Inferring Context from Pixels for Multimodal Image Classification

Manan Shah\textsuperscript{*}  
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  
zhennli@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.

\textsuperscript{*}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:


1 INTRODUCTION

In recent years, we have witnessed the emergence of ever-improving convolutional network (ConvNet) architectures for image classification \cite{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 misinformation. We further demonstrate that if we predict textual information from pixels, we can subsequently use the predicted text to train models that improve overall performance.

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.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

\textsuperscript{©} 2019 Copyright held by the owner/author(s).


https://doi.org/10.1145/3357384.3357987
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 then 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
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:

\[
p_c = \frac{\exp(-d(w_c, x))}{\sum_i \exp(-d(w_i, x))}
\]

where \( x \) is the image embedding (from the bottleneck layer in ResNet-101), \( w_c \) 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:

\[
\phi_c = \frac{1}{|S_c|} \sum_{i \in S_c} x_i
\]

where \( S_c \) is the set of images that are associated with query \( c \). We can now replace \( w_c \) 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:

$$q^*_c = \arg \min_x d(\phi_c(x)).$$  

(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 (subsampled 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 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
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.

<table>
<thead>
<tr>
<th>ImageNet Fraction</th>
<th>Inception v3</th>
<th>Multimodal</th>
<th>Improvement</th>
<th>Inception v3</th>
<th>Multimodal</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text Only</td>
<td>-</td>
<td>57.74</td>
<td>-</td>
<td>76.76</td>
<td>83.80</td>
<td>+7.04%</td>
</tr>
<tr>
<td>10%</td>
<td>54.41</td>
<td>63.84</td>
<td>+9.43%</td>
<td>80.87</td>
<td>86.39</td>
<td>+5.52%</td>
</tr>
<tr>
<td>30%</td>
<td>60.09</td>
<td>67.81</td>
<td>+7.72%</td>
<td>82.61</td>
<td>87.57</td>
<td>+4.96%</td>
</tr>
<tr>
<td>50%</td>
<td>62.45</td>
<td>69.41</td>
<td>+6.76%</td>
<td>85.01</td>
<td>88.72</td>
<td>+3.71%</td>
</tr>
<tr>
<td>70%</td>
<td>65.94</td>
<td>71.02</td>
<td>+5.08%</td>
<td>88.33</td>
<td>89.73</td>
<td>+0.90%</td>
</tr>
<tr>
<td>100%</td>
<td>72.37</td>
<td>73.15</td>
<td>+0.78%</td>
<td>90.01</td>
<td>91.45</td>
<td>+1.44%</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>ImageNet Fraction</th>
<th>Small Output Space</th>
<th>Large Output Space</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text Only</td>
<td>47.02</td>
<td>57.74</td>
</tr>
<tr>
<td>10%</td>
<td>60.35</td>
<td>63.84</td>
</tr>
<tr>
<td>30%</td>
<td>63.05</td>
<td>67.81</td>
</tr>
<tr>
<td>50%</td>
<td>63.43</td>
<td>69.41</td>
</tr>
<tr>
<td>70%</td>
<td>66.22</td>
<td>71.02</td>
</tr>
<tr>
<td>100%</td>
<td>72.90</td>
<td>73.15</td>
</tr>
</tbody>
</table>

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 ($\sim$20% of ImageNet) is used to pre-train the visual features. The first model yielded a 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
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