DeViSE: A Deep Visual-Semantic Embedding Model

Andrea Frome*, Greg S. Corrado*, Jon Shlens*, Samy Bengio
Jeffrey Dean, Marc’Aurelio Ranzato, Tomas Mikolov
* These authors contributed equally.

{afrome, gcorrado, shlens, bengio}@google.com
{jeff, ranzato, tmikolov}@google.com

Google
Mountain View, CA, USA

Abstract

Modern visual recognition systems are often limited in their ability to scale to large numbers of object categories. This limitation is in part due to the increasing difficulty of acquiring sufficient training data in the form of labeled images as the number of object categories grows. One remedy is to leverage data from other sources—such as text data—both to train visual models and to constrain their predictions. In this paper we present a new deep visual-semantic embedding model trained to identify visual objects using both labeled image data as well as semantic information gleaned from unannotated text. We demonstrate that this model matches state-of-the-art performance on the 1000-class ImageNet object recognition challenge while making more semantically reasonable errors, and also show that the semantic information can be exploited to make predictions about tens of thousands of image labels not observed during training. Semantic knowledge improves such zero-shot predictions by up to 65%, achieving hit rates of up to 10% across thousands of novel labels never seen by the visual model.

1 Introduction

The visual world is populated with a vast number of objects, the most appropriate labeling of which is often ambiguous, task specific, or admits multiple equally correct answers. Yet state-of-the-art vision systems attempt to solve recognition tasks by artificially categorizing objects into a small number of rigidly defined classes. This has led to building labeled image data sets according to these artificial categories and in turn to building visual recognition systems based on N-way discrete classifiers. While growing the number of labels and labeled images has helped improve the utility of visual recognition systems [7], scaling such systems beyond a limited number of discrete categories remains an unsolved problem. This problem is exacerbated by the fact that N-way discrete classifiers treat all labels as disconnected and unrelated, resulting in visual recognition systems that cannot transfer information about learned labels to unseen words or phrases. One way of dealing with this issue is to respect the natural continuity of visual space instead of artificially partitioning it into disjoint categories [17].

We propose an approach that addresses these shortcomings by training a visual recognition model with both labeled images and a comparatively large and independent dataset—semantic information from unannotated text data. This deep visual-semantic embedding model (DeViSE) leverages textual data to learn semantic relationships between labels, and explicitly maps images into a rich semantic embedding space. We show that this model performs comparably to state-of-the-art visual object classifiers when trained and evaluated on flat 1-of-N metrics, while simultaneously making fewer semantically unreasonable mistakes along the way. Furthermore, we show that the model leverages
the learned visual and semantic similarity to correctly predict object category labels for unseen categories (i.e. “zero-shot” learning) – even when the number of unseen visual categories is 20,000 for a model initially trained on just 1000 categories. We argue that this architecture offers several benefits including (1) seamless fine-to-coarse generalization, (2) natural compatibility with large un-manicured datasets, and (3) innate scalability to very large, highly overlapping, and dynamic label sets.

2 Previous Work

The current state-of-the-art approach to image classification is a deep convolutional neural network trained with a softmax output layer (i.e. multinomial logistic regression) that has as many units as the number of classes (see, for instance [11]). However, as the number of classes grows, the distinction between classes blurs, and it becomes increasingly difficult to obtain sufficient numbers of training images for rare concepts.

One solution to this problem, termed WSABIE [17], is to train a joint embedding model of both images and labels, by employing an online learning-to-rank algorithm. The proposed model contained two sets of parameters (1) a linear mapping from image features to the joint embedding space, and (2) an embedding vector for each possible label. Compared to the proposed approach, WSABIE only explored linear mappings from image features to the embedding space, and the available labels were only those provided in the image training set. It could thus not generalize to new classes.

