applied sciences

MDPI

Article

Incorporating Concreteness in Multi-Modal Language Models with Curriculum Learning

Erhan Sezerer and Selma Tekir *

Department of Computer Engineering, Izmir Institute of Technology, 35430 Izmir, Turkey; erhansezerer@iyte.edu.tr

* Correspondence: selmatekir@iyte.edu.tr

Abstract: Over the last few years, there has been an increase in the studies that consider experiential (visual) information by building multi-modal language models and representations. It is shown by several studies that language acquisition in humans starts with learning concrete concepts through images and then continues with learning abstract ideas through the text. In this work, the curriculum learning method is used to teach the model concrete/abstract concepts through images and their corresponding captions to accomplish multi-modal language modeling/representation. We use the BERT and Resnet-152 models on each modality and combine them using attentive pooling to perform pre-training on the newly constructed dataset, which is collected from the Wikimedia Commons based on concrete/abstract words. To show the performance of the proposed model, downstream tasks and ablation studies are performed. The contribution of this work is two-fold: A new dataset is constructed from Wikimedia Commons based on concrete/abstract words, and a new multi-modal pre-training approach based on curriculum learning is proposed. The results show that the proposed multi-modal pre-training approach contributes to the success of the model.

Keywords: multi-modal dataset; Wikimedia Commons; multi-modal language model; concreteness; curriculum learning

check for updates

Citation: Sezerer, E.; Tekir, S. Incorporating Concreteness in Multi-Modal Language Models with Curriculum Learning. *Appl. Sci.* **2021**, *11*, 8241. https://doi.org/10.3390/app11178241

Academic Editors: Emanuele Carpanzano, Arturo Montejo-Ráez and Salud María Jiménez-Zafra

Received: 1 August 2021
Accepted: 27 August 2021
Published: 6 September 2021

1. Introduction

After the success of contextual representations, language model pre-training and fine-tuning the model for downstream tasks have been common practices in natural language processing (NLP) . The wide-spread adoption of BERT [1] led to several pre-trained language models that are described as BERT variants [2–5]. Putting BERT at the core, these models provide extensions with different viewpoints, cross-lingual, multi-task, multi-modal, and world knowledge, to name a few. Among these models, Albert [3] targets efficiency by using weight sharing and decreasing memory consumption, RoBERTa [2] increases the amount of training data and times and removes the next sentence prediction objective, XLNet [4] uses permutation instead of masking to capture the bidirectional context and combines BERT with autoregressive language modeling, and ERNIE [5] aims to exploit world knowledge by masking named entities and phrases rather than random words, and, in its updated version [6], the pre-training task is organized as a multi-task objective to capture different relations, such as lexical, syntactic, and semantic.

The earlier approaches to bridge vision and language relied on architectures with a visual feature extractor, a text encoder, a multi-modal fusion component, and a classification layer to perform the given multi-modal task, e.g., visual question answering. The robust pre-trained language models have caused a shift from a task-specific perspective to a task-agnostic one, multi-modal language model pre-training.

Multi-modality, especially with vision and language, has been implemented in some BERT variants [7–9], as well. VisualBERT [7] and VideoBERT [8] use similar transformer-based architectures. The former processes image captions together with image regions to discover implicit alignments between language and vision. On the other hand, the latter works with spoken words paired with a series of images to learn a similar alignment. Distinctively, ViLBERT [9] has a two-stream transformer model, which processes vision and language separately but learns their relationships through co-attentions between them.

The primary motivation for combining vision and language in these models has been visual grounding to learn visual features under the guidance of textual descriptions. Apart from it, we can leverage visual and language features to mimic human language acquisition.

There have been studies that indicate we can mainly attribute language acquisition in children to experiential information in early ages [10–12]. It is mentioned in those works that the language acquisition in children starts with experiential information, where we mostly learn about concrete concepts in languages and continue with the textual information in later ages where we mostly know about abstract concepts. Thus, many researchers tried to build language models with multi-modal information (Refs. [9,13,14], and many more), leveraging both textual and visual inputs.

This work aims to create a multi-modal language model that uses both textual and visual features, similar to what humans do. First, we feed the image model concrete examples. Then, we train the textual model with all of the samples concrete and abstract combined, in a curriculum learning fashion [15,16]. We rely on University of Western Australia The Medical Research Council (UWA MRC) Psycholinguistic Dataset [17] for the lists of the abstract/concrete words. The contribution of this work is two-fold: A new dataset is constructed from Wikimedia Commons based on concrete/abstract terms, and a new multi-modal pre-training approach that is based on curriculum learning [15,16] is proposed.

The results show that the proposed multi-modal pre-training method contributes to the success of the model in downstream tasks, e.g., visual question answering. In addition, it can be seen from the ablation study that this increase in performance is consistent among all fusion techniques used in this work. We obtained the best results when the multi-modal pre-training scheme is used with attentive pooling as the fusion mechanism. In addition to the tests mentioned above, we performed several tests for measuring the informativeness of the newly constructed dataset.

The rest of the manuscript is structured as follows: In Section 2, we give background information on the task of language modeling/representation. Model details and the new dataset are explained in Section 3. We share the experimental results in Section 4, along with the descriptions of the datasets used. In addition, finally, in Section 5, final remarks are made with possible future directions.

2. Related Work

The idea of building word representations from frequency statistics comes from the Distributional Hypothesis [18,19]. The distributional hypothesis states that one can determine the meaning of a word through the words that co-occur with it in the same context. Famously, Harris (1954 [19]) states that the “words that occur in the same context tend to have similar meanings”.

Although the count-based methods can leverage the distributional model to learn the representations of words, they suffer from several drawbacks: lack of word order, unable to retrieve representations from partial information (generalization power), and the curse of dimensionality (they create millions, if not trillions, of different possible n-grams which are very unlikely to be observed in the training data, which leads to a very sparse matrix with a lot of uninformative zero entries).

Neural network solutions emerged to solve these issues. In such a first attempt, Hinton et al., in 1986 [20], utilized the idea of distributed representations for concepts. They proposed to use patterns of hidden layer activations (which are only allowed to be 0 or 1) as the representation of meanings instead of representing words with discrete entities, such as the number of occurrences, together. They argued that the most critical evidence of distributed representations is their degree of similarity to the weaknesses and strengths of the human mind.

Elman (1990) [21] was the first to implement the distributional model proposed by Reference [20] in a language model. He presents a specific recurrent neural network structure with memory, called the Elman network, to predict bits in temporal sequences. Memory is provided to the network through context units that are fully connected with hidden units.

Although these models build the basis of neural word representations, Bengio et al., in 2003 [22], popularized the distributional representation idea by realizing it through a language model and lead to numerous other studies that are built on it. Their model architecture uses a feed-forward network with a single hidden layer and optional direct connections from the input layer to the softmax layer. The weights of the hidden layer are then taken as the representations of words.

Once it is shown that neural language models are efficiently computable by Bengio et al., as in 2003 [22], newer language models, along with better word embeddings, are developed successively. In such an effort, Mikolov et al., in 2013 [23], proposed word2vec to learn high-quality word vectors. The authors removed the non-linearity in the hidden layer in the proposed model architecture of Bengio et al., in 2003 [22], to gain an advantage in computational complexity. Due to this change, the system can be trained using billions of words efficiently. Thus, it is considered as the initiator of early word embeddings [24].

