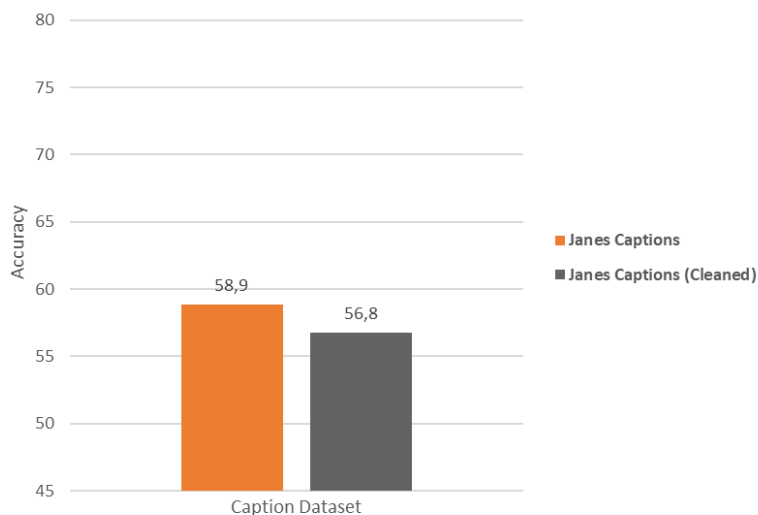


Figure 21: **Janes Captions versus Cleaned Captions versus Increased Token Vocabulary:** Pre-training using image captioning on the “standard” Janes Captions (58,9%) dataset outperforms pre-training using a cleaned version of the Janes captions (56,8%) and pre-training using an increased token vocabulary size (57,1%).

4 Discussion

Like various other domains, the military intelligence domain is investigating the use of deep learning methods to automate specific processes. One suitable task for automation through deep learning is the recognition of very specific entities in large sets of images. Using carefully collected and manually annotated datasets, deep learning methods have found ways to learn visual representations of images that can be used to achieve high performances on various vision tasks, such as image classification and object detection. However, collecting large scale annotated datasets is far from trivial. In fact, it is labour-intensive and often requires expert knowledge. As a result, these datasets often concern very generic topics. This causes limitations when applying models trained on these generic datasets to vision tasks in more specific domains since the learned features are less generalizable to the highly specific class taxonomy present in, for example, the military intelligence domain. Recent developments in vision-language modelling have led to using natural language supervision to pre-train vision models, which requires less careful annotation efforts. In line with these advancements, this project aimed to investigate the use of natural language supervision for domain-specific visual representation learning. Specifically, we investigated using a domain-specific image-caption dataset to learn domain-specific visual features, using three different vision-language pre-training tasks. Besides evaluating on a domain-specific image classification task, we compared our methods against two benchmarks.

4.1 Vision-language pre-training datasets and tasks

The first experiment hypothesized that image captioning pre-training using a domain-specific dataset would lead to domain-specific visual representations, leading to better performances on the domain-specific image classification task. However, we found that pre-training using the Janes Captions dataset did not directly lead to better performances when compared to the full COCO Captions dataset.

However, the Janes Captions dataset differs from the COCO captions not only in domain coverage but also in size. Accordingly, we performed an additional analysis in which the model was pre-trained using only 20% of the COCO captions dataset, resulting in two similarly-sized image-caption datasets. This analysis showed that pre-training on the Janes captions outperformed pre-training on the reduced COCO captions dataset, indicating that domain-specific datasets can increase the learned visual features, given that the generic dataset is of equal size. Showing the added benefit of using in-domain captions indicates that it could be a good starting point for pre-training compared to a generic dataset. Additionally, this result emphasizes one of the main problems with current academic research that often opts for generic image-captions datasets and tasks [13,62,75], not allowing for a comparison of domain-specific datasets and tasks. As an example for another domain, in the study by Zhang and colleagues, a multimodal contrastive training for medical domain visual representations was used, indicating that a medical-domain dataset improved the results compared to pre-training using generic datasets [77]. Future research could investigate whether the observed difference in performance could increase even further when a strongly-aligned domain-specific dataset is used.