More recently, Socher et al [15] presented a model for zero-shot learning where a deep neural network was first trained in an unsupervised manner from many images in order to obtain a rich image representation [3]; in parallel, a neural network language model [2] was trained in order to obtain embedding representations for thousands of common terms. The authors trained a linear mapping between the image representations and the word embeddings representing a few classes (8) for which they had labeled images, thus linking the image representation space to the embedding space. This last step was performed using a mean-squared error criterion. They also trained a simple model to determine if a given image was from any of the 8 original classes or not (i.e., an outlier detector). When the model determined an image to be in the set of 8 classes, a separately trained softmax model was used to perform the 8-way classification; otherwise the model predicted the nearest class in the embedding space (in their setting, only 2 outlier classes were considered). Their model differs from our proposed approach in several ways: first and foremost, the scale, as our model considers 1,000 known classes for the image model and up to 20,000 unknown classes, instead of respectively 8 and 2; second, by using a different visual model, different language model, and different training objective, we were able to train a unified model that uses only embeddings without the need for a separate softmax model nor an outlier detector.

Another attempt at zero-shot learning was presented in [14], where the authors proposed a few methods to cope with images of classes that were unknown at training time. All of them rely on the availability of a provided hierarchy among classes, including the unknown ones (in their case, WordNet) in order to derive either examples or attributes from surrounding classes. By contrast, our proposed approach does not require a manually constructed class hierarchy to reason about unseen classes; it instead learns an appropriate semantic representation directly from unannotated data.

3 Proposed Approach

Our objective is to leverage semantic knowledge learned in the text domain, and transfer it to a model trained for visual object recognition. We begin by pre-training a simple neural language model well-suited for learning semantically-meaningful, dense vector representations of words [12]. In parallel, we pre-train a state-of-the-art deep neural network for visual object recognition [11], complete with a traditional softmax output layer. We then construct a deep visual-semantic model by taking the lower layers of the pre-trained visual object recognition network and re-training them to predict the vector representation of the image label text as learned by the language model. These three training phases are detailed below.
3.1 Language Model Pre-training

The skip-gram text modeling architecture introduced by Mikolov et al. [12] has been shown to efficiently learn semantically-meaningful floating point representations of words from unannotated text corpora. The model learns to represent a word (or phrase) as a fixed length embedding vector by predicting adjacent words in the document from this vector representation (Figure 1a, right). We call these vector representations embedding vectors. Because synonyms tend to appear in similar contexts, this simple objective function drives the model to learn similar embedding vectors for semantically related words.

We trained such a skip-gram text model on a text corpus of 5.7 million documents (5.4 billion words) extracted from a popular online encyclopedia (wikipedia.org). The text of the webpages was tokenized into a lexicon of roughly 155,000 single-word and multi-word terms. The lexicon of terms was selected from common English words and phrases as well as terms used in commonly used visual object recognition datasets [7]. Our skip-gram model used a hierarchical softmax layer for predicting adjacent terms [13, 12]. Adjacent terms were chosen within a window of 20 terms before or after the source term, where nearby terms were sampled more frequently than distant terms. The model was trained in a single pass through the corpus with a linearly decaying learning rate and no further regularization.

We trained skip-gram models of varying hidden dimensions, ranging from 100-D to 2000-D, and found 500- and 1000-D embeddings to be a good compromise between training speed, semantic quality, and the ultimate performance of the deep visual-semantic embedding model described below. The semantic quality of the embedding representations learned by these models is impressive. For example, the ten nearest terms to “tiger shark” (as measured by cosine similarity between vectors) are “bull shark”, “blacktip shark”, “shark”, “oceanic whitetip shark”, “sandbar shark”, “dusky shark”, “blue shark”, “requiem shark”, and “great white shark”. The ten nearest terms to “car” are “cars”, “muscle car”, “sports car”, “compact car”, “automobile”, “racing car”, “pickup truck”, “dealership”, “sedans”, and “passenger car”. A visualization of the language embedding space over a subset of the training image labels indicates that the language model learned a rich semantic structure that could be exploited in vision tasks (Figure 1b).

