

# Inferring Context from Pixels for Multimodal Image Classification

Manan Shah\*  
Stanford University  
manans@stanford.edu

Krishnamurthy Viswanathan  
Google Research  
kvis@google.com

Chun-Ta Lu  
Google Research  
chunta@google.com

Ariel Fuxman  
Google Research  
afuxman@google.com

Zhen Li  
Google Research  
zhenli@google.com

Aleksei Timofeev  
Waymo  
altimofeev@google.com

Chao Jia  
Google Research  
chaojia@google.com

Chen Sun  
Google Research  
chensun@google.com

## ABSTRACT

Image classification models take image pixels as input and predict labels in a predefined taxonomy. While contextual information (e.g. text surrounding an image) can provide valuable orthogonal signals to improve classification, the typical setting in literature assumes the unavailability of text and thus focuses on models that rely purely on pixels. In this work, we also focus on the setting where only pixels are available in the input. However, we demonstrate that if we predict textual information from pixels, we can subsequently use the predicted text to train models that improve overall performance.

We propose a framework that consists of two main components: (1) a phrase generator that maps image pixels to a contextual phrase, and (2) a multimodal model that uses textual features from the phrase generator and visual features from the image pixels to produce labels in the output taxonomy. The phrase generator is trained using web-based query-image pairs to incorporate contextual information associated with each image and has a large output space.

We evaluate our framework on diverse benchmark datasets (specifically, the WebVision dataset for evaluating multi-class classification and OpenImages dataset for evaluating multi-label classification), demonstrating performance improvements over approaches based exclusively on pixels and showcasing benefits in prediction interpretability. We additionally present results to demonstrate that our framework provides improvements in few-shot learning of minimally labeled concepts. We further demonstrate the unique benefits of the multimodal nature of our framework by utilizing intermediate image/text co-embeddings to perform baseline zero-shot learning on the ImageNet dataset.

\*Work conducted while author was at Google Research.

## CCS CONCEPTS

• **Computing methodologies** → **Image representations**; *Object identification*.

## KEYWORDS

computer vision; multimodal models; image classification; convolutional networks

### ACM Reference Format:

Manan Shah, Krishnamurthy Viswanathan, Chun-Ta Lu, Ariel Fuxman, Zhen Li, Aleksei Timofeev, Chao Jia, and Chen Sun. 2019. Inferring Context from Pixels for Multimodal Image Classification. In *The 28th ACM International Conference on Information and Knowledge Management (CIKM '19)*, November 3–7, 2019, Beijing, China. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3357384.3357987>

## 1 INTRODUCTION

In recent years, we have witnessed the emergence of ever-improving convolutional network (ConvNet) architectures for image classification [7, 13, 21, 41–43]. These ConvNets take image pixels as input and predict labels in a predefined taxonomy, and have enjoyed significant success with unparalleled predictive power and generalizability to unseen data. However, classifiers that rely solely on pixel information are often unable to recognize the context of an image, leading to misinformed predictions, a lack of interpretability, and an inability to generate meaningful outcomes for labels associated with few training examples.

Consider the example image in Figure 1, which is obtained from the ImageNet dataset [35]. The image resembles a salt shaker as well as a milk can, both of which are labels in the ImageNet taxonomy. As the image has the same shape and structure as a salt shaker, networks trained purely with pixel information often produce the incorrect label “salt shaker.” Now, suppose we are told that the image is related to the phrase “cartoon cow and milk” (see Figure 2)—this phrase certainly hints that the image is more likely to be related to the label “milk can” than the label “salt shaker.” While such textual information (e.g. text surrounding an image) can provide valuable orthogonal signals, the typical setting in image understanding literature assumes the unavailability of text and focuses on models that rely purely on pixels. In this work, we also focus on the setting where only pixels are available in the input.



**Figure 1: Traditional convolutional network misclassification example.** Inception v3 [43] produces the label “salt shaker,” while using web-sourced textual information more readily points to the correct outcome “milk can.”

However, we show that if we *predict* text from pixels, we can then use the predicted text to train models that lead to better overall performance.

But how do we obtain contextual text if we are only given the pixels for the image? While images are sometimes associated with existing surrounding text (e.g. captions), such contextual information is unavailable in most cases. This is particularly true given the large number of images captured from video frames and photographs taken using smartphones. Furthermore, even when surrounding text is available, the text rarely deterministically maps to the target taxonomy of the classification problem (e.g. “cartoon cow and milk” does not trivially map to the “milk can” label). In this paper, we propose a multimodal framework that can infer contextual information from image pixels to address these two challenges. To address the absence of surrounding text, we present a *phrase generation model* capable of predicting contextual phrases (such as “cartoon cow and milk”) exclusively from image pixels. The image pixels and the textual phrases are taken together as input to a multimodal image classifier for predicting labels in the target taxonomy.

This ability to generate text from images is at the crux of our approach: the generated phrases allow our framework to “bridge the gap” between visual (input) and textual modalities. The generation of suitable phrases is challenging, however, as generated phrases should have a rich vocabulary that expresses the fine-grained content of an image independent of target taxonomy labels. Although image-to-text approaches such as image captioning [33, 50, 51] and text retrieval from similar images currently exist, they cannot be directly applied as a phrase generator. This is because image captioning methods rely on an underlying image classification model, limiting their vocabulary to the taxonomy used to train the model. Furthermore, text retrieval based methods require carefully designed text aggregation and cleaning to infer context in an image, and such methods are not scalable at inference time.

We propose to utilize web-based query-image pairs to train a phrase generation model. Specifically, we use 260 million images and 40 million unique queries, in which the query is treated as the class annotating the image, and we apply ResNet [13] to train the phrase generation model. We further apply techniques inspired by prototypical networks [39] to increase the size of the phrase generation model output space. We subsequently reformulate the

query prediction problem as a nearest neighbor search problem, for which we adopt quantization techniques [12, 48].