In the second experiment, it was shown that pre-training tasks which require the model to predict the complete caption, including its sequential structure, outperform pre-training tasks, which only required the prediction of the semantics of the captions. The effect of these different pre-training tasks was evaluated on an image classification task. Since image classification only requires the visual backbone to label the image and does not require more

complex operations, it could be that for other downstream vision tasks, the visual backbone could benefit from other pre-training tasks. Recent studies have already investigated how different image captioning pre-training tasks affected the performance on object-detection compared to image classification [13,62,74], showing that generic image classification and object-detection benefit from pre-training tasks differently. Future research should try to extend the scope of evaluation to other visual tasks, such as military object detection.

Additional analysis also showed that by increasing the complexity of the textual head, the performance on the image classification task increased. A difference between the pre-training tasks concerns the complexity of the textual head. The architecture pre-trained on image captioning, which achieved the highest performance on the image classification task, has both a forward as well as a backward transformer component in the textual heads, while the architecture pre-trained on ICMLM, which came in second, only consisted of one transformer component in its textual head. The Token-Classification task achieved the lowest performance with only a fully connected layer as its textual head. Research into natural language processing indicated that more complex language models with larger transformers learn better textual features [3,14,48]. Studies investigating natural language supervision for visual representation learning also indicated a performance difference when comparing pre-training tasks [13,62]. However, these studies claimed that the difference in performance was due to the model needing to learn the sequential structure of the captions, while the performance differences (largely) followed the complexity of the textual head. In combination with the results of our study, these findings indicate that the more complex supervisory signal originating from a more complex textual head can be beneficial for the visual backbone's training, no matter the implemented pre-training task. We hypothesize that with a complex textual head, the visual backbone can restrict itself to representing more generalizable features, which the complex textual head can properly reason over.

The relationship between Image captioning performance and image classification performance was also investigated. During the pre-training, It was observed that our architecture quickly overfitted on the training split of the Janes captions dataset. However, the performance of the visual backbone on the image classification was not affected by this overfitting. This indicates that the textual head is overfitting, while the visual backbone is slightly improving and not affected in learning generalizable domain-specific visual representations. A multimodal multitask model implemented by Hu & Singh [32] showed that it could achieve high performance on both single modal and multimodal model tasks at the same time by implementing a training scheme that simultaneously learns seven different tasks. They also showed that the multimodal tasks of VQA and visual entailment benefit from being jointly trained on uni-modal tasks such as object detection and sentiment analysis. Future research into visual representation learning through natural language supervision should investigate using a more complex training scheme instead of focusing on one multimodal task during pre-training and one vision task during evaluation.

Finally, we compared these achieved performances against two benchmarks, which implemented state-of-the-art pre-training methods: supervised and unsupervised pre-training. These pre-training methods outperformed pre-training on image captioning, using either the domain-specific or generic image caption datasets. These benchmarks performed their task only on the visual modality, not using any information from the textual modality. This allows these models to pre-train using very large datasets. The supervised and unsupervised methods were pre-trained on 1.24M images, whereas our dataset contained only 25K images, almost 50 times smaller. This

difference could explain the performance gap between these datasets. Using a generic image-caption dataset, researchers compared supervised image classification and image captioning pre-training using a similar number of samples [13]. They found that a similar number of samples during pre-training led to higher performance when pre-trained using image captioning than pre-training using image classification. Increasing the number of captions per image increased the performance even further. We hypothesize that extending the Janes caption dataset with more images and captions is crucial for future studies to learn better domain-specific visual representations.