3.2 Visual Model Pre-training

The visual model architecture we employ is based on the winning model for the ImageNet Large Scale Visual Recognition Challenge 2012 [11, 6]. The deep neural network model consists of several convolutional filtering, local contrast normalization, and max-pooling layers, followed by several fully connected neural network layers trained using the dropout regularization technique [10]. We trained this model with a softmax output layer, as described in [11], to predict one of a 1000 object categories in the ImageNet 2012 1K Challenge dataset [7], and were able to reproduce their results. This trained model serves both as our benchmark for performance comparisons, as well as the initialization for our joint model.
3.3 Deep Visual-Semantic Embedding Model

Our deep visual-semantic embedding model is initialized from these two pre-trained neural network models (Figure 1a). The embedding vectors learned by the language model are unit normed and used to map label terms into target vector representations. The core visual model, with its softmax prediction layer now removed, is trained to predict these vectors for each image, by means of a projection layer and a similarity metric. The projection layer is necessary because the dimensionality of the topmost layer of the visual network may not match the dimensionality of the semantic label space. In our case we used a trainable linear transformation layer to map the 4096 dimensional representation at the top of our core visual model into the 1000 dimensional representation native to our language model.

The choice of similarity function proved to be important. We found that training the model to minimize the mean-squared difference between visual network output and the target label vector, as suggested by Socher et al [15], produced poor results. We achieved much better results with a combination of dot-product similarity and hinge rank loss (similar to [17]). In particular, the model was trained to produce a higher dot-product similarity between the visual model output and the vector representation of the correct label than between the visual output and other randomly chosen text terms. We defined the per training example hinge rank loss:

\[
\text{loss}(\text{image}, \text{label}) = \sum_{j \neq \text{label}} \max[0, \text{margin} - \tilde{r}_{\text{label}} M \bar{v}(\text{image}) + \tilde{r}_j M \bar{v}(\text{image})]
\]

where \( \bar{v}(\text{image}) \) is a column vector denoting the output of the top layer of our core visual network for the given image, \( M \) is the matrix of trainable parameters in the linear transformation layer, \( \tilde{r}_{\text{label}} \) is a row vector denoting learned embedding vector for the provided text label, and \( \tilde{r}_j \) are the embeddings of other text terms. In practice, we found that it was expedient to randomize the algorithm both by (1) restricting the set of false text terms to possible image labels, and (2) truncating the sum after the first margin-violating false term was encountered. The \( \tilde{r} \) vectors were constrained to be unit norm, and a fixed \text{margin} of 0.1 was used in all experiments.

The DeViSE model was trained by asynchronous stochastic gradient descent on a distributed computing platform similar to [4]. As above, the model was presented only with images drawn from the ImageNet 2012 1K Challenge training set, but now trained to predict the term strings as text\(^1\). The parameters of the projection layer \( M \) were first trained while holding both the core visual model and the text representation fixed. In the later stages of training the derivative of the loss function was back-propagated into the core visual model to fine-tune its output\(^2\). Adagrad per-parameter dynamic learning rates were utilized to keep gradients well scaled at the different layers of the network [9].

At test time, when a new image arrives, one first computes its vector representation using the visual model and the transformation layer; then one needs to look for the nearest labels in the embedding space; this last step can be done efficiently using either a tree or a hashing technique, in order to be faster than the naive linear search approach (see for instance [1]). The nearest labels are then mapped back to ImageNet synsets for scoring (see Supplementary Materials for details).

4 Results

The goals of this work are to develop a vision model that makes semantically relevant predictions even when it makes errors and generalizes to classes outside of its labeled training set, so called zero-shot learning. We compare our DeViSE model to two models that employ the same high-quality core vision model, but lack the semantic structure imparted by our language model: (1) a softmax baseline model – a state-of-the-art vision model [11] which employs a 1000-way softmax classifier; (2) a random embedding model – a version of our model that uses random unit-norm

---

\(^1\)ImageNet image labels are synsets, a set of synonymous terms, where each term is a word or phrase. We found training the model to predict the first term in each synset to be sufficient, but sampling from the synset terms might work equally well.

\(^2\)In principle the gradients can also be back-propagated into the vector representations of the text labels. In this case, the language model should continue to train simultaneously in order to maintain the global semantic structure over all terms in the vocabulary.