The resulting multimodal classification framework, where the predicted phrase from the phrase generation model is used together with image pixels to classify input images, is illustrated in Figure 2. Given an image of a milk can, the phrase generation model produces phrases such as “cartoon cow and milk” and “cow with milk” from its large output space. These phrases are in turn fed into a model that produces a textual embedding. In parallel, the image is input to a convolutional network whose bottleneck layer produces a visual embedding. Finally, the visual and textual embeddings are input to a multimodal model that predicts labels in the target taxonomy.

Notice that our framework is able to produce a textual embedding although the input to the framework consists solely of image pixels. The textual embedding of the phrase generation model’s output is necessary as the phrase generator does not necessarily consider semantic similarity between queries in the textual space. For example, if the target label was “dairy farm” rather than “milk can,” the similarity between those concepts may not be captured in spite of their similarity in semantic textual space. Therefore, we subsequently apply a text embedding model trained with query-query associations to the generated phrases to produce a representation that incorporates orthogonal textual signals and captures additional similarity in the textual space.

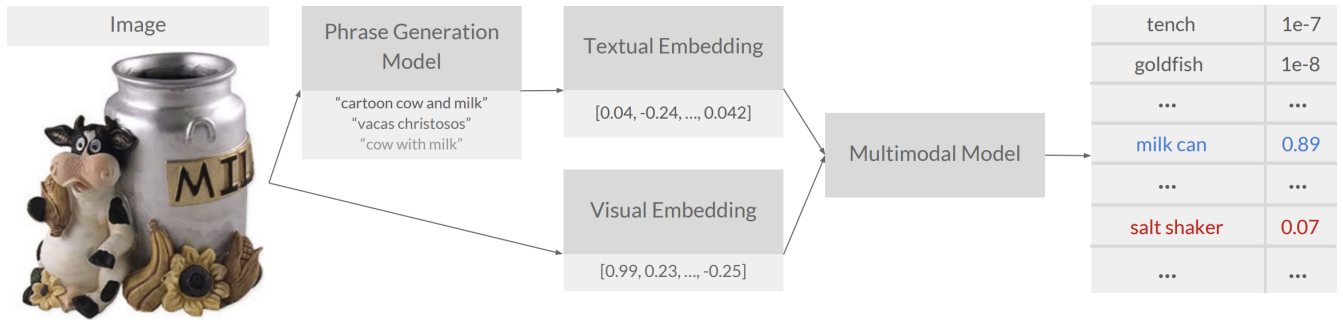
We evaluate the effectiveness of our approach on the WebVision [24] and OpenImages [22] benchmarks, showcasing significant improvements in classification accuracy, particularly in cases of few-shot learning of minimally labeled concepts. Furthermore, we demonstrate that the inclusion of textual information improves interpretability of the classification outcomes. In summary, the contributions of our work include the following:

- A multimodal image classification framework combining *predicted* textual signals with visual signals for context-aware prediction. Our model is distinguished from other multimodal frameworks in its ability to bridge the gap between visual and textual modalities via the prediction of text, even when only pixels are provided for the image.
- A phrase generation model trained on web query-image pairs that takes arbitrary images as input. In addition to producing output contextual signals using input images alone, this model facilitates the efficient extraction of text-based information from input images for multimodal prediction.
- An analysis of the viability of our model across diverse benchmark datasets, showcasing improvements in classification performance, model interpretability, and few-shot learning. We additionally present, to our knowledge, the best reported result on the WebVision dataset, further bolstering the effectiveness of our approach.

The rest of the paper is organized as follows. In Section 2, we discuss related work. In Section 3, we describe the proposed framework and methodology. We provide an experimental evaluation of the approach in Section 4, and conclude in Section 5.

## 2 RELATED WORK

**Representation Learning from Images.** The enormous progress of convolutional networks has demonstrated the effectiveness of



**Figure 2: Illustration of our proposed pipeline flow on Figure 1, with outputs of intermediate models represented below model names. Note that the generation and embedding of contextual phrases containing “milk” and “cow” aid the multimodal model in producing the correct classification.**

representation learning from large-scale image datasets [22, 24, 35]. Ever since the pioneering work by Krizhevsky et al. [21] on ImageNet, there have been numerous efforts to make ConvNets deeper [38, 42], more accurate [13, 41] and faster [14, 36]. Recently, there have been attempts to automate the neural network architecture design process [25, 32, 52] using reinforcement learning. Once image representations are learned, they can be “transferred” to other related perception tasks including object detection [10, 16, 34] and semantic segmentation [3, 4].

In this paper, we use Inception v3 [43] for image representation learning and residual networks [13] for the phrase generation model. However, our proposed framework does not make assumptions about the underlying ConvNet, which can easily be replaced with the latest architectures.

**Learning from Web Data.** We are not the first to resort to large-scale web data for visual recognition tasks. For example, there is a line of work which aims to learn image representations from web images [5, 6, 8] with the goal of using noisy web tags (e.g. from the YFCC-100M [45] dataset) to conduct supervised training on web images. However, the learned representations from such models perform worse than ImageNet pre-trained counterparts when applied on target tasks. Recently, it has been shown that scaling up to 300 million [40] or even 1 billion [28] images allows the learned representations to outperform their ImageNet counterpart when trained from noisy web labels.

Unlike previous work that uses web supervision for pre-training, we propose to use a webly-supervised phrase generator to predict web phrases for arbitrary images and subsequently build a multimodal framework for image classification.