4.2 Visual modality in multimodal transformers

The architecture implemented in this study implemented a multimodal transformer to create a supervisory learning signal for a visual backbone to learn visual representation through natural language. Although we showed that domain-specific representations could be learned using natural language supervision, much is still unknown about the interaction of modalities in multimodal transformers. Recent research has already hinted at problems that can arise in multimodal transformers. These studies have indicated problems that concern (1) the bias in vision-language datasets, (2) the modality interaction in the multimodal transformer, and (3) the use of multimodal losses.

4.2.1 Bias in datasets

Firstly, studies investigating image captioning and VQA found that complex multimodal models could be outperformed by more straightforward methods that take advantage of the implicit structure and bias present in our language, and therefore also in vision-language datasets, such as the COCO captions dataset.

A study into image captioning was able to outperform, at the time, complex methods that generated novel captions by taking advantage of the typical structure and vocabulary in image-caption datasets [15]. Instead of generating a unique caption, their method identified the K most similar images, given a queried image, and copied a caption from one of these images. They showed that this approach benefits from the bias towards generic vocabulary and structure present in the caption datasets. The researchers indicated that these approaches would become less beneficial when applied to datasets with more diversity or capture more rare occurrences.

Another study that highlighted similar problems with the vision-language dataset investigated a popular VQA dataset [27]. They highlighted that the dataset contains various biases, such as specific answers being present a lot more than others ('tennis' is the correct answer on 41% of the questions starting with 'What sport is'), and that people will often mention an object that is present in the image ('yes' is the correct answer to 87% of the questions starting with 'Do you see a ..'). The researchers implemented a new, more balanced version of the VQA dataset. They showed that existing state-of-the-art VQA models performed poorly on their more balanced dataset, indicating that these models exploited the bias present in the language modality to achieve higher accuracy.

A final study showed that pre-trained vision-language models are prone to abuse the bias towards quantities in the available datasets [56]. They investigated this by asking the model to judge whether a question or statement about the number of visible entities in the image is correct, which is challenging for VL models. They found that the models performed sub-optimal on the counting task, again exploiting the bias concerning quantities present in the VL datasets.

These studies have highlighted that the language modality in vision-language datasets can provide a strong signal which can be used to get a good, albeit superficial, performance. In turn, this can result in models ignoring visual

information and researchers “claiming” multimodal capability of models that is not present. Since our study implemented image captioning using a new dataset, our architecture could have taken advantage of the bias present in our dataset to perform well at image captioning, limiting the pre-training of the visual backbone. Since we only acquired our captions from one source, the captions could show a bias towards a specific vocabulary, referring to specific entities with the same vocabulary and structure since all the captions were created for use in a magazine. Future work could investigate the amount of bias present in the Janes captions dataset and, if necessary, implement adaptations to diversify the dataset further.

4.2.2 *Modality interaction in multimodal transformers*

More recently, multimodal transformers have been used in almost all pre-trained vision-language models, achieving state-of-the-art performance on multiple vision-language tasks. Suggesting that these multimodal transformers can obtain accurate multimodal representations that encode a substantial amount of both the visual and textual knowledge from each modality. Because of this success, various studies investigated pre-trained vision-language models that implemented multimodal transformers, revealing various problems with the multimodal representations.

One study designed a set of tasks to reveal the workings of the inner mechanisms of two popular vision-language models, LXMERT and UNITER [4]. These so-called probing tasks were designed to provide insights into various aspects of multimodality, such as modality importance and cross-modal interaction. They showed that the textual modality plays a more critical role than the visual modality in making the final predictions. Moreover, by investigating the cross-modal and single-modal representations through different single-modal and multimodal probing tasks, they also showed that the pre-trained vision-language models are much better at encoding linguistic than visual information.

Another study proposed a different method to investigate how pre-trained vision-language models use cross-modal information for their final prediction [21]. They showed that although the models have learned to fuse information of both modalities, the cross-modal interaction is not symmetrical. This means that the visual representations are much more influenced by text representations than vice versa.