Despite the success of these earlier word embeddings, there were still many limitations in terms of the accuracy of representations (lack of polysemy, unable to account for morphology, antonymy/synonymy problem). Many methods have been proposed for solving the deficiencies of embedding methods. Each of them is specialized on a single problem, such as sense representations [25,26], morpheme representations [27,28], etc., while none of them could combine different aspects into a single model, a single solution. It is the idea of contextual representations to provide a solution that covers each element successfully. The main idea behind contextual representations is that words should not have a single representation to be used in every context. Instead, one should calculate a representation separately for different contexts. Contextual representation methods calculate the embedding of a word from the surrounding words each time the word is seen. This characteristic leads to an implicit solution to many problems, such as sense representations, since multi-sense words can now have different representations according to their contexts. Furthermore, character-level processing has been proposed to incorporate the sub-word information into embeddings. Therefore, contextual representation models described below can incorporate different aspects together into a single model.

In such a first attempt to create contextual representations, Melamud et al., in 2016 [29], developed a neural network architecture based on bidirectional-LSTMs to learn context embeddings with the target word embeddings jointly. CoVe [30] uses Glove [24] as the initial word embeddings and feeds them into a machine translation architecture to learn contextual representations. The authors argue that pre-training the contextual representations on machine learning tasks, where there are vast amounts of data, can lead to better contextual representations to transfer learning to other downstream tasks. Using language modeling and learning word representations as a pre-training objective then fine-tuning the architecture to downstream tasks is first proposed by References [31,32]. ELMO [33] improves on the character-aware neural language model by Reference [34]. The architecture takes characters as input to a CNN network from where it is fed to a 2-layer bidirectional-LSTM network to predict a target word. They show that this architecture can learn various aspects of semantic, syntactic, and sub-word information. Instead of using words as input, Flair [35] uses a character-level language model to learn contextual word representations. Unlike ELMO, where character-level inputs are later converted into word features, authors propose using characters only in this work. BERT [1] uses a bidirectional transformer [36] architecture to learn contextual word representations. XLNet [4] is an autoregressive method that combines the advantages of two language modeling methods:

Autoregressive models (i.e., transformer-XL [37]) and autoencoder models (i.e., BERT). ALBERT [3] aims at lowering the memory consumption and training times of BERT [1]. To accomplish this, they perform two changes on the original BERT model: They factorize the embeddings into two matrices to use smaller dimensions, and they apply weight sharing to decrease the number of parameters.

The success of uni-modal language models drives the researchers into studies that examine the use of visual information for training language models. They base this decision on the advances in cognitive science where it is shown that language acquisition in children mostly relies on experiential data [10–12]. While some of those studies focused on producing better representations, [12,38–42], most of these models produce multi-modal embeddings as a side-product of a multi-modal task. These tasks include image retrieval with text and caption [43,44], image-text alignment [45,46], image segmentation using a target text [47], visual question answering [13,14,48], visual common-sense reasoning [49], and image captioning [42]. Some other studies also contributed to the field of multi-modal language modeling by encompassing many of these models similar to contextual embeddings [9] or by enhancing the existing models [50]. As the field is relatively new, most of these works focus on the fusion of modalities more than the individual models.

Curriculum learning [15,16] used in this study is a progressive training method that puts the samples in a meaningful order instead of random shuffling. Training is done in learning steps where, in each step, the difficulty of the examples is increased. Curriculum learning provides two benefits: faster convergences of neural methods and finding a better local minimum. Many aspects of multi-modal language models are well studied, and curriculum learning methods are applied to other NLP subjects. However, to the best of our knowledge, there has not been a study that explored curriculum learning approaches in multi-modal language modeling.

3. Method

In this section, we introduce the details of the proposed model and dataset. First, a newly created dataset from Wikimedia Commons is described in Section 3.1. In the following Sections 3.2 and 3.3, the proposed model, along with the training method, is explained.

3.1. Wikimedia Commons Dataset

Wikimedia Commons (https://commons.wikimedia.org/wiki/Main_Page, accessed on through 1 January 2020 to 13 April 2020) is a repository of free-to-use images that is a part of Wikimedia Foundation. Wikimedia Commons files are used across all Wikimedia projects in all languages, including Wikipedia, Wiktionary, Wikibooks, Wikivoyage, Wikispecies, Wikisource, Wikinews, or downloaded offsite use. It comprises approximately 65 million images that take about 250 TB of space. The images also contain captions, descriptions, and timestamps.

To retrieve the images, one must send queries to the Wikimedia Commons website. To this end, we have used two different sets of query words to construct datasets. For retrieving the entire dataset, the dictionary of the BERT model [1] is used. As for getting the subset that we primarily used in this work, UWA MRC psycholinguistic dataset words are used.

UWA MRC Psycholinguistic Dataset [17] contains 98538 words and their properties, such as type, meaningfulnes, concreteness, part-of-speech, familiarity, and many more. Concreteness scores which are used in this research are derived from merging the two datasets provided by References [51,52].

In this dataset, 4293 out of 98538 words have a concreteness rating, rated by human annotators. Human annotators are asked to rate the concreteness of words between (including) 1 and 7, where the higher the score, the more concrete the word is. The mean of all users' scores is the final concreteness rating of the word, which is scaled between 100 and 700. Overall, the most abstract term in the dataset is "as" with a rating of 158, and the most concrete word is "milk" with a score of 670. The mean rating of all terms is 438, and the standard deviation is 120.

To successfully integrate this dataset into our task, some processing is required. Although the UWA MRC Psycholinguistic dataset successfully identifies the concreteness of words, it considers the words in isolation, unlike this work, where contextual embeddings and language models regard words in their context. Therefore, all the stop-words are removed (stop-words from the NLTK library are used) from the dataset, considering that they can appear in various contexts with different levels of concreteness and therefore can lead to misleading results. It is observed from the dataset that the lowest-rated words are usually stop-words, such as "as", "therefore", and "and". Thus, a lot of abstract words are removed in the lower bound. The most abstract word in the dataset after the removal is "apt" with a rating of 183. The final version of the dataset contains 1674 abstract and 2434 concrete words.

For each word, a query is sent to the Wikimedia Commons website with 1000 as a maximum threshold for the number of results. As a result, we have images, their corresponding captions, descriptions, and concreteness labels. Figure 1 shows the number of images returned for each query word in UWA MRC psycholinguistic dataset. As seen from the graph, most of the query words returned less than 100 results despite a large threshold. Only around a hundred words have more than 500 images associated with them. The number of samples collected is shown in Table 1. More than 43 million images are collected using the dictionary of BERT, while approximately 3.2 million images are collected using the words in UWA MRC psycholinguistic dataset. We can also observe that not all images have a description and/or caption associated with them. Some images contain only captions, some images contain descriptions but no caption, and, finally, some images do not contain any textual information at all. In total, 630,000 images contain captions, and approximately 2 million images contain descriptions. Overall, there is an overlap between both sets which means that some images contain both captions and descriptions.

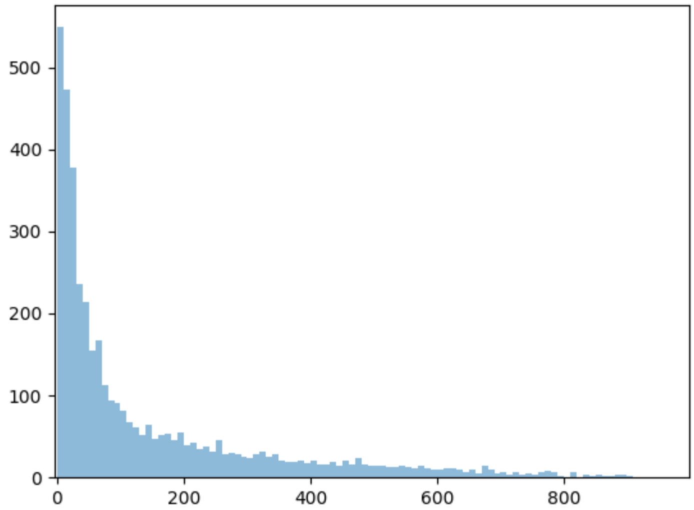


Figure 1. Histogram of the samples retrieved for words. Horizontal axis shows the number of images retrieved, while the vertical axis shows the amount of words which have that many images associated with them.