---

4
embedding vectors in place of those learned by the language model. Both use the trained visual model described in Section 3.2.

In order to demonstrate parity with the softmax baseline on the most commonly-reported metric, we compute “flat” hit@$k$ metrics – the percentage of test images for which the model returns the one true label in its top $k$ predictions. To measure the semantic quality of predictions beyond the true label, we employ a hierarchical precision@$k$ metric based on the label hierarchy provided with the ImageNet image repository [7]. In particular, for each true label and value of $k$, we generate a ground truth list from the semantic hierarchy, and compute a per-example precision equal to the fraction of the model’s $k$ predictions that overlap with the ground truth list. We report mean precision across the test set. Detailed descriptions of the generation of the ground truth lists, the hierarchical scoring metric, and train/validation/test dataset splits are provided in the Supplementary Materials.

4.1 ImageNet 2012 1K Results

This section presents flat and hierarchical results on the ImageNet 2012 1K dataset, where the classes of the examples presented at test time are the same as those used for training. Table 4.1 shows results for the DeViSE model for 500- and 1000-dimensional skip-gram models compared to the random embedding and softmax baseline models, on both the flat and hierarchical metrics.

<table>
<thead>
<tr>
<th>Model type</th>
<th>dim</th>
<th>Flat hit@$k$ (%)</th>
<th>Hierarchical precision@$k$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Softmax baseline</td>
<td>N/A</td>
<td>55.6</td>
<td>67.4</td>
</tr>
<tr>
<td>DeViSE</td>
<td>500</td>
<td>55.2</td>
<td>65.2</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>54.9</td>
<td>66.9</td>
</tr>
<tr>
<td>Random embeddings</td>
<td>500</td>
<td>52.4</td>
<td>63.9</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>50.5</td>
<td>62.2</td>
</tr>
<tr>
<td>Chance</td>
<td>N/A</td>
<td>0.1</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Table 1: ImageNet 2012 1K test set, using both flat and hierarchical metrics. This compares the DeViSE model for 500-D and 1000-D language models against the softmax baseline and a model using random embeddings. Note that hierarchical precision@$1$ is equivalent to flat hit@$1$.

On the flat metric, the softmax baseline shows higher accuracy for $k = 1, 2$. At $k = 5, 10$, the 1000-D joint vision-language model has reached parity, and at $k = 20$ (not shown) the 1000-D joint model performs slightly better. We expected the softmax model to be the best performing model on the flat metric, given that its cross-entropy training objective is most well matched to the evaluation metric, and are surprised that the performance of the DeViSE model is so close to softmax performance.

On the hierarchical metric, the DeViSE models show better semantic generalization than the softmax baseline, especially for larger $k$. At $k = 5$, the 500-D joint vision-language model shows a 3% relative improvement over the softmax baseline, and at $k = 20$ almost a 7% relative improvement. This is a surprisingly large gain, considering that the softmax baseline is a reproduction of the best published model on these data. The gap that exists between the DeViSE model and softmax baseline on the hierarchical metric showcases the value of semantic information that goes beyond simple visual similarity [8]. The gap between the DeViSE model and the random embeddings model establishes that the source of the gain is indeed the well-structured embeddings learned by our language model, and not some other property of our architecture.

4.2 Generalization and Zero-Shot Learning

A distinct advantage of our model is its ability to make reasonable inferences about candidate labels it has never visually observed. For example, a DeViSE model that was trained on images with labels like “tiger shark”, “bull shark”, and “blue shark”, but never with images labeled simply “shark”, would likely have the ability to generalize to this more coarse-grained descriptor because the lan-

---

3Note that our softmax baseline results differ from the results in [11] due to a simplification in the evaluation procedure: [11] creates several distorted versions of each test image and aggregates the results for a final label, whereas in our experiments, we evaluate using only the original test image. Our softmax baseline is able to reproduce the performance of the model in [11] when evaluated with the same procedure.
To test this hypothesis, we extracted images from the ImageNet 2011 21K dataset with labels that were not included in the ImageNet 2012 1K dataset on which our model was trained. We termed these datasets zero-shot data, since our model has no visual knowledge of these labels. The softmax baseline is only able to predict labels from ImageNet 2012 1K, whereas DeViSE predicts labels from both ImageNet 2012 1K and the zero-shot label set.