These studies reveal that, contrary to popular belief, multimodal transformers experience problems integrating the textual and visual modality. The multimodal representations show an asymmetric preference on both modalities, leaning towards the textual modality, which affects both the final decision and the influence both modalities have on each other throughout the transformer. Since our study also implemented the multimodal transformer, our results could be affected by this. These findings could explain why overfitting on the image captioning task did not affect the performance on the image classification task. The multimodal transformer does not require specific features from the visual modality, resulting in a less specific learning signal propagated back through the visual backbone and a visual backbone that focuses on learning more generic and generalizable features.

4.2.3 *Optimizing on multimodal losses*

A final problem discussed in research concerns the specific loss functions used for pre-training vision-language models and multimodal transformers architectures. A study that investigated the use of different loss functions, one for the textual and visual modality separately or a combined loss, found that the absence of a visual loss does

not harm the performance on vision-language tasks, indicating that these models are not tapping into the useful signal in the visual modality [31].

Our study implemented a textual loss during the evaluation phase, which only computed the loss as a function of the correctness of the predicted captions. It could be that the visual backbone could have benefitted from a visual loss during pre-training, such as Masked Region Modelling loss. The visual backbone was purely trained using a learning signal based on the textual modality, which might have limited the learning of visual features by the backbone. Future research could focus on implementing multimodal losses and image losses during pre-training.

The studies discussed above show that a lot remains unsure about the role of the visual modality in multimodal transformers. Future work should consider the bias present in vision-language datasets, the modality interaction in the transformer component, and the use of other losses during pre-training. Especially the interaction between these three issues should be taken into account. The tendency of a vision-language model to exploit the bias in language could perhaps be overcome by using a strict multimodal loss. Alternatively, using only a textual loss, such as the case in our study, would benefit by improved integration of the different modalities in the multimodal transformer.

4.3 Conclusion

This study showed the feasibility of learning domain-specific visual features through natural language learning. We improved the performance on a domain-specific image classification task by pre-training on image captioning with a domain-specific image-caption dataset, compared to a similar-sized generic image-caption dataset. We could not outperform current state-of-the-art methods using large labelled datasets or image captioning methods with an increased amount of generic captions. The results of our study, combined with other studies into natural language supervision and multimodal transformers, have hinted at potential solutions that could increase the learning of domain-specific visual features, such as increased and improved domain-specific image-caption datasets and improvements in multimodal transformers. Ultimately, natural language supervision for pre-training visual models is a promising concept that, if applied correctly, could solve the problems of current state-of-the-art methods, especially for application in specific domains.