Table 1. Wikimedia Commons dataset statistics.

Dataset	# of Images	# of Captions	# of Descriptions
Complete Dataset	43,726,268	1,022,829	17,767,000
Subset (queried w/UWA MRC words)	3,206,765	629,561	1,961,567

The retrieved images have many formats, such as .jpeg, .jpg, .jpe .png, .apng, .gif, .tif, .tiff, .xcf, .webp, and many image modes, such as RGB (3 × 8-bit pixels, true color), CMYK (4 × 8-bit pixels, color separation), I (32-bit signed integer pixels), I;16 (16-bit unsigned integer pixels). Although many of these formats and modes are supported, we eliminated some of them. Images with the extension .xcf and .webp are filtered because mainstream image processing libraries do not support them. In addition to this, images with mode I (and other modes of I, such as I;16, I;16L, I16B, and so on) are eliminated because they are single-channel image modes, and the neural network models that process these images run with multi-channel inputs. Nearly 26,000 images are eliminated after this filtering. In the final version of the dataset, there are approximately 603,000 images with captions, where 177,000 belongs to abstract concepts, while 425,000 belongs to concrete concepts.

Many images in Wikimedia Commons have a very high resolution (resolutions, such as 3000 × 5000, 6000 × 6000, are very common), therefore requiring huge storage space. In addition to the filters applied above, a resize operation is performed to cope with this storage problem. All images are converted to a resolution of 224 × 224 since all the image models (GoogleNet [53], VGG [54], Resnet [55]) run with those.

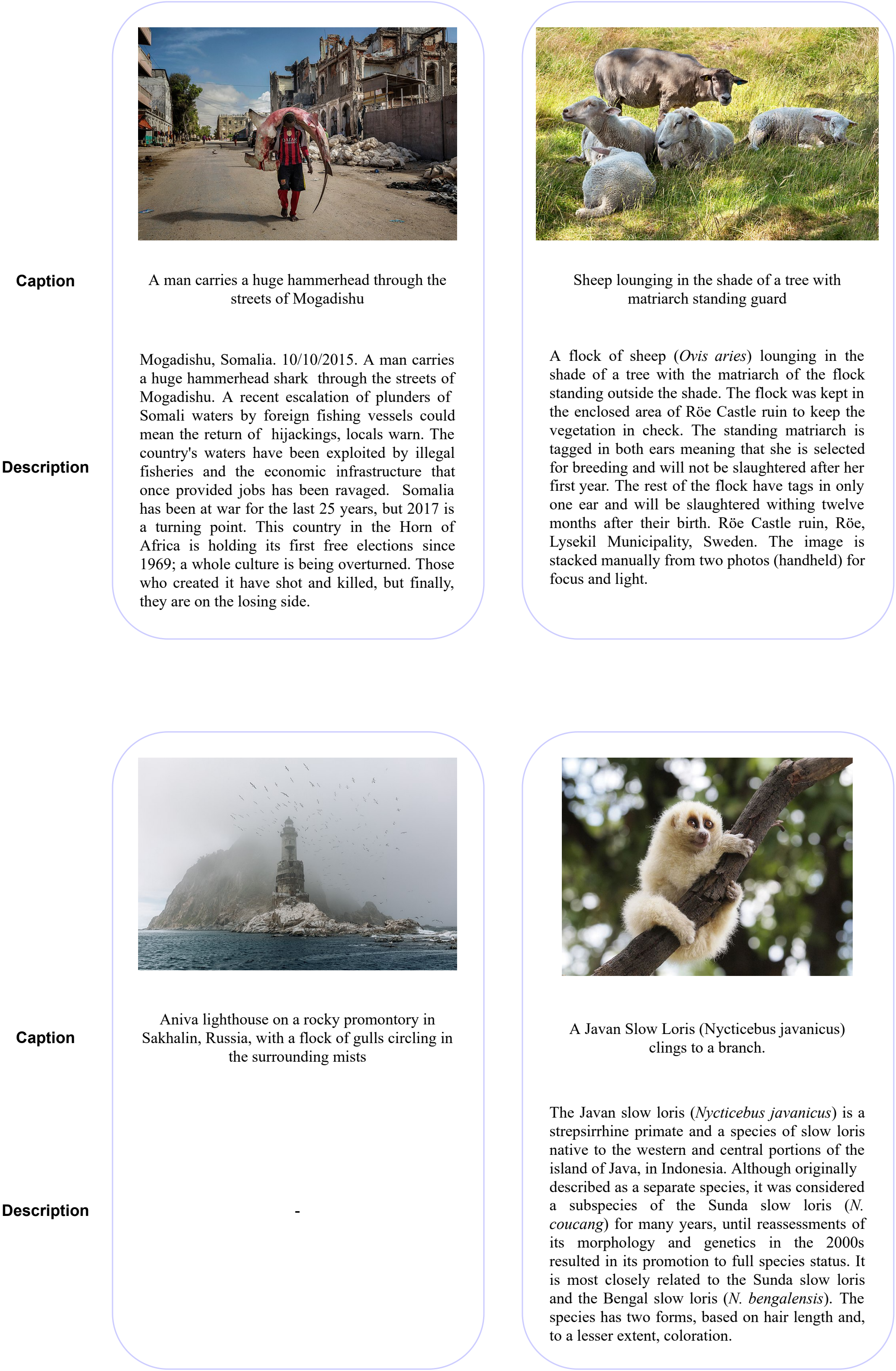
Figure 2 shows some example images and their corresponding captions and descriptions from the collected Wikimedia Commons dataset. The selected images have captions and descriptions, except for the bottom-left image where a description does not exist.

One thing to be observed from these images is, indeed, the images and the texts convey different information on the relationship of concepts. For example, there is no textual information in the top-left image, neither in the caption nor in the description, about the buildings that can be seen in the image. However, streets are primarily located near buildings (almost 70% of all images from Wikimedia Commons contains buildings when you search for the keyword “street”), which is captured by the image. Therefore the system can learn a relationship of concrete concepts, such as “street” and “building”, from the pictures without relying on the text. Similarly, the image contains no clue about its location, but it is understandable from both the caption and the description that it is in Mogadishu, Somalia. In the same vein, in the bottom-left image; there is no mention of a sea/lake in the text, but the lighthouse and the sea/lake can be seen together (which occur with almost no exception in real life) in the image, which will help the model to learn their relationships better. So, a language model trained with both images and text can help to improve the performances of language models.

Although the collected dataset contains captions and descriptions, captions are used to train the multi-modal language model. The reason is two-fold. We observed that descriptions in Wikimedia Commons are unclean. They include many additional texts, such as copyright notices, information about the photographer, or information about how the photograph is taken (such an example can be seen in the last sentence of the top-right image of Figure 2). On the other hand, captions are already cleaned and contain information only about the picture itself. Because of the requirement of tedious cleaning, we relied on captions.

The second but most important reason is the image-text alignment issues. Captions are written to describe the images briefly without giving any other information or making any further comment classified as common-sense knowledge or real-world knowledge. Contrarily, descriptions contain much information that cannot be seen in or referred from the images. Although these additional pieces of knowledge can be essential and valuable in other tasks, they break the image-text alignment and lead to learning noisy contexts in language modeling. If we take the top-right image in Figure 2 as an example, we can see how this can affect the language models. The description of the top-right image provides many semantically similar words to the context of the image, which is sheep lounging in a field, such as “breeding”, “slaughtered”, and “vegetation”. However, it also provides a lot of different or unrelated words, such as “castle”, “ruin”, “municipality”, which has very little to do with the image itself. Consequently, this leads to learning from an accidental relationship, for example, between the context of “sheep” and the context of “municipality”.