**Multimodal Image Classification.** Multimodal learning is concerned with leveraging information from multiple distinct sources, such as images, text and video [26, 27, 37]. For image classification problems, metadata such as image tags, keywords, and captions have been used as additional textual features to either train ensemble classifiers [47] or generate pseudo labels for training a visual classifier [11]. However, these approaches often require the availability of surrounding textual information at training time (e.g. text captions for an image). Instead of relying on surrounding text, our input consists exclusively of image pixels, and textual signals are generated by a phrase generation model in our framework.

Apart from combining visual and textual signals to train a image classifier, several research directions instead focus on zero-shot learning tasks, in which evaluation classes may not have been seen at training time. Techniques for such tasks leverage curated sources of semantic information for the labels, such as the WordNet hierarchy [29] and Wikipedia [23], as well as a knowledge base containing descriptive properties for each class [30]. Although these approaches have shown promising performance in the zero-shot learning setting, it has been observed [49] that they do not generalize well to all classes. In contrast, our proposed method significantly outperforms the state-of-the-art baseline in scenarios with minimally labeled images.

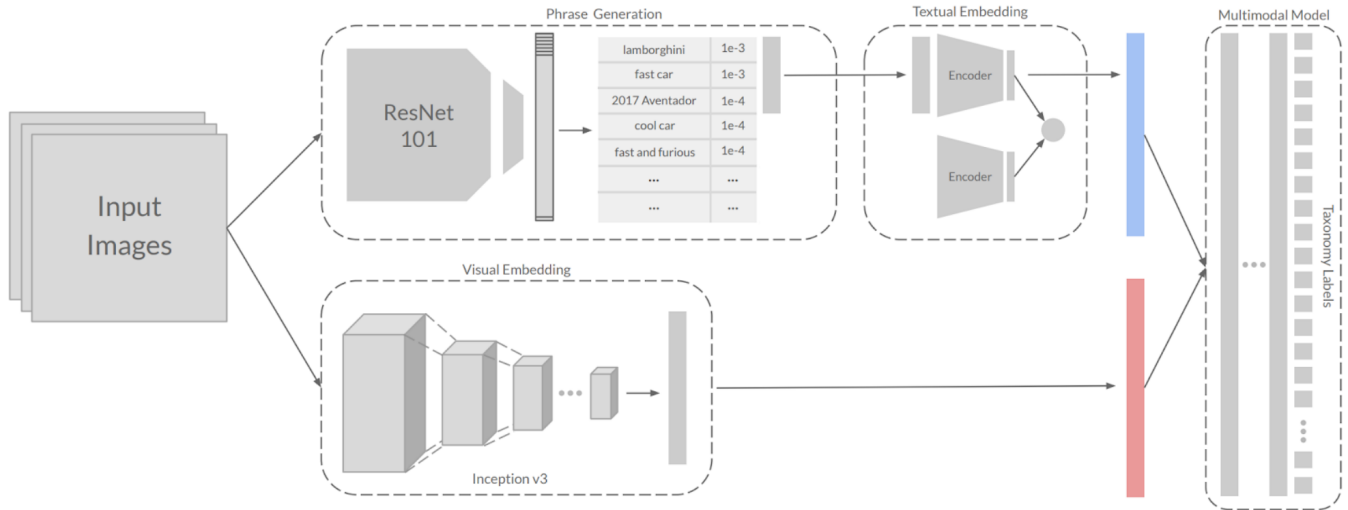
### 3 METHODOLOGY

In this section, we provide a detailed description of our approach. In particular, we give an overview of our framework in Section 3.1. We then describe the phrase generator (Section 3.2), the model to embed text produced by the phrase generator (Section 3.3), the visual embedding model (Section 3.4), and the multimodal model that combines visual and textual signals (Section 3.5).

#### 3.1 Framework Overview

A detailed representation of the inference-time flow of our pipeline can be seen in Figure 3. The input consists of image pixels and the output is a probability distribution over labels from a predefined taxonomy. Such a distribution is produced by a multimodal model that incorporates features from visual and textual embeddings, with the blue (top) vector representing a textual embedding and the red (bottom) vector representing a visual embedding.

The ability to generate text is a critical aspect of our approach as it bridges the gap between visual and textual modalities. In order to produce textual signals, solely given image pixels, we present a phrase generation model that predicts phrases from a large output space. This phrase generator is trained using visual signals (image-query associations), but we subsequently obtain orthogonal textual signals from the generated text by embedding the text using a model trained with query-query associations. As a result, the predicted text incorporates representations learned from two corpora reflecting complementary signals, namely the labeled image dataset and a set of query-query associations. In this way, images that are visually



**Figure 3: Detailed depiction of inference-time pipeline flow for multimodal predictions. Input images are processed using two pipelines, with the former generating and embedding text to produce blue vectors consisting of textual embeddings and the latter embedding pixels to produce red vectors consisting of visual embeddings. These vectors are subsequently combined to generate multimodal representations, which are used as input to the last stage of the pipeline to produce a probability distribution over a desired label taxonomy.**

distinct but textually similar are embedded closer to one another than they would be in models that simply employ image-text co-embeddings, allowing our framework to more effectively represent images. To the best of our knowledge, the proposed framework is the first to generate context-aware predictions for multimodal image classification.

### 3.2 Phrase Generation Model

In this section, we describe the model that generates phrases from image pixels. A key challenge in training this model is that the phrases it generates must capture fine-grained image semantics. Although image captions can be used to provide image semantics, the vocabulary of image captioning models [33, 50, 51] is often limited and their quality relies on pre-trained image classification models. In order to infer rich and diverse context from images, we instead consider using web query-image pairs to train a phrase generation model.