References

- [1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. *VQA: Visual question answering*. DOI:<https://doi.org/10.1109/ICCV.2015.279>
- [2] Philip Bachman, R. Devon Hjelm, and William Buchwalter. 2019. Learning representations by maximizing mutual information across views. *Adv. Neural Inf. Process. Syst.* 32, (2019). Retrieved January 12, 2022 from <https://github.com/Philip-Bachman/amdim-public>.
- [3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* 2020-Decem, (2020).
- [4] Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen Chun Chen, and Jingjing Liu. 2020. Behind the Scene: Revealing the Secrets of Pre-trained Vision-and-Language Models. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* 12351 LNCS, (2020), 565–580. DOI:https://doi.org/10.1007/978-3-030-58539-6_34
- [5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Adv. Neural Inf. Process. Syst.* 2020-Decem, (2020). Retrieved June 24, 2021 from <https://github.com/facebookresearch/swav>
- [6] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts. (2021), 3557–3567. DOI:<https://doi.org/10.1109/cvpr46437.2021.00356>
- [7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *37th International Conference on Machine Learning, ICML 2020*, International Machine Learning Society (IMLS), 1575–1585. Retrieved July 5, 2021 from <http://arxiv.org/abs/2002.05709>
- [8] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. 2020. Big self-supervised models are strong semi-supervised learners. *Adv. Neural Inf. Process. Syst.* 2020-Decem, (2020). Retrieved December 21, 2021 from <https://github.com/google-research/simclr>.
- [9] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. 2015. Microsoft COCO Captions: Data Collection and Evaluation Server. (2015). Retrieved June 24, 2021 from <http://arxiv.org/abs/1504.00325>
- [10] Yen Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: UNiversal Image-TExt Representation Learning. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 104–120. DOI:https://doi.org/10.1007/978-3-030-58577-8_7
- [11] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying Vision-and-Language Tasks via Text Generation. (2021). Retrieved May 17, 2021 from <http://arxiv.org/abs/2102.02779>
- [12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2010. ImageNet: A large-scale hierarchical image database. *Institute of Electrical and Electronics Engineers (IEEE)*, 248–255. DOI:<https://doi.org/10.1109/cvpr.2009.5206848>
- [13] Karan Desai and Justin Johnson. 2021. VirTex: Learning Visual Representations from Textual Annotations. (2021), 11157–11168. DOI:<https://doi.org/10.1109/cvpr46437.2021.01101>
- [14] Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 4171–4186. Retrieved May 5, 2021 from <https://github.com/tensorflow/tensor2tensor>
- [15] Jacob Devlin, Saurabh Gupta, Ross Girshick, Margaret Mitchell, and C. Lawrence Zitnick. 2015. Exploring Nearest Neighbor Approaches for Image Captioning. (2015). Retrieved December 7, 2021 from <http://arxiv.org/abs/1505.04467>
- [16] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. 2015. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, 1422–1430. DOI:<https://doi.org/10.1109/ICCV.2015.167>
- [17] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. 2014. DeCAF: A deep convolutional activation feature for generic visual recognition. In *31st International Conference on Machine Learning, ICML 2014*, 988–996. Retrieved June 24, 2021 from <https://github.com/>
- [18] Ye Du, Zehua Fu, Qingjie Liu, and Yunhong Wang. 2021. Visual Grounding with Transformers. (May 2021). Retrieved June 4, 2021 from <http://arxiv.org/abs/2105.04281>