On account of this fact, captions are used in all language modeling tasks in this work to provide a better image-text alignment in training samples.



Caption

Description

A man carries a huge hammerhead through the streets of Mogadishu

Mogadishu, Somalia. 10/10/2015. A man carries a huge hammerhead shark through the streets of Mogadishu. A recent escalation of plunders of Somali waters by foreign fishing vessels could mean the return of hijackings, locals warn. The country's waters have been exploited by illegal fisheries and the economic infrastructure that once provided jobs has been ravaged. Somalia has been at war for the last 25 years, but 2017 is a turning point. This country in the Horn of Africa is holding its first free elections since 1969; a whole culture is being overturned. Those who created it have shot and killed, but finally, they are on the losing side.

Sheep lounging in the shade of a tree with matriarch standing guard

A flock of sheep (*Ovis aries*) lounging in the shade of a tree with the matriarch of the flock standing outside the shade. The flock was kept in the enclosed area of Röe Castle ruin to keep the vegetation in check. The standing matriarch is tagged in both ears meaning that she is selected for breeding and will not be slaughtered after her first year. The rest of the flock have tags in only one ear and will be slaughtered withing twelve months after their birth. Röe Castle ruin, Röe, Lysekil Municipality, Sweden. The image is stacked manually from two photos (handheld) for focus and light.

Caption

Description

Aniva lighthouse on a rocky promontory in Sakhalin, Russia, with a flock of gulls circling in the surrounding mists

-

A Javan Slow Loris (Nycticebus javanicus) clings to a branch.

The Javan slow loris (*Nycticebus javanicus*) is a strepsirrhine primate and a species of slow loris native to the western and central portions of the island of Java, in Indonesia. Although originally described as a separate species, it was considered a subspecies of the Sunda slow loris (*N. coucang*) for many years, until reassessments of its morphology and genetics in the 2000s resulted in its promotion to full species status. It is most closely related to the Sunda slow loris and the Bengal slow loris (*N. bengalensis*). The species has two forms, based on hair length and, to a lesser extent, coloration.

Figure 2. Example images and their corresponding captions and descriptions from the Wikimedia Commons Dataset.

There have been several other multi-modal datasets proposed in the literature that consist of image-text pairs, such as Flickr [56], MS COCO [57], Wikipedia, British Library, and ESP Game[58]. Table 2 shows the collected dataset in comparison with these multi-modal datasets. The Flickr dataset and MS COCO dataset contain image-caption pairs, while the Wikipedia dataset provides the images in Wikipedia with their corresponding articles. The British Library book dataset, on the other hand, contains historical books and the pictures depicted in them. Finally, the ESP game dataset consists of 5 words for each image labeled by human annotators. Although both Wikipedia and BL datasets provide much longer texts, they lack the image-text alignment of caption datasets. Therefore, caption datasets, such as MS COCO, Flickr, or the proposed dataset in this work, are more suited to the task of multi-modal language modeling. Compared with these image captioning datasets, the size of the collected dataset is much greater. As deep neural representations have massive data requirements, it is preferable to have such a large amount of data. Recently, the WIT [59] dataset was also proposed, with a large number of image-text pairs that can be used for multi-lingual, multi-modal pre-training. It contains 11.4 million unique images with captions and descriptive text from Wikipedia articles for various languages. Among them, 3.98 million images have textual information in English, where 568,000 of them have captions. In addition to captions, the collection also includes contextual data, such as page titles, page descriptions, section titles, etc., with their descriptions. However, the most significant benefit of the proposed dataset is the concreteness labels provided for each image-text pair which might be very useful for various tasks, especially for the multi-modal language modeling. The other datasets mentioned in this section, including WIT, do not contain that information.

Table 2. Comparison of Wikimedia Commons to other multi-modal datasets.

Dataset	# of Images	Textual Source	Ave. Word Length	Additional Info.
Flickr [56]	32,000	Captions	9	-
COCO [57]	123,000	Captions	10.5	-
Wikipedia	549,000	Articles	1397.8	-
BL	405,000	Books	2269.6	-
ESP[58]	100,000	Object Annotations	5	-
WIT[59]	11.4 million	Captions/Articles	-	-
	3.98 million	Captions/Article (En)	-	-
	568,000	Captions (En)	-	-
Wikimedia Commons (ours)	3.2 million	-	-	Concreteness Ratings
	629,000	Captions	10.2	Concreteness Ratings
	1.96 million	Descriptions	57.4	Concreteness Ratings

3.2. Model

The overall architecture of the proposed model can be seen in Figure 3. The model is comprised of three main parts: text processing part, image processing part, and a fusion mechanism where the outputs of text and image models are combined. Each piece is explained below in its respective subsection.

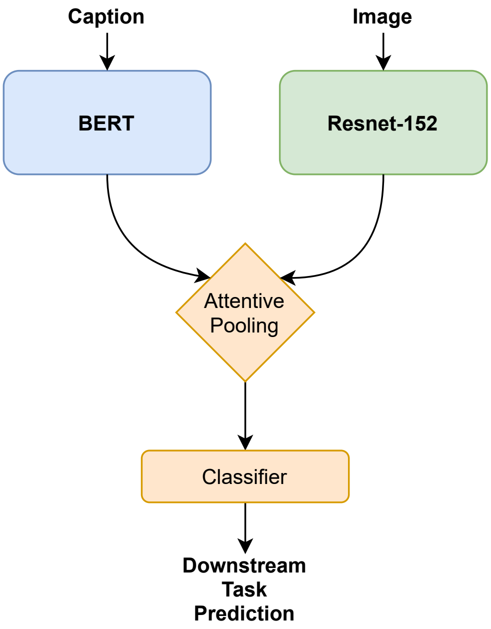


Figure 3. Proposed black-box model architecture.

3.2.1. Text Model

In this work, BERT is primarily used for processing text input, while we also utilized DistilBERT in some of the tests.

BERT [1] is a neural network model that uses a bidirectional transformer architecture [36], a self-attention mechanism to learn contextual word embeddings. It has multiple layers of transformers (12 in BERT-base, 24 in BERT-large) where each layer has 12 attention heads that span the entire sentence from both right-to-left and left-to-right, learning "where to look" by producing probabilistic weights for each word.

Different from the earlier language modeling approaches, BERT does not use next word prediction as an objective. Instead, it uses two training objectives: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). For the MLM objective, randomly selected words are occluded from the model and labeled as masks. The model tries to predict the masked word as the training objective. Attention heads do not span these masked words since it would create a bias for the prediction. Using MLM enables the model to learn contextual dependencies among words very successfully. The embedding of a word is computed depending on the surrounding terms instead of using the same vector in the embedding space for every instance of that word. For the NSP objective, the model tries to predict whether the two sentences provided to the model belong to the same context or not. It helps BERT to consider multiple sentences as context and to represent inter-sentence relations.

In addition to the token (word) embeddings, BERT also uses segment (sentence) embeddings and position embeddings (words' position in segments) as input. While sentence embedding determines which sentence the word is in, positional embedding acknowledges the word order. Therefore, a word's embedding is fed to the model as the average of its token embedding, sentence embedding, and positional embedding. This input structure has many benefits: Positional embeddings raise the model's awareness of word order, while segment embeddings help the NSP objective. In addition, giving multiple sentences as input helps BERT be integrated into most downstream tasks requiring inter-sentence connections, such as Question Answering and Natural Language Inference (NLI), easily, without requiring any other architecture.