Specifically, our phrase generation model consists of a ResNet-101 architecture [13] trained with a dataset of query-image pairs, which contains approximately 260 million images and 40 million unique queries (used as classes) [19]. The distinctive feature of this network is its ability to extract image semantics in an expressive vocabulary by predicting the most relevant queries from input images. In order to accommodate this large output space, we shrink the ResNet output to a 64-dimensional bottleneck layer. The choice of the dimensionality of the bottleneck layer was driven both by computational constraints as well as the need to construct an efficient nearest-neighbor index based on that layer. Our network can thus be conceptualized as containing three stages: the primary training mechanism consisting of learned weights and biases, a 64-dimensional feature layer representing image embeddings, and a softmax layer that produces a probability distribution across the

40 million queries. Such a decomposition is depicted in the phrase generation component of Figure 3. Since computing the softmax loss over such a large output space is computationally demanding, we use sampled softmax loss, a sampling technique for handling large output spaces that was first introduced in [17].

After training the phrase generation model using query-image pairs, the predicted probability of query  $q_c$  is computed via softmax:

$$p_c = \frac{\exp(-d(\mathbf{w}_c, \mathbf{x}))}{\sum_i \exp(-d(\mathbf{w}_i, \mathbf{x}))} \quad (1)$$

where  $\mathbf{x}$  is the image embedding (from the bottleneck layer in ResNet-101),  $\mathbf{w}_c$  are the weights from the fully connected layer after the bottleneck layer that correspond to the query  $q_c$  among 40 million queries, and  $d(\cdot, \cdot)$  is the distance function. We adopt the cosine distance function for use in experiments.

To further increase the vocabulary of the predicted phrases from the phrase generation model, we adopt techniques from prototypical networks [39] to construct prototypes of a significantly larger number of queries. The prototype  $\phi_c$  of each query  $c$  is defined as the mean vector of the embeddings of the images associated with that query:

$$\phi_c = \frac{1}{|S_c|} \sum_{i \in S_c} \mathbf{x}_i \quad (2)$$

where  $S_c$  is the set of images that are associated with query  $c$ . We can now replace  $\mathbf{w}_c$  by  $\phi_c$  in Equation 1 to increase the output size of the phrase generation model from 40 million queries to a much larger output space. As will be demonstrated in Section 4.5, this provides substantial performance improvements.

In order to integrate the phrase generation model with our multimodal pipeline, we require an efficient way to predict the top related queries for a given image. In fact, most queries in the output

space are irrelevant to the given image, and it is unnecessary to compute their softmax probability to retrieve the top queries. Thus, we can reformulate the classification problem as a top- $k$  nearest neighbor search problem for phrase generation:

$$q_c^* = \arg \min_c (d(\phi_c, \mathbf{x})). \quad (3)$$

Our problem can now be solved by fast similarity search techniques in high-dimensional spaces [12, 48].

### 3.3 Textual Embedding Model

We next describe the model that produces textual embeddings from the phrases generated as per Section 3.2. One could consider using the 64-dimensional embedding produced from visual features directly instead of generating and separately embedding contextual phrases. To motivate the need for a separate textual embedding, consider a taxonomy that contains the term “dairy farm” as a label instead of “milk can.” The object in Figure 1 is an item that can be found in a “dairy farm” and should perhaps be mapped to it. However, the method that we used to generate the embedding in Section 3.2 treats the output labels merely as non-overlapping classes with no regard to semantic similarity they might possess in the text space. Therefore, it is unlikely that the image in Figure 1 will get mapped to “dairy farm” purely based on the produced embedding. On the other hand, if one had access to an additional text embedding model that embeds semantically similar textual concepts close together, then it is likely that the text produced by the image (e.g. “cow with milk”) in Figure 1 will be closer to “dairy farm” in this text embedding space. We take advantage of such a powerful text embedding model that is trained on query-query associations to encode semantic similarity.

Our textual embedding model is based on a dual-encoder as described in [9]. It consists of a Siamese network (as in [15, 44]) pre-trained on a dataset of query-query associations; to improve embedding quality, attention mechanisms described in [31, 46] were additionally used. The resulting embeddings have the critical property that related queries are situated close together in the output embedding space. The embedding model was subsequently utilized to embed the top-3 English text results obtained from the phrase generation model, with phrase embeddings obtained via a bag-of-words model consisting of unigrams and bigrams. The embeddings for the top three phrases were subsequently averaged, resulting in a single 200-dimensional textual feature vector.

Notice that at this stage of the pipeline, we have generated an embedding from a network trained on textual signals (query-query associations) that represents an image for which we were initially given solely pixels as input. The independent textual embedding of the generated phrases thus bridges the gap between the visual and textual modalities, as desired.

### 3.4 Visual Embedding Model

In addition to the generation and extraction of textual embeddings in the prior sections, visual embeddings were obtained from the bottleneck layer of a pre-trained convolutional network as per the visual embedding section in Figure 3. We apply the canonical procedure of training a convolutional network on an input dataset, identifying the bottleneck layer (i.e., the layer prior to outputs),

and extracting the output of that layer [2]. In our case, we use the Inception v3 network trained on the ImageNet dataset.

### 3.5 Multimodal Model

Once both visual and textual embeddings are generated, the last stage of the pipeline is a multimodal model that fuses both signals to produce a final image classification. The input to this model consists of the concatenation of features from the 1024-dimensional visual embedding and the 200-dimensional textual embedding. The resulting vector is input to a multi-stage fully connected neural network, with the output softmax layer predicting probabilities associated with classes according to the target taxonomy. Notice that while we use a rather simple model to demonstrate the benefit of leveraging textual information generated from image pixels, other fusion techniques such as MLB [20] and MUTAN [1] could be easily applied.

In Section 4, we present experiments where multimodal models were trained using the WebVision and OpenImages training sets. In each case, input images are fed at training time through the entire pipeline of Figure 3 to produce the visual and textual embeddings that constitute the input to the multimodal model.