- [19] Image-text Embeddings and Liwei Wang. 2016. *Learning Deep Structure-Preserving*.
- [20] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2019. VSE++: Improving visual-semantic embeddings with hard negatives. In *British Machine Vision Conference 2018, BMVC 2018*. Retrieved June 30, 2021 from <https://github.com/fartashf/vsepp>
- [21] Stella Frank, Emanuele Bugliarello, and Desmond Elliott. 2021. Vision-and-Language or Vision-for-Language? On Cross-Modal Influence in Multimodal Transformers. (2021). DOI:<https://doi.org/10.18653/v1/2021.emnlp-main.775>
- [22] Andrea Frome, Greg S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc’auelio Ranzato, and Tomas Mikolov. 2013. DeViSE: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*.
- [23] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. 2018. Unsupervised representation learning by predicting image rotations. In *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*. Retrieved June 24, 2021 from <https://github.com/gidariss/FeatureLearningRotNet>.
- [24] Lluís Gomez, Yash Patel, Marçal Rusinol, Dimosthenis Karatzas, and C. V. Jawahar. 2017. Self-supervised learning of visual features through embedding images into text topic spaces. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017–2026*. DOI:<https://doi.org/10.1109/CVPR.2017.218>
- [25] Raul Gomez, Lluís Gomez, Jaume Gibert, and Dimosthenis Karatzas. 2019. Self-supervised learning from web data for multimodal retrieval. In *Multimodal Scene Understanding: Algorithms, Applications and Deep Learning*. 279–306. DOI:<https://doi.org/10.1016/B978-0-12-817358-9.00015-9>
- [26] Yash Goyal, Tejas Khot, Aishwarya Agrawal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2019. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. *Int. J. Comput. Vis.* 127, 4 (2019), 398–414. DOI:<https://doi.org/10.1007/s11263-018-1116-0>
- [27] Yash Goyal, Tejas Khot, Aishwarya Agrawal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2019. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. *Int. J. Comput. Vis.* 127, 4 (2019), 398–414. DOI:<https://doi.org/10.1007/s11263-018-1116-0>
- [28] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 9726–9735. DOI:<https://doi.org/10.1109/CVPR42600.2020.00975>
- [29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 770–778. DOI:<https://doi.org/10.1109/CVPR.2016.90>
- [30] Olivier J. Henaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aaron Van den Oord Eslami. 2020. Data-Efficient image recognition with contrastive predictive coding. In *37th International Conference on Machine Learning, ICML 2020*, 4130–4140.
- [31] Lisa Anne Hendricks, John Mellor, Rosalia Schneider, Jean Baptiste Alayrac, and Aida Nematzadeh. 2021. Decoupling the role of data, attention, and losses in multimodal transformers. *Trans. Assoc. Comput. Linguist.* 9, (2021), 570–585. DOI:<https://doi.org/10.1162/tacl.00385>
- [32] Ronghang Hu and Amanpreet Singh. 2021. *UniT: Multimodal Multitask Learning with a Unified Transformer*. Retrieved May 7, 2021 from <http://arxiv.org/abs/2102.10772>
- [33] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. 2020. Pixel-BERT: Aligning Image Pixels with Text by Deep Multi-Modal Transformers. (April 2020). Retrieved June 4, 2021 from <http://arxiv.org/abs/2004.00849>
- [34] Yuqi Huo, Manli Zhang, Guangzhen Liu, Haoyu Lu, Yizhao Gao, Guoxing Yang, Jingyuan Wen, Heng Zhang, Baogui Xu, Weihao Zheng, Zongzheng Xi, Yueqian Yang, Anwen Hu, Jinming Zhao, Ruichen Li, Yida Zhao, Liang Zhang, Yuqing Song, Xin Hong, Wanqing Cui, Danyang Hou, Yingyan Li, Junyi Li, Peiyu Liu, Zheng Gong, Chuhao Jin, Yuchong Sun, Shizhe Chen, Zhiwu Lu, Zhicheng Dou, Qin Jin, Yanyan Lan, Wayne Xin Zhao, Ruihua Song, and Ji-Rong Wen. 2021. WenLan: Bridging Vision and Language by Large-Scale Multi-Modal Pre-Training. (2021). Retrieved May 7, 2021 from <http://arxiv.org/abs/2103.06561>
- [35] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. (2021), 139. Retrieved June 30, 2021 from <http://arxiv.org/abs/2102.05918>
- [36] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. 2019. *Supervised Multimodal Bitransformers for Classifying Images and Text*. Retrieved June 1, 2021 from <http://arxiv.org/abs/1909.02950>
- [37] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. 2020. Big Transfer (BiT): General Visual Representation Learning. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 491–507. DOI:https://doi.org/10.1007/978-3-030-58558-7_29
- [38] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual Genome: Connecting