To integrate BERT to downstream tasks, an additional fully connected layer is used on top of transformer layers to predict the given text's class instead of the target (masked) word. Usually, the Wikipedia dataset is used to pre-train the model on MLM and NSP objectives. The resulting parameters are fine-tuned on the downstream task with the addition of the aforementioned fully connected layer.

In this study, we performed some tests using the DistilBERT language model. DistilBERT [60] is based on the original BERT model. It is a more efficient version of BERT in expense for a minor deficiency in classification performance. It retains 97% of BERT's performance while using 40% fewer parameters. To accomplish this, they use knowledge distillation, where a small model is trained to reproduce the behavior of a larger model (DistilBERT and BERT, respectively, in this case). Knowledge distillation aims to make the student model (DistilBERT) predict the same values as the teacher model (BERT) using fewer parameters. This way, one can transfer the knowledge learned by the teacher model to more efficient student models. Parameter reduction from BERT to DistilBERT comes from the removal of some of the transformer layers in BERT. The authors of DistilBERT show that some of the parameters of BERT are not used in the prediction, therefore, do not contribute to learning downstream tasks. Consequently, they suggest removing some layers and use the knowledge distillation technique to create a more efficient language model.

3.2.2. Image Model

We used Resnet [55] as the image model due to its success in many image processing tasks. It is a very deep neural network model that relies on convolutional neural network architecture. At the time it is published, it was the state-of-the-art model in the ImageNet [61] object classification challenge.

Resnet has several different variations in network depth: 34-layered model Resnet34, 50-layered model Resnet50, 101-layered model Resnet101, and, finally, the largest model with 152-layers Resnet152. Each layer consists of several 1×1 and 3×3 convolutions. Each model starts and ends with an average pooling operation before the first layer and after the last layer.

Stacking so many layers in deep neural networks naively does not immediately lead to better results; instead, it causes performance degradation problems. An increase in the depth of a model causes an increase in training errors, and accuracy is saturated. To deal with this issue and build substantially deeper networks, authors needed a workaround. Therefore, shortcut connections called residual connections are used. These shortcut connections are used after every two layers in the architecture, propagating the inputs to the outputs of those two layers. They are parameter-free, which means that they do not perform any operation on the inputs, such as pooling, convolution, or multiplication; therefore, they do not contain any learnable parameters. It is shown that these shortcut connections can overcome the performance degradation problem in very deep neural network architectures, making models, such as Resnet, very successful at stacking many layers and capturing more features than the prior models.

In this work, Resnet152 is used because it outperforms the smaller Resnet models, and the Wikimedia Commons dataset was large enough to tune such a large model.

3.2.3. Text-Image Combination Method

Combining multiple modalities can be problematic and risks breaking the learned semantic relationship of words by individual models. Thus, many studies in this field focus on the fusion of modalities.

We used attentive pooling networks [62] to combine the text and vision parts of the model. It is a two-way attention mechanism that is aware of both modalities and jointly learns to attend over them through matrix multiplications and pooling operations.

Attentive pooling takes the hidden states of each word in BERT as textual input and takes the last layer of Resnet in the form of a matrix as visual input. These inputs are multiplied with the matrix U , which is composed of parameters to learn and passed through *tanh* activation. The result is a single matrix of visual features on the rows and textual features on the columns. This representation scheme allows features from different modalities to be jointly represented in a single matrix where max-pooling operation is performed over each row and column to find out the most important feature dependent upon the other modality. Two vectors, I_{output} and T_{output}, are the outcomes of the attentive pooling mechanism. For fine-tuning this model on downstream tasks, these two outputs are concatenated and passed through an additional fully connected layer to reduce the dimension to the number of classes.

3.3. Multi-Modal Language Model Training

The idea of pre-training neural language models is borrowed from the advances in image processing models [32]. It is shown in both vision and text models that pre-training a model on a preliminary image/text understanding task improves the performance vastly.

For image processing, the pre-training task is usually the object classification task on the ImageNET dataset [61]. ImageNET dataset has 1.2 million images that are hand-labeled into 1000 categories. Respective models are trained to predict the objects in each image by adding a fully connected layer on top to reduce the feature vectors' size to 1000. The aim here is to teach the model basic image understanding: Identifying objects and entities in images. It is shown by many vision models that they are even able to differentiate images of 120 different dog breeds in the imageNET dataset, such as "Australian terrier" and "Airedale terrier". They manage to do this by using the shapes and colors of entities in the pictures.

The process is similar for language models, with the only difference in pre-training objectives. Earlier models (before BERT) used next word prediction in huge unlabeled text, such as Wikipedia and Common Crawl text. The aim was to predict the next word given the previous set of words. Starting from BERT and onward, the pre-training objective changed from the next word prediction to masked language modeling. This method allowed the text models to successfully grasp language understanding by training them on massive datasets containing billions of words. They learned the meaning and semantic/syntactic relations of words (due to distributional hypothesis), which are fundamental to any downstream task.

Once the pre-training objective is completed and the image/text model gained basic image/language understanding, respectively, the last fully connected layer is removed from the model and replaced with an appropriate classification layer according to the task at hand. The model is, then, fine-tuned for the downstream task. For image models, downstream tasks can be object detection, semantic segmentation, etc., while, on the textual models, they are composed of sentiment analysis, sentence classification, natural language inference, and so on.

In this work, we adopt a novel multi-modal pre-training objective. The idea is inspired from the advances in cognitive psychology. It is shown that language acquisition in children starts with experiential information and continues with textual information [11,12]. As Kiela et al., in 2015 [63], stated, perceptual information is more relevant for, e.g., elephants than it is for happiness. In other words, we first learn the language through images and learn concrete concepts, and then we start learning abstract concepts from textual sources.

Advancements in computational linguistics also reinforce this idea by showing that concrete examples in language are easier to learn, while abstract ones are more challenging. Hessel et al., in 2018 [64], showed that the more concrete the downstream task gets, the easier it becomes for language models. Bruni et al., in 2014 [38], showed that the semantic/syntactic similarities of concrete examples on the MEN dataset are easier to learn, while the abstract words can get ambiguous. They prove this by showing that the concrete examples have a 0.78 Spearman correlation rank, while the abstract examples have 0.52 (contributing to an overall 0.76).

To adopt this learning scheme to this project, the Wikimedia Commons Dataset (see Section 3.1) is divided into two categories: Abstract samples and concrete samples. We determined concrete/abstract examples based on the concreteness levels of words from the UWA MRC Psycholinguistic Database. First, we fed the image model concrete examples. Then, we trained the textual model with all of the samples concrete and abstract combined, in a curriculum learning fashion [15,16]. Therefore, the learning model mimics humans through this pre-training process.

4. Experiments

The first step of experimentation was to measure the informativeness of the collected dataset. To meet this objective, we selected concreteness classification and tested the performance of captions in this task. Moreover, to show the expressiveness of captions relative to regular texts, we did the same classification with the regular Wikipedia articles. We worked with the June 2020 version of wikidumps, which consists of $6,957,578$ documents in total.

To prepare the dataset for comparison, we search for articles in the Wikipedia dataset using UWA MRC Psycholinguistic dataset words. Specifically, each article titled with the corresponding words is retrieved. We concatenated the captions that corresponded to the same word and removed the terms that do not have a Wikipedia article to match captions with the Wikipedia articles further. After this, there are 4108 samples remaining in the dataset, which is partitioned into the train (70%), dev (10%), and test (20%) sets randomly.