## 4 EXPERIMENTAL EVALUATION

In this section, we compare our work against state-of-the-art methods for image classification using standard benchmarks. We additionally quantify the benefits of our multimodal approach when provided a limited amount of training data, and we further analyze the performance of the phrase generation model.

### 4.1 Experimental Setup

In addition to ImageNet, we employed the following standard benchmarks for evaluation:

- **WebVision.** [24] This dataset contains an imbalanced set of 2.4 million training examples and 50,000 test examples, with blurry images and label ambiguity mimicking the poor quality of real-world data. It uses the same taxonomy as ImageNet, consisting of 1,000 labels, but is much noisier than ImageNet data due to its curation via weak supervision.
- **OpenImages v4.** [22] To the best of our knowledge, OpenImages is the largest image classification benchmark currently available. It consists of 9 million images and 21,000 classes. Unlike WebVision, OpenImages data are labeled for multi-label prediction, where each image may be associated with multiple labels in the ground truth.

We train the multimodal model using training data from the aforementioned benchmarks. Its input, consisting of visual and textual embeddings, is obtained by processing training images with the pipeline in Figure 3. The visual embedding model is pre-trained with data from ImageNet, the phrase generation model is trained with image-query associations, and the textual embedding model is trained with query-query associations.

### 4.2 Ablating Textual Signals

In this section, we validate our hypothesis that generating textual signals as an intermediate artifact leads to better performance. For this purpose, we compare our multimodal framework against a

baseline consisting of an ablation of textual signals from the framework. In particular, the baseline consists of a fully connected neural network with the same architecture as the multimodal model, except for the input layer, which consists only of the visual signals. The visual signals are the same that we use in our framework; that is, pre-trained visual embeddings.

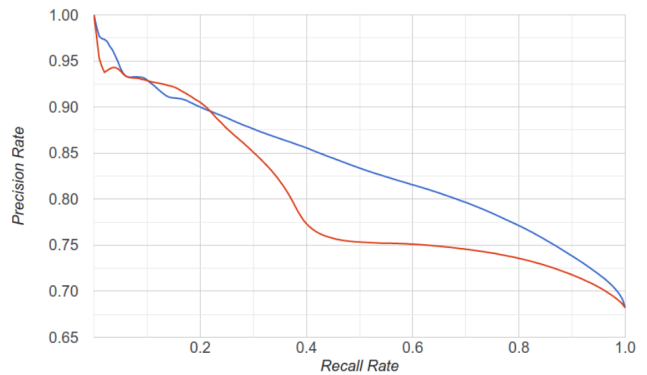
We also validate the hypothesis that the weaker the visual signals, the more effective our approach based on generating textual signals becomes. For this purpose, we conduct experiments where we vary the “strength” of the visual embeddings by controlling the amount of data that is used to pre-train them (using Inception v3 for a visual embedding generator as in Section 3.4).

**4.2.1 Results on the ImageNet benchmark.** We begin by presenting results on the ImageNet dataset. In particular, our framework utilizes textual embeddings and visual embeddings pre-trained with the full ImageNet training dataset, while the baseline consists of a model solely using the visual embeddings. In this setting, our multimodal model achieves a top-1 accuracy of 77.74% and top-5 accuracy of 93.62%, compared to the baseline model top-1 accuracy of 76.93% and top-5 accuracy of 92.84%. The ability of our multimodal model to improve upon the results of a highly optimized, state-of-the-art visual model on ImageNet data is evidence of the merit of our approach.

**4.2.2 Results on the WebVision benchmark.** We next present results on the WebVision dataset. Our framework and the baseline are trained and evaluated using WebVision data, and both use visual embeddings pre-trained with image datasets of varying sizes (subsampling from ImageNet). As evident in Table 1, our multimodal method consistently outperforms the baseline approach, with significant gains on all fractions of ImageNet used for pre-training visual embeddings. Furthermore, improvements become more pronounced when the visual embeddings are “weaker”; that is, when the visual embedding framework is pre-trained with smaller fractions of ImageNet data. For example, if the visual embeddings are pre-trained with 10% of ImageNet, we observe a multimodal gain of 9.43% in top-1 accuracy; if they are trained with 70% of ImageNet, the top-1 accuracy gain is 5.08%.

Even when all of ImageNet is used to pre-train visual embeddings, we continue to see significant gains. In fact, the result that we obtain in this case outperforms the best result reported so far in the literature for the WebVision benchmark (a curriculum-based approach by Jiang et al. [18]). In particular, Jiang et al. report a 72.60% top-1 accuracy when training with WebVision data and pre-training their teacher network with ImageNet. Replicating the same setting (training on WebVision and pre-training our visual embedding model with ImageNet), we obtain a 73.15% top-1 accuracy.

**4.2.3 Results on the OpenImages benchmark.** In addition to conducting multi-class evaluation with WebVision and ImageNet, we performed evaluation in a multi-label setting using the OpenImages benchmark. The setting for OpenImages evaluation is the same as for WebVision, where the baseline consists of an ablated model that uses only visual signals. As we show in Figure 4, our method significantly outperforms the baseline at most points on the precision-recall curve. For example, at 80% precision, we observe a 31.1% improvement in recall (37.2% and 68.3% recall for



**Figure 4: Precision-recall curve representing the efficacy of the multimodal model (blue) and baseline visual-only model (red) on OpenImages hierarchy verticals.**

the baseline and our multimodal model, respectively). Furthermore, on trainable classes (defined in [22] as “classes with at least 100 positive human-verified labels in the train split”), the multimodal model achieves an mAP of 0.741, a 7.2% relative improvement over the baseline model’s mAP of 0.691.