Figure 2 shows label predictions for a handful of selected examples from this dataset to qualitatively illustrate model behavior. The middle column lists the top 5 predictions of our model for the image at the left. Correct predictions are highlighted in green. Note that the model successfully predicts a wide range of labels outside the its training set. Furthermore, note that the incorrect predictions are generally semantically “close” to the desired label. The right column of the figure shows the predictions of the softmax baseline model for the same example images. Figure 2 (a), (b), (c), and

Figure 2: Zero-shot predictions of the joint semantic visual model and a vision model trained on ImageNet 2012 1K. Predictions ordered by decreasing score. Correct predictions labeled in green. Ground truth: (a) telephoto lens, zoom lens; (b) English horn, cor anglais; (c) babbler, cackler; (d) pineapple, pineapple plant, Ananas comosus; (e) salad bar; (f) spacecraft, ballistic capsule, space vehicle.
Table 2: Zero Shot Learning, flat hit@$k$ metric. Performance of our model on ImageNet-based zero-shot datasets of increasing difficulty from top to bottom. Note that none of the labels in these test sets occurred in the visual training set. For comparison, we also show chance performance for each test set and value of $k$.

(d) show cases where our model makes significantly better top-5 predictions than the softmax-based model. For example, in Figure 2 (a), the DeViSE model is able to predict a number of lens-related labels even though it has never been trained on images in any of the predicted categories. Figure 2 (d) illustrates a case where the top softmax prediction is quite good, but where it is unable to generalize to new labels, and where its remaining predictions are off the mark, while our model’s predictions are more plausible. Figure 2 (e) highlights a case where although neither model gets the exact true label, both models are giving plausible or related labels. Figure 2 (f) shows a case where the softmax model emits more nearly correct labels than the DeViSE model.

To quantify the performance of the model on zero-shot data, we constructed three datasets of labeled images of increasing difficulty. The first dataset consisted of images from ImageNet 2011 21K set whose labels are visually and semantically similar to the training images in the ImageNet 2012 1K set. 1589 labels were selected for this dataset by identifying all ImageNet 2011 21K labels within 2 tree hops of the ImageNet 2012 1K labels on the combined ImageNet label hierarchy [7]. We termed this the “2-hops from ImageNet 1K” dataset. A more difficult “3-hops from ImageNet 1K” dataset was constructed in the same manner. Finally, we built a third, particularly challenging dataset consisting of all the labels in ImageNet 2011 21K that are not part of ImageNet 2012 1K.

We calculated the flat hit@$k$ measure employed in Section 4.1 to measure how frequently the joint semantic visual model predicted the correct label for each of these data sets. The joint semantic visual model predicted the correct label as its top prediction 0.8% of the time across 1589 novel labels. Given 5 and 20 predictions per image, the model produced the correct label as one of its predictions 7.9% and 22.7% of the time, respectively. For the the more difficult dataset (7860 novel labels), the model predicted 3.4% and 9.7% of the labels correctly in 5 and 20 predictions, respectively. Finally, for the most challenging dataset (20,900 novel labels), the model predicted 1.9% and 5.3% of the labels given 5 and 20 predictions, respectively. At best these predictions rates are almost two orders of magnitude above chance performance. Since a traditional softmax visual model can never produce the correct label on zero-shot data, its performance would be 0% for all $k$.

To provide a stronger baseline for comparison, we compared the performance of our model and the softmax model on the hierarchical metric we employed above (see Supplementary Materials for details). Although the softmax baseline model can never predict exactly the correct label, the hierarchical metric will give the model credit for predicting labels that are in the neighborhood of the correct label in the ImageNet hierarchy (at least for $k > 1$). The softmax model can thus leverage the object’s visual similarity to labels it has been trained on to make predictions that will be scored favorably (e.g. Figure 2d). Visual similarity is strongly correlated with semantic similarity for nearby object categories [8]. That said, because the softmax model does not directly leverage semantic similarity, the errors in its predictions can be surprisingly unreasonable in some cases (e.g. Figure 2c).