Table 3 shows the results of DistilBERT and BERT along with the random baselines on these datasets. The results show that, although the Wikimedia captions give us worse than the Wikipedia articles, results are not far off, making the Wikimedia captions almost as informative as the Wikipedia text itself.

Table 3. Results comparing the informativeness of the proposed dataset.

Model	Wikimedia Captions				Wikipedia Articles			
	Accuracy	**Precision**	**Recall**	**F1**	**Accuracy**	**Precision**	**Recall**	**F1**
Random	0.5171	0.5171	0.5171	0.5171	0.5255	0.5255	0.5255	0.5255
DistilBERT	80.91 (−1.47 + 2.28)	80.89 (−1.47 + 2.31)	80.91 (−1.47 + 2.28)	80.83 (−1.41 + 2.36)	86.54 (−1.97 + 0.53)	86.69 (−1.08 + 0.83)	86.54 (−1.97 + 0.53)	86.58 (−1.99 + 0.50)
BERT	82.37 (−1.88 + 1.19)	82.35 (−1.96 + 1.10)	82.37 (−1.88 + 1.19)	82.31 (−1.97 + 1.12)	85.60 (−1.91 + 1.35)	85.69 (−1.89 + 1.24)	85.60 (−1.91 + 1.35)	85.45 (−1.07 + 1.49)

Table 4 shows the experimental results of the multi-modal pre-training task on the test set. As stated before, we performed this pre-training in a curriculum learning fashion. Our image model is further pre-trained with concrete samples of the training set, and then the text model is trained on all the samples on the training set, concrete, and abstract combined. The results show the performance of each model on the test set of the pre-training dataset. While the image model obtained 0.8147 F1 on the concrete samples, the text model obtained 0.8707 and 0.6518 F1 on the concrete and abstract samples. Although we did not pre-train the image model on abstract samples, we also show its results to give an idea.

Table 4. Experimental results of the multi-modal pre-training task.

Model	Accuracy	Precision	Recall	F1	F1-abs	F1-Conc
Bert	0.8116	0.8057	0.8116	0.8069	0.6518	0.8708
Resnet	0.7001	0.6472	0.7001	0.6383	0.2144	0.8147

We can draw several conclusions from the results. Firstly, the results comply with References [38,64]: Identifying concrete concepts is much easier than identifying abstract concepts. Both the Resnet and BERT models perform above 0.8 in terms of F1 scores for the concrete class. On the other hand, the F1 score of Resnet on the abstract class turns out to be significantly lower, with a value of 21.5. These results show that both image and text models struggle more with abstract concepts than concrete ones.

Secondly, the results of Resnet agree with the scientific work (i.e., References [11,12]) on human language acquisition. Thus, they also comply with the curriculum learning objectives in this work: Experiential information is used early in language acquisition on concrete concepts, while leaving its place to textual information for learning abstract ones.

It can be argued that, no matter how abstract an idea is, one needs to find a concrete example to show that in an image. For example, the image/caption pairs returned for the search word "dream" frequently contain pictures of places. Although the term itself can safely be considered abstract, one needs to find a particular and concrete idea/object to represent it as an image. Therefore, we can conclude that images almost always contain concrete concepts. To determine abstractness, one should use a diverse set of images belonging to a particular concept instead of individual images (the variance in images for the word "tomato" is very low, with the first 25 results are all images of single or a couple of red tomatoes, while the variance in images for the word "dream" is very high, ranging from the picture of places, famous people to screenshots of literary work).

To validate the effectiveness of the proposed multi-modal pre-training scheme, we tested the model's performance on a downstream NLP task. As a multi-modal task, Visual Question Answering fits nicely with our objective. Visual Question Answering dataset is a multi-modal dataset that was proposed by Antol et al., in 2015 [65]. It includes approximately 200,000 images from the COCO dataset [57]. Each image in this dataset has multiple questions associated with it in various forms, such as yes/no questions and open-ended questions. Yes/No questions are binary questions, such as "Is the umbrella upside down?", while the open-ended questions, such as "Who is wearing glasses?", require more diverse answers. Close to 40% of all questions are yes/no questions, and the rest is open-ended. Open-ended questions have a variety of types, including but not limited to "What is . . . ?", "How many . . . ?", and "Who is . . . ?".

Although the dataset requires a lot of inference between modalities, Agrawal et al., in 2018 [13], stated that the dataset includes bias towards some question/answer pairs. In their work, they showed that questions related to colors ("What is the color of . . . ?" or "is . . . white?") almost always lead to the answers of white/no for open-ended and yes/no questions, respectively. Similarly, Goyal et al., in 2017 [66], suggested that answering the questions that are starting with the phrase "Do you see a ...?" with yes blindly leads to an accuracy of 87% among those questions. Therefore, using language priors alone, a model can correctly predict a significant amount of questions. The authors develop the second version of the dataset to overcome this problem, which has additional samples to balance the biased question/answer pairs. This update increased the dataset size to 443 thousand, 214 thousand, and 453 thousand pairs (question, image) for train, dev, and test sets, respectively. The results reported in this manuscript refer to this new dataset as v2, while they refer to the former as v1.

Table 5 shows the model's performance on VQA. The best result is obtained when both multi-modal pre-training and attentive pooling mechanisms are used, although the performance is consistent across all configurations. In terms of accuracy, there is a 1.01% difference between the best performing model (with multi-modal pre-training and attentive pooling) and the worst (with fully connected layer and without multi-modal pre-training). Performance difference becomes more significant in F1: a 3.37% increase can be observed between the best and worst-performing models (model with multi-modal pre-training and attentive pooling, and model without multi-modal pre-training with a fully connected layer, respectively, similar to the previous case).

Table 5. Model performance on VQA dataset v2. (FC = Fully-connected, AP = Attentive pooling).

Model	Multi-Modal Pre-Training	Combination Method	Accuracy	F1	Precision	Recall
Bert + Resnet	✗	FC	53.12	50.71	54.07	53.12
Bert + Resnet	✓	FC	53.17	52.79	53.34	53.17
Bert + Resnet	✗	AP	53.56	52.91	53.69	53.56
Bert + Resnet	✓	AP	54.13	54.08	54.07	54.13

One can better analyze performance differences with ablation studies. Table 6 reports the relative improvements of each component. Each column represents the percentage increase in relative performance when the feature/component in the row is replaced or enhanced by the feature/component in the column. The results show that multi-modal pre-training increases the model's performance regardless of the underlying fusion mechanism (Fully-connected or attentive pooling). It leads to a 4.1% increase when used with fully connected layers and leads to a 2.21% increase when used with attentive pooling networks. Similarly, the attentive pooling mechanism improves the performance of the model in both cases: When the fully-connected layer is replaced with attentive pooling, it amounts to an increase of 4.34% without multi-modal pre-training and an increase of 2.44% with multi-modal pre-training. Additionally, from the first row, we can conclude that replacing FC with an attentive pooling mechanism is slightly more beneficial than using FC together with multi-modal pre-training. Overall, as the results suggest, using both attentive pooling and multi-modal pre-training proved to be useful and led to an increase in performance up to 6.65% compared to the baseline model.

Table 6. Results of the ablation study. Relative performance improvements (%) of each component in terms of F1. MMPT = Multi-modal pre-training, FC = Fully-connected, AP = Attentive pooling.

	FC	**MMPT + FC**	**AP**	**MMPT + AP**
FC	0	4.10	4.34	6.65
MMPT + FC	-	0	0.23	2.44
AP	-	-	0	2.21
MMPT + AP	-	-	-	0

Table 7 shows the performance of the multi-modal models described in Section 2 on the VQA task. We share the results on version 1 and version 2, though it would only be fair to compare the models that run on the same version. The models that run on both versions (stacked attention network (SAN) and GVQA) suggest that a performance difference between 3–7% can be expected between the versions, most likely due to the effect of language priors. Human baselines, obtained on the 3000 samples in the training set of the v1 dataset, are also provided in the top part.