We further analyzed model performance on a per-label basis, observing larger multimodal improvements on labels with fewer training examples. For instance, “fictional character” is the label with the least number of examples (1,069 examples, accounting for just 0.009% of the training set). For this label, the average precision of our method is 0.63 as opposed to 0.58 for the baseline. In other words, we observe a 9% relative improvement over the baseline for this label, as opposed to a 7.2% relative improvement across all trainable classes. These results reinforce our intuition that textual information improves classification when insufficient training data is available for pixel-based classifiers. In the next section, we study this hypothesis in more detail.

### 4.3 Ablating Visual Features and the Multimodal Model via a Zero-Shot Baseline

Given the construction of the phrase generation model, a natural question involves the value of the visual features and multimodal model. In particular, can one obtain a similar performance by using a simpler mapping from the phrase generation model to the target taxonomy? We perform another ablation to quantify the improvement of our approach over such a mapping.

Recall that the phrase generation model produces a 64-dimensional bottleneck layer representing image feature embeddings. These bottleneck layer embeddings are subsequently input to a softmax layer that outputs a probability distribution over 40 million queries. We will refer to these queries as our base query set. Since each column of the weight matrix of the softmax layer represents the queries, the weight matrix yields an implicit co-embedding of textual and visual information.

The aforementioned co-embedding allows us to generate embeddings for images and taxonomy labels in the same space, and suggests the following zero-shot baseline comparison. We generate query embeddings for each label in the taxonomy by looking at

ImageNet Fraction	WebVision (top-1)			WebVision (top-5)		
	Inception v3	Multimodal	Improvement	Inception v3	Multimodal	Improvement
Text Only	-	57.74	-	-	75.09	-
10%	54.41	63.84	+9.43%	76.76	83.80	+7.04%
30%	60.09	67.81	+7.72%	80.87	86.39	+5.52%
50%	62.45	69.41	+6.76%	82.61	87.57	+4.96%
70%	65.94	71.02	+5.08%	85.01	88.72	+3.71%
100%	72.37	73.15	+0.78%	88.83	89.73	+0.90%

**Table 1: Visual embedding quality vs. multi-class accuracy on WebVision. The Inception v3 column represents the performance of our baseline model, using only visual embeddings pre-trained with the specified fraction of ImageNet in the “ImageNet Fraction” column. The Multimodal column represents the performance of our multimodal framework, which utilizes textual embeddings alongside the same pre-trained visual embeddings employed in the corresponding baseline model.**

the terms in the synset corresponding to the label and selecting the embedding corresponding to a term if it is present in our base query set. If the synset is not present, we generate embeddings for the terms with a model trained on the base query set to produce an embedding for arbitrary input text. At inference time, we compute the bottleneck layer embedding of the input image in the phrase generation model and predict the output label to be the one whose term embedding is closest to the image embedding in cosine distance. If there is more than one term in the synset for a label, we consider the embedding that is closest to the inference image embedding.

For evaluation, we used test data from ImageNet. The proposed simple zero-shot framework achieves a top-1 accuracy of 48.29% on the ImageNet validation set, compared to the top-1 accuracy of 76.93% achieved by our multimodal model trained with ImageNet data. These results suggest the necessity of our multimodal framework on top of the phrase generation model in order to achieve state-of-the-art performance.

#### 4.4 Ablating Textual Embeddings

Recall that our framework explicitly generates textual embeddings with an encoder framework described in Section 3.3 after obtaining the predicted queries from the phrase generation model. The phrase generation model produces these queries according to Equation (3), where each prototype  $\phi_c$  represents a query, and the query corresponding to the prototype vector closest to the bottleneck layer embeddings  $x$  in cosine distance is selected as the output phrase. This raises the question of whether the explicit generation of the phrase is necessary and whether working with the bottleneck layer embeddings is sufficient.

To evaluate the benefits of separately embedding textual information instead of utilizing the bottleneck embeddings from the phrase generation model, we trained and compared two supplementary models on the WebVision dataset. The first model takes visual features and the bottleneck embedding from the phrase generation model as input, while the second includes textual embeddings of the generated phrases in addition to visual features and the phrase generator’s bottleneck embedding as input. To diminish the impact of visual features, a fair amount of data (~20% of ImageNet) is used to pre-train the visual features. The first model yielded a

ImageNet Fraction	Small Output Space	Large Output Space
Text Only	47.02	57.74
10%	60.35	63.84
30%	63.05	67.81
50%	63.43	69.41
70%	66.22	71.02
100%	72.90	73.15







**Table 2: Multimodal top-1 accuracies on WebVision with differing phrase generation output spaces.**

top-1 accuracy of 66.90% and the second a top-1 accuracy of 67.80%, indicating that the explicit generation of the text phrases and using their textual embeddings improves model performance by effectively bridging gaps between images that are visually distinct but contextually similar.

#### 4.5 Understanding the Phrase Generation Output Space

Having verified each component of our multimodal model via ablation studies in the previous subsections, we now turn to understanding the implications of the very large output space of the phrase generation model after constructing query prototypes. Recall in Section 3.2 that our initial phrase generation model is trained on a dataset of 260 million images and 40 million unique queries, and its output space is subsequently expanded to a much larger space using techniques derived from prototypical networks. Here, we consider a series of experiments on WebVision and OpenImages to verify the benefits of expanding the phrase generation model output space. In particular, we consider identical experimental setups as in Section 4.2, but we compare two versions of the phrase generation model: the first utilizing the initial output space of 40 million queries, and the second utilizing the expanded output space obtained via computing query prototypes. In doing so, we hope to understand the importance of the scale of the phrase generation output space in the quality of multimodal predictions.