The easiest dataset, “2-hops from ImageNet 1K”, contains object categories that are as visually and semantically similar to the training set as possible. For this dataset the softmax model actually outperforms the DeViSE model for hierarchical precision@2, demonstrating just how large a role visual similarity plays in predicting labels “nearby” semantic labels. However, for $k = 5, 10, 20$, our model produces superior predictions relative to the ImageNet hierarchy, even on this easiest dataset. For the two more difficult datasets, where there are both more novel categories and the novel categories are less closely related to those in the training data set, our joint semantic visual model outperforms the softmax model at all measured hierarchical precisions. The gains can be quite
Hierarchical precision@k

<table>
<thead>
<tr>
<th>Label Set and Model</th>
<th>Hierarchical precision@k</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>2-hops from ImageNet 1K</td>
<td></td>
</tr>
<tr>
<td>Softmax baseline</td>
<td>0.000</td>
</tr>
<tr>
<td>DeViSE</td>
<td>0.008</td>
</tr>
<tr>
<td>Chance</td>
<td>0.0004</td>
</tr>
<tr>
<td>3-hops from ImageNet 1K</td>
<td></td>
</tr>
<tr>
<td>Softmax baseline</td>
<td>0.000</td>
</tr>
<tr>
<td>DeViSE</td>
<td>0.005</td>
</tr>
<tr>
<td>Chance</td>
<td>0.0001</td>
</tr>
<tr>
<td>All of ImageNet 21K</td>
<td></td>
</tr>
<tr>
<td>Softmax baseline</td>
<td>0.000</td>
</tr>
<tr>
<td>DeViSE</td>
<td>0.003</td>
</tr>
<tr>
<td>Chance</td>
<td>0.00005</td>
</tr>
</tbody>
</table>

Table 3: Relative Performance on Zero-Shot Learning. Performance of the joint semantic visual model compared to a softmax model based on hierarchical precision at k across same datasets in Figure 2. Relative performance show in parenthesis. Note that the softmax model can never directly predict the correct label so the precision@1 = 0 (see text for details).

large, as much as a 65% relative improvement over softmax performance. The random embeddings model we described above performed substantially worse than either of the real models. These results indicate that our architecture succeeds in leveraging the semantic knowledge captured by the language model to make reasonable predictions, even as test images become increasingly dissimilar from those used in the training set.

Taken together, these zero-shot experiments indicate that the DeViSE model can exploit both visual and semantic information to predict novel classes never before observed. Furthermore, the presence of semantic information in the model substantially improves the quality of its predictions, achieving surprisingly good performance on a very large visual knowledge transfer task.

5 Conclusion

In contrast to previous attempts in this area [15], we have shown that our joint semantic visual model can be trained to give performance comparable to a state-of-the-art softmax based model on a flat object classification metric, while simultaneously making more semantically reasonable errors, as indicated by its improved performance on a hierarchical label metric. We have also shown that this model is able to make correct predictions across thousands of previously unseen classes by leveraging semantic knowledge elicited only from unannotated text.

The advantages of this architecture, however, extend beyond the experiments presented here.

First, we believe that our model’s unusual compatibility with larger, less manicured data sets will prove to be a major strength moving forward. In particular, the skip-gram language model we constructed included only a modestly sized vocabulary, and was exposed only to the text of a single online encyclopedia; we believe that the gains available to models with larger vocabularies and trained on vastly larger text corpora will be significant, and easily outstrip methods which rely on manually constructed semantic hierarchies (e.g. [14]). Perhaps more importantly, though here we trained on a curated academic image dataset, our model’s architecture naturally lends itself to being trained on all available images that can be annotated with any text term contained in the (larger) vocabulary. We believe that training