Table 7. Experimental results on VQA task. Top part shows human baselines.

Model	**Dataset Version**	**Accuracy**
Question	v1	40.81
Question + Caption	v1	57.47
Question + Image	v1	83.30
SAN [67]	v1	58.9
GVQA [13]	v1	51.12
SAN [67]	v2	52.2
GVQA [13]	v2	48.24
Anderson et al., 2018 [14]	v2	70.34
DFAF [48]	v2	70.34
VilBERT [9]	v2	70.92
ours	v2	54.13

Although human baselines are on v1 and our performance is on the v2 version of the dataset, our 54.13% accuracy indicates that the model can perform similarly to humans when given only questions and corresponding captions without images. Compared to the other models, ours performed better than the earlier models but cannot reach the success obtained by the state-of-the-art model (VilBERT), which has 70.92% accuracy. VilBERT processes paired visiolinguistic data in the architecture of BERT to exploit visual grounding in a task-agnostic way.

It should be noted that there are subtle but vital differences between our model and the VilBERT model. The main focus of VilBERT is to process text and image streams in parallel under the transformer architecture to encode their relationship in a pre-trained model to have optimized performance in downstream tasks. On the other hand, the main focus of this work is to optimize the model for the fusion of modalities and curriculum learning. Although our work is much similar to earlier multi-modal works in this regard, our model is a language pre-training model, not a task-specific architecture. The main difference in our work is to add curriculum learning methodology on top of the pre-trained models.

Other than the main focus described above, several reasons might lead to the performance discrepancy between the proposed model and the state-of-the-art models, such as VilBERT. First, the number of learnable parameters in VilBERT is much greater than the proposed model (~600 million versus ~170 million). Second, VilBERT uses the Faster-RCNN [68] model to match each word in the text with the corresponding image patch, while our model uses the Resnet-152 model on the entire image. One could argue that the better alignment provided by the faster-RCNN method might lead to better learning since the model also learns which part in the image a particular word corresponds to. Providing such an alignment could also benefit the proposed model for catching up with the performance of the state-of-the-art models.

5. Conclusions

This study aims to contribute to one of the oldest and most predominant subjects in computer science: language modeling. Since the distributional hypothesis in the early 1950s, many models with many different architectures and methodologies have been introduced in this field. Until recently, models focused on a single modality where a language learner is trained with plain text. Lately, however, the focus is shifted from single modality to multi-modal language models. An increase in the success of neural models, cheaper and more powerful hardware sources, and advances in cognitive science were the major driving forces behind this change.

Similar to this latest trend, this work aims to create a language model/representation technique inspired by the advances in cognitive science, which states that language acquisition in humans starts with the experiential information for concrete concepts and continues with distributional information for abstract concepts. To this end, we combined the BERT and Resnet models with the attentive pooling mechanism to construct a multi-modal language model and embeddings. The image model is trained with the concrete samples from Wikimedia samples first, and then the text model is trained with concrete and abstract examples combined in a curriculum learning fashion. Additionally, we constructed a new dataset composed of image caption pairs from Wikimedia Commons based on concrete/abstract metadata.

The contribution of this work is two-fold: First, a new dataset, created from Wikimedia Commons, is introduced, which has approximately 3.2 million images, with 630,000 captions, 1.96 million descriptions, and concreteness labels. Second, a new training scheme for multi-modal pre-training is introduced. We inspired this novel learning scheme from the curriculum learning approaches in artificial intelligence. The results show that, although the model could not outperform state-of-the-art results, the multi-modal pre-training objective can significantly increase the models' performance. Our results also confirm the findings in the literature by showing that it is harder to detect and classify abstract samples.

Author Contributions: Conceptualization, E.S. and S.T.; methodology, E.S. and S.T.; software, E.S.; validation, E.S. and S.T.; formal analysis, E.S.; investigation, E.S.; resources, E.S. and S.T.; data curation, E.S.; writing—original draft preparation, E.S. and S.T.; writing—review and editing, E.S. and S.T.; visualization, E.S.; supervision, S.T.; project administration, S.T. Both authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: UWA MRC dataset is available from here: https://websites.psychology.uwa.edu.au/school/MRCDatabase/uwa_mrc.htm (accessed on 31 March 2020), Wikipedia dumps can be downloaded from here: https://dumps.wikimedia.org/backup-index.html (accessed on 16 June 2020). The data collected in this study are available on request from the corresponding author.

Acknowledgments: The Titan V used for the experiments in this work is donated by the NVIDIA Corporation.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019), Minneapolis, MN, USA, 2–7 June 2019; Volume 1, pp. 4171–4186.
2. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
3. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 26–30 April 2020.
4. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q.V. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Advances in Neural Information Processing Systems 32*; Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2019; pp. 5753–5763.
5. Sun, Y.; Wang, S.; Li, Y.; Feng, S.; Chen, X.; Zhang, H.; Tian, X.; Zhu, D.; Tian, H.; Wu, H. ERNIE: Enhanced Representation through Knowledge Integration. *arXiv* **2019**, arXiv:1904.09223.
6. Sun, Y.; Wang, S.; Li, Y.; Feng, S.; Tian, H.; Wu, H.; Wang, H. ERNIE 2.0: A Continual Pre-training Framework for Language Understanding. *arXiv* **2020**, arXiv:1907.12412.
7. Li, L.H.; Yatskar, M.; Yin, D.; Hsieh, C.; Chang, K. VisualBERT: A Simple and Performant Baseline for Vision and Language. *arXiv* **2019**, arXiv:1908.03557.
8. Sun, C.; Myers, A.; Vondrick, C.; Murphy, K.; Schmid, C. VideoBERT: A Joint Model for Video and Language Representation Learning. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 7463–7472. [CrossRef]
9. Lu, J.; Batra, D.; Parikh, D.; Lee, S. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Advances in Neural Information Processing Systems*; Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2019; Volume 32.
10. Griffiths, T.L.; Tenenbaum, J.B.; Steyvers, M. Topics in semantic representation. *Psychol. Rev.* **2007**, *114*, 2007.
11. Vigliocco, G.; Meteyard, L.; Andrews, M.; Kousta, S. Toward a theory of semantic representation. *Lang. Cogn.* **2009**, *1*, 219–247.
12. Andrews, M.; Vigliocco, G.; Vinson, D. Integrating experiential and distributional data to learn semantic representations. *Psychol. Rev.* **2009**, *116*, 463–498.
13. Agrawal, A.; Batra, D.; Parikh, D.; Kembhavi, A. Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.
14. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6077–6086.
15. Elman, J.L. Learning and development in neural networks: The importance of starting small. *Cognition* **1993**, *48*, 71–99. [CrossRef]
16. Bengio, Y.; Louradour, J.; Collobert, R.; Weston, J. Curriculum Learning. In Proceedings of the 26th Annual International Conference on Machine Learning, Association for Computing Machinery (ICML'09), New York, NY, USA, 19–24 June 2009; pp. 41–48. [CrossRef]
17. Coltheart, M. The MRC Psycholinguistic Database. *Q. J. Exp. Psychol. Sect. A* **1981**, *33*, 497–505.[CrossRef]
18. Wittgenstein, L. *Philosophical Investigations*; Basil Blackwell: Oxford, UK, 1953.
19. Harris, Z.S. Distributional Structure. *Word* **1954**, *10*, 146–162. [CrossRef]
20. Hinton, G.E.; McClelland, J.L.; Rumelhart, D.E. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*; Chapter Distributed Representations; MIT Press: Cambridge, MA, USA, 1986; Volume 1, pp. 77–109.
21. Elman, J.L. Finding structure in time. *Cogn. Sci.* **1990**, *14*, 179–211.
22. Bengio, Y.; Ducharme, R.; Vincent, P.; Janvin, C. A Neural Probabilistic Language Model. *J. Mach. Learn. Res.* **2003**, *3*, 1137–1155.
23. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv* **2013**, arXiv:1301.3781.
24. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global Vectors for Word Representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.