On both benchmark datasets, multimodal models utilizing the phrase generation model with a 40 million query output space produced poorer results. In particular, WebVision results with differing fractions of ImageNet to train the visual model are reported in

Multimodal Correct		Multimodal Incorrect	
 <p>cartoon cow and milk vacas christosos cow with milk</p> <p>"Salt shaker" vs "Milk can"</p>	 <p>keyboard hd image computer keyboard keypad image</p> <p>"Space bar" vs "Keyboard"</p>	 <p>corvette agents of shield 1957 corvette convertible old corvette</p> <p>"Sports car" vs "convertible"</p>	
 <p>masjidil aqsa hd jerusalem mezquita dome of the rock pictures</p> <p>"Dome" vs "Mosque"</p>	 <p>darth mole stuffed animal karate outfit for bear nibbler costume for dogs</p> <p>"Bonnet" vs "Teddy bear"</p>	 <p>g3 vs ak47 ak 47 22 rifle century arms catamount</p> <p>"Rifle" vs "assault rifle"</p>	

**Figure 5: Comparison of multimodal and baseline (Inception v3) predictions on ImageNet. Contextual information extracted as per Section 3.2 is included to the right of each image, and the final predictions of the baseline model (left) versus the multimodal model (right) are presented at bottom, with the correct prediction highlighted in blue.**

Table 2, with the larger phrase generation output space providing significant accuracy boosts regardless of the strength of visual embeddings. On the OpenImages dataset, the multimodal model utilizing a phrase generation model with an output space of 40 million queries achieved an mAP of 0.702, while the multimodal model with a larger phrase generation output space achieved an mAP of 0.741. These substantial improvements suggest that a larger diversity of predicted phrases allows our multimodal model to uniquely identify image nuances that are missed by a visual-only model, cementing the importance of computing query prototypes to expand the output space of our phrase generation model.

#### 4.6 Qualitative Multimodal Prediction Analysis

Finally, to better understand the benefits and drawbacks of our multimodal model, Figure 5 shows anecdotal examples of both successful and erroneous classifications on ImageNet data (as in the experimental evaluation described in Section 4.2.1). In particular, images with a green background were predicted correctly by the multimodal model and incorrectly by the baseline, while images with a red background were incorrectly predicted by the multimodal model and correctly classified by the baseline. While the presented examples are selected from ImageNet data, they are representative of trends observed in all benchmark datasets.

As can be seen in the milk can and mosque pictures in the first column, while the visual structural similarity of the milk can to a salt shaker and the mosque to a large dome fool traditional convolutional networks, the context-aware multimodal model is able to incorporate external information to augment and improve pixel-only predictions. The examples in the second column support such intuition: while black color and structure of the first image is similar to that of a space bar, contextual information such as “keyboard hd image” and “keypad image” aid the multimodal model in recognizing that the image is one of a computer keyboard as opposed to a singular space bar. The generated phrases for the second image (e.g. “karate outfit for bear”) similarly aid in the multimodal model’s correct classification of the image as a teddy bear instead of a bonnet.

Finally, while the multimodal framework yields erroneous results on images in the third column of Figure 5, these results are easily explainable. In particular, the image at the top right of the table does indeed represent a convertible, with the context provided by textual information producing a prediction more precise than the one assigned in the label taxonomy. Similarly, the multimodal model correctly predicted that the rifle at the bottom right of the table represented an assault rifle (although the ground truth label was “rifle”), further emphasizing the additional level of detail provided by the generated contextual phrases.

## 5 CONCLUSIONS

In this work, we develop a multimodal framework that bridges the gap between visual and textual modalities by inferring contextual semantics from images when no text is provided. The developed framework contains two major components: (1) a phrase generation model that infers contextual semantics from image pixels, and (2) a multimodal model that combines the text embedding mapped from the phrase generation model’s output and visual embedding features learned from image pixels to predict labels in the target taxonomy. Training the phrase generation model on 40 million unique queries and expanding its output space by computing query prototypes allows it to express fine-grained image semantics in a rich vocabulary, regardless of the target taxonomy for the multimodal classification problem. Moreover, in order to provide an efficient way to produce contextual signals from an output space of a very large size, we transform the phrase generation problem to a nearest neighbor search problem and solve it by quantization techniques. By fusing the generated textual information with orthogonal visual signals, the proposed multimodal model can better distinguish image content.

In experimental evaluation on the WebVision and OpenImages benchmark datasets, we show that contextual information improves prediction quality and provides significant benefits with weak visual signals. Demonstrations of our results indicate that textual signals further provide promising implications for few-shot learning and zero-shot learning while enhancing model interpretability. We hope to continue pursuing viable approaches to optimize the presented



algorithms in the future, primarily by developing improvements for the multimodal model and by utilizing language-agnostic textual embeddings to further improve predictive performance.