- Language and Vision Using Crowdsourced Dense Image Annotations. *Int. J. Comput. Vis.* 123, 1 (2017), 32–73. DOI:<https://doi.org/10.1007/s11263-016-0981-7>
- [39] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (2017), 84–90. DOI:<https://doi.org/10.1145/3065386>
- [40] Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *EMNLP 2018 - Conf. Empir. Methods Nat. Lang. Process. Syst. Demonstr. Proc.* (2018), 66–71. DOI:<https://doi.org/10.18653/v1/d18-2012>
- [41] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. 2020. The Open Images Dataset V4: Unified Image Classification, Object Detection, and Visual Relationship Detection at Scale. *Int. J. Comput. Vis.* 128, 7 (2020), 1956–1981. DOI:<https://doi.org/10.1007/s11263-020-01316-z>
- [42] Kuang Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked Cross Attention for Image-Text Matching. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 212–228. DOI:https://doi.org/10.1007/978-3-030-01225-0_13
- [43] Ang Li, Allan Jabri, Armand Joulin, and Laurens Van Der Maaten. 2017. Learning Visual N-Grams from Web Data. In *Proceedings of the IEEE International Conference on Computer Vision*, 4193–4202. DOI:<https://doi.org/10.1109/ICCV.2017.449>
- [44] Chenliang Li, Ming Yan, Haiyang Xu, Fuli Luo, Wei Wang, Bin Bi, and Songfang Huang. 2021. SemVLP: Vision-Language Pre-training by Aligning Semantics at Multiple Levels. (2021). Retrieved June 29, 2021 from <http://arxiv.org/abs/2103.07829>
- [45] Junnan Li, Pan Zhou, Caiming Xiong, and Steven C. H. Hoi. 2020. Prototypical Contrastive Learning of Unsupervised Representations. (2020). Retrieved June 24, 2021 from <http://arxiv.org/abs/2005.04966>
- [46] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. *VisualBERT: A Simple and Performant Baseline for Vision and Language*. Retrieved June 1, 2021 from <http://arxiv.org/abs/1908.03557>
- [47] Junyang Lin, An Yang, Yichang Zhang, Jie Liu, Jingren Zhou, and Hongxia Yang. 2020. InterBERT: Vision-and-Language Interaction for Multi-modal Pretraining. (March 2020). Retrieved June 30, 2021 from <http://arxiv.org/abs/2003.13198>
- [48] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. (2019). Retrieved December 21, 2021 from <http://arxiv.org/abs/1907.11692>
- [49] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems* 32.
- [50] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 10434–10443. DOI:<https://doi.org/10.1109/CVPR42600.2020.01045>
- [51] Fuli Luo, Pengcheng Yang, Shicheng Li, Xuancheng Ren, and Xu Sun. 2020. CAPT: Contrastive Pre-Training for Learning Denoised Sequence Representations. (2020). Retrieved June 29, 2021 from <http://arxiv.org/abs/2010.06351>
- [52] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. 2020. UniVL: A Unified Video and Language Pre-Training Model for Multimodal Understanding and Generation. *arXiv*. Retrieved May 17, 2021 from <http://arxiv.org/abs/2002.06353>
- [53] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. 2018. Exploring the Limits of Weakly Supervised Pretraining. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 185–201. DOI:https://doi.org/10.1007/978-3-030-01216-8_12
- [54] Jong Hak Moon, Hyungyung Lee, Woncheol Shin, and Edward Choi. 2021. Multi-modal Understanding and Generation for Medical Images and Text via Vision-Language Pre-Training. (2021). Retrieved June 4, 2021 from <http://arxiv.org/abs/2105.11333>
- [55] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation Learning with Contrastive Predictive Coding. (2018). Retrieved June 24, 2021 from <http://arxiv.org/abs/1807.03748>
- [56] Letitia Parcalabescu, Albert Gatt, Anette Frank, and Iacer Calixto. 2020. Seeing past words: Testing the cross-modal capabilities of pretrained V&L models on counting tasks. (December 2020). Retrieved December 2, 2021 from <http://arxiv.org/abs/2012.12352>
- [57] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. 2012. Cats and dogs. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 3498–3505. DOI:<https://doi.org/10.1109/CVPR.2012.6248092>
- [58] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury Google, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf Xamla, Edward Yang, Zach Devito, Martin Raison Nabla, Alykhan Tejani, Sasank Chilamkurthy, Qure Ai, Benoit Steiner, Lu Fang Facebook, Junjie Bai