25. Reisinger, J.; Mooney, R.J. Multi-prototype Vector-space Models of Word Meaning. In *Human Language Technologies, Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, HLT'10, Los Angeles, CA, USA, 1–6 June 2010*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2010; pp. 109–117.
26. Huang, E.H.; Socher, R.; Manning, C.D.; Ng, A.Y. Improving Word Representations via Global Context and Multiple Word Prototypes. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL'12), Jeju, Korea, 8–14 July 2012; Volume 1, pp. 873–882.
27. Luong, T.; Socher, R.; Manning, C. Better Word Representations with Recursive Neural Networks for Morphology. In Proceedings of the Seventeenth Conference on Computational Natural Language Learning, Sofia, Bulgaria, 8–9 August 2013; pp. 104–113.
28. Rothe, S.; Schütze, H. AutoExtend: Extending Word Embeddings to Embeddings for Synsets and Lexemes. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, 26–30 July 2015; Volume 1, pp. 1793–1803. [CrossRef]
29. Melamud, O.; Goldberger, J.; Dagan, I. context2vec: Learning Generic Context Embedding with Bidirectional LSTM. In Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL 2016), Berlin, Germany, 11–12 August 2016; pp. 51–61.
30. McCann, B.; Bradbury, J.; Xiong, C.; Socher, R. Learned in Translation: Contextualized Word Vectors. In Proceedings of the Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 6297–6308.
31. Dai, A.M.; Le, Q.V. Semi-Supervised Sequence Learning. In Proceedings of the 28th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2015; Volume 2, pp. 3079–3087.
32. Howard, J.; Ruder, S. Universal Language Model Fine-tuning for Text Classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; Volume 1, pp. 328–339. [CrossRef]
33. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep Contextualized Word Representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018), New Orleans, LA, USA, 1–6 June 2018; Volume 1, pp. 2227–2237.
34. Kim, Y.; Jernite, Y.; Sontag, D.A.; Rush, A.M. Character-Aware Neural Language Models. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; pp. 2741–2749.
35. Akbik, A.; Blythe, D.; Vollgraf, R. Contextual String Embeddings for Sequence Labeling. In Proceedings of the COLING 2018, 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 21–25 August 2018; pp. 1638–1649.
36. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.u.; Polosukhin, I. Attention is All you Need. In *Advances in Neural Information Processing Systems 30*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 5998–6008.
37. Dai, Z.; Yang, Z.; Yang, Y.; Carbonell, J.; Le, Q.; Salakhutdinov, R. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 2978–2988. [CrossRef]
38. Bruni, E.; Tran, N.K.; Baroni, M. Multimodal Distributional Semantics. *J. Artif. Int. Res.* **2014**, *49*, 1–47.
39. Kiros, R.; Salakhutdinov, R.; Zemel, R. Multimodal Neural Language Models. In Proceedings of the 31st International Conference on International Conference on Machine Learning (ICML'14), Beijing, China, 21–26 June 2014; Volume 32, p. II-595–II-603.
40. Liu, Y.; Guo, Y.; Bakker, E.M.; Lew, M.S. Learning a Recurrent Residual Fusion Network for Multimodal Matching. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 4127–4136. [CrossRef]
41. Hill, F.; Korhonen, A. Learning Abstract Concept Embeddings from Multi-Modal Data: Since You Probably Can't See What I Mean. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 255–265. [CrossRef]
42. Kiros, R.; Salakhutdinov, R.; Zemel, R. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. *arXiv* **2014**, arXiv:1411.2539.
43. Karpathy, A.; Joulin, A.; Li, F.-F. Deep Fragment Embeddings for Bidirectional Image Sentence Mapping. In Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS'14), Montreal, QC, Canada, 8–13 December 2014; Volume 2, pp. 1889–1897.
44. Wang, L.; Li, Y.; Lazebnik, S. Learning Deep Structure-Preserving Image-Text Embeddings. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 5005–5013. [CrossRef]
45. Lee, K.H.; Chen, X.; Hua, G.; Hu, H.; He, X. Stacked cross attention for image-text matching. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 201–216.
46. Socher, R.; Li, F.-F. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 966–973. [CrossRef]
47. Yu, L.; Lin, Z.; Shen, X.; Yang, J.; Lu, X.; Bansal, M.; Berg, T.L. MAttNet: Modular Attention Network for Referring Expression Comprehension. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.
48. Gao, P.; Jiang, Z.; You, H.; Lu, P.; Hoi, S.C.H.; Wang, X.; Li, H. Dynamic Fusion With Intra- and Inter-Modality Attention Flow for Visual Question Answering. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 6632–6641.[CrossRef]

49. Zellers, R.; Bisk, Y.; Farhadi, A.; Choi, Y. From Recognition to Cognition: Visual Commonsense Reasoning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
50. Shi, H.; Mao, J.; Xiao, T.; Jiang, Y.; Sun, J. Learning Visually-Grounded Semantics from Contrastive Adversarial Samples. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 21–25 August 2018; pp. 3715–3727.
51. Paivio, A.; Yuille, J.C.; Madigan, S.A. Concreteness, imagery, and meaningfulness values for 925 nouns. *J. Exp. Psychol.* **1968**, *76*, 1.
52. Gilhooly, K.J.; Logie, R.H. Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1944 words. *Behav. Res. Methods Instrum.* **1980**, *12*, 395–427.
53. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. In Proceedings of the Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
54. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2015**, arXiv:1409.1556.
55. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.[CrossRef]
56. Young, P.; Lai, A.; Hodosh, M.; Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguist.* **2014**, *2*, 67–78. [CrossRef]
57. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
58. von Ahn, L.; Dabbish, L. Labeling Images with a Computer Game. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'04), Vienna, Austria, 25 April 2004; pp. 319–326. [CrossRef]
59. Srinivasan, K.; Raman, K.; Chen, J.; Bendersky, M.; Najork, M. WIT: Wikipedia-Based Image Text Dataset for Multimodal Multilingual Machine Learning. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'21), Montreal, QC, Canada, 11–15 July 2021; pp. 2443–2449. [CrossRef]
60. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv* **2019**, arXiv:1910.01108.
61. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Li, F.-F. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
62. Santos, C.d.; Tan, M.; Xiang, B.; Zhou, B. Attentive pooling networks. *arXiv* **2016**, arXiv:1602.03609.
63. Kiela, D.; Rimell, L.; Vulić, I.; Clark, S. Exploiting Image Generality for Lexical Entailment Detection. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, 26–30 July 2015; Volume 2, pp. 119–124. [CrossRef]
64. Hessel, J.; Mimno, D.; Lee, L. Quantifying the Visual Concreteness of Words and Topics in Multimodal Datasets. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LA, USA, 1–6 June 2018; Volume 1, pp. 2194–2205. [CrossRef]
65. Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C.L.; Parikh, D. VQA: Visual Question Answering. In Proceedings of the International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
66. Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; Parikh, D. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6325–6334. [CrossRef]
67. Yang, Z.; He, X.; Gao, J.; Deng, L.; Smola, A. Stacked Attention Networks for Image Question Answering. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 21–29. [CrossRef]
68. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS'15), Montreal, QC, Canada, 8–13 December 2015; pp. 91–99.