## REFERENCES

- [1] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. 2017. Mutan: Multimodal tucker fusion for visual question answering. In *Proc. IEEE Int. Conf. Comp. Vis.*, Vol. 3.
- [2] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531* (2014).
- [3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. *CoRR abs/1802.02611* (2018). arXiv:1802.02611 <http://arxiv.org/abs/1802.02611>
- [4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 2016. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915* (2016).
- [5] Xinlei Chen and Abhinav Gupta. 2015. Webly supervised learning of convolutional networks. In *ICCV*.
- [6] Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. 2013. Neil: Extracting visual knowledge from web data. In *ICCV*.
- [7] Francois Chollet. 2016. Xception: Deep Learning with Depthwise Separable Convolutions. *arXiv:1610.02357* (2016).
- [8] Santosh Divvala, Ali Farhadi, and Carlos Guestrin. 2014. Learning Everything about Anything: Webly-Supervised Visual Concept Learning. In *CVPR*.
- [9] Daniel Gillick, Alessandro Presta, and Gaurav Singh Tomar. 2018. End-to-End Retrieval in Continuous Space. In *preparation* (2018).
- [10] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2013. Rich feature hierarchies for accurate object detection and semantic segmentation. *arXiv:1311.2524* (2013).
- [11] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. 2010. Multimodal semi-supervised learning for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 902–909.
- [12] Ruiqi Guo, Sanjiv Kumar, Krzysztof Choromanski, and David Simcha. 2016. Quantization based fast inner product search. In *Artificial Intelligence and Statistics*. 482–490.
- [13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*.
- [14] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv:1704.04861* (2017).
- [15] Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. 2014. Convolutional neural network architectures for matching natural language sentences. In *NIPS*.
- [16] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, and Kevin Murphy. 2017. Speed/accuracy trade-offs for modern convolutional object detectors. In *CVPR*.
- [17] S. Jean, K. Cho, R. Memisevic, and Y. Bengio. 2014. On using very large target vocabulary for neural machine translation. In *Proceedings of ACL-ICJNLP*.
- [18] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. 2018. MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning*. 2309–2318.
- [19] Da-Cheng Juan, Chun-Ta Lu, Zhen Li, Futang Peng, Aleksei Timofeev, Yi-Ting Chen, Yaxi Gao, Tom Duerig, Andrew Tomkins, and Sujith Ravi. 2019. Graph-RISE: Graph-Regularized Image Semantic Embedding. (2019). arXiv:1902.10814
- [20] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. 2017. Hadamard product for low-rank bilinear pooling. In *International Conference on Learning Representations*.
- [21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS*.
- [22] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Tom Duerig, et al. 2018. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982* (2018).
- [23] Jimmy Lei Ba, Kevin Swersky, Sanja Fidler, et al. 2015. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *Proceedings of the IEEE International Conference on Computer Vision*. 4247–4255.
- [24] Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. 2017. WebVision Database: Visual Learning and Understanding from Web Data. *arXiv:1708.02862* (2017).
- [25] Chenxi Liu, Barret Zoph, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan L. Yuille, Jonathan Huang, and Kevin Murphy. 2018. Progressive Neural Architecture Search. In *ECCV*.
- [26] Chun-Ta Lu, Lifang He, Hao Ding, Bokai Cao, and Philip S Yu. 2018. Learning from multi-view multi-way data via structural factorization machines. In *Proceedings of the 2018 World Wide Web Conference*. International World Wide Web Conferences Steering Committee, 1593–1602.
- [27] Chun-Ta Lu, Lifang He, Weixiang Shao, Bokai Cao, and Philip S Yu. 2017. Multi-linear factorization machines for multi-task multi-view learning. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM, 701–709.
- [28] Dhruv Mahajan, Ross B. 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 *ECCV*.
- [29] Thomas Mensink, Jakob Verbeek, Florent Perronnin, and Gabriela Csurka. 2012. Metric learning for large scale image classification: Generalizing to new classes at near-zero cost. In *Computer Vision—ECCV 2012*. Springer, 488–501.
- [30] Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. 2009. Zero-shot learning with semantic output codes. In *Advances in neural information processing systems*. 1410–1418.
- [31] Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of EMNLP*.
- [32] Hieu Pham, Melody Y. Guan, Barret Zoph, Quoc V. Le, and Jeff Dean. 2018. Efficient Neural Architecture Search via Parameter Sharing. *arXiv:1802.03268* (2018).
- [33] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. 2016. Learning Deep Representations of Fine-Grained Visual Descriptions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NIPS*.
- [35] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. 2014. ImageNet Large Scale Visual Recognition Challenge. *arXiv:1409.0575* (2014).
- [36] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Inverted Residuals and Linear Bottlenecks: Mobile Networks for Classification, Detection and Segmentation. *arXiv:1801.04381* (2018).
- [37] Weixiang Shao, Lifang He, Chun-Ta Lu, Xiaokai Wei, and S Yu Philip. 2016. Online unsupervised multi-view feature selection. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*. IEEE, 1203–1208.
- [38] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556* (2014).
- [39] Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*. 4077–4087.
- [40] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2017. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. In *ICCV*.
- [41] Christian Szegedy, Sergey Ioffe, and Vincent Vanhoucke. 2016. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *arXiv:1602.07261* (2016).
- [42] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going Deeper with Convolutions. In *CVPR*. <http://arxiv.org/abs/1409.4842>
- [43] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the inception architecture for computer vision. *arXiv:1512.00567* (2015).
- [44] Wen tau Yih, Kristina Toutanova, John C Platt, and Christopher Meek. 2011. Learning discriminative projections for text similarity measures. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*.
- [45] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2015. The New Data and New Challenges in Multimedia Research. *arXiv:1503.01817* (2015).
- [46] Gaurav Singh Tomar, Thyago Duque, Oscar Täckström, Jakob Uszkoreit, and Dipanjan Das. 2017. Neural Paraphrase Identification of Questions with Noisy Pretraining. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*.
- [47] Gang Wang, Derek Hoiem, and David Forsyth. 2009. Building text features for object image classification. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 1367–1374.
- [48] Xiang Wu, Ruiqi Guo, Ananda Theertha Suresh, Sanjiv Kumar, Daniel N Holtmann-Rice, David Simcha, and Felix Yu. 2017. Multiscale quantization for fast similarity search. In *Advances in Neural Information Processing Systems*. 5745–5755.
- [49] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. 2018. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence* (2018).
- [50] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044* (2015).

[51] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image Captioning With Semantic Attention. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[52] Barret Zoph and Quoc V. Le. 2016. Neural Architecture Search with Reinforcement Learning. *arXiv:1611.01578* (2016).