- Facebook, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. (2019).
- [59] Ariadna Quattoni, Michael Collins, and Trevor Darrell. 2007. Learning visual representations using images with captions. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. DOI:<https://doi.org/10.1109/CVPR.2007.383173>
- [60] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. CNN features off-the-shelf: An astounding baseline for recognition. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 512–519. DOI:<https://doi.org/10.1109/CVPRW.2014.131>
- [61] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* 115, 3 (2015), 211–252. DOI:<https://doi.org/10.1007/s11263-015-0816-y>
- [62] Mert Bulent Sariyildiz, Julien Perez, and Diane Larlus. 2020. Learning Visual Representations with Caption Annotations. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Science and Business Media Deutschland GmbH, 153–170. DOI:https://doi.org/10.1007/978-3-030-58598-3_10
- [63] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypemymed, image alt-text dataset for automatic image captioning. In *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, Association for Computational Linguistics, 2556–2565. DOI:<https://doi.org/10.18653/v1/p18-1238>
- [64] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2017. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. In *Proceedings of the IEEE International Conference on Computer Vision*, 843–852. DOI:<https://doi.org/10.1109/ICCV.2017.97>
- [65] Hao Tan and Mohit Bansal. 2020. LXMert: Learning cross-modality encoder representations from transformers. In *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 5100–5111. DOI:<https://doi.org/10.18653/v1/d19-1514>
- [66] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive Multiview Coding. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 776–794. DOI:https://doi.org/10.1007/978-3-030-58621-8_45
- [67] Trieu H. Trinh, Minh-Thang Luong, and Quoc V. Le. 2019. Selfie: Self-supervised Pretraining for Image Embedding. (2019). Retrieved June 24, 2021 from <http://arxiv.org/abs/1906.02940>
- [68] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. *The Caltech-UCSD Birds-200-2011 Dataset*. Retrieved June 24, 2021 from <http://www.birdfieldguide.com>
- [69] Jianfeng Wang, Xiaowei Hu, Pengchuan Zhang, Xiujun Li, Lijuan Wang, Lei Zhang, Jianfeng Gao, and Zicheng Liu. 2020. MiniVLM: A Smaller and Faster Vision-Language Model. (December 2020). Retrieved June 4, 2021 from <http://arxiv.org/abs/2012.06946>
- [70] Qizhe Xie, Minh Thang Luong, Eduard Hovy, and Quoc V. Le. 2020. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 10684–10695. DOI:<https://doi.org/10.1109/CVPR42600.2020.01070>
- [71] I. Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. 2019. Billion-scale semi-supervised learning for image classification. (2019). Retrieved June 24, 2021 from <http://arxiv.org/abs/1905.00546>
- [72] Mang Ye, Xu Zhang, Pong C. Yuen, and Shih Fu Chang. 2019. Unsupervised embedding learning via invariant and spreading instance feature. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 6203–6212. DOI:<https://doi.org/10.1109/CVPR.2019.00637>
- [73] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguist.* 2, (2014), 67–78. DOI:https://doi.org/10.1162/tacl_a_00166
- [74] Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. 2021. Multimodal Contrastive Training for Visual Representation Learning. (2021), 6991–7000. DOI:<https://doi.org/10.1109/cvpr46437.2021.00692>
- [75] Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. 2021. Multimodal Contrastive Training for Visual Representation Learning. (April 2021), 6991–7000. DOI:<https://doi.org/10.1109/cvpr46437.2021.00692>
- [76] Richard Zhang, Phillip Isola, and Alexei A. Efros. 2016. Colorful image colorization. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 649–666. DOI:https://doi.org/10.1007/978-3-319-46487-9_40
- [77] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. 2020. Contrastive

- Learning of Medical Visual Representations from Paired Images and Text. (2020). Retrieved June 10, 2021 from <http://arxiv.org/abs/2010.00747>
- [78] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7W: Grounded question answering in images. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 4995–5004. DOI:<https://doi.org/10.1109/CVPR.2016.540>
- [79] Chengxu Zhuang, Alex Zhai, and Daniel Yamins. 2019. Local aggregation for unsupervised learning of visual embeddings. In *Proceedings of the IEEE International Conference on Computer Vision*, 6001–6011. DOI:<https://doi.org/10.1109/ICCV.2019.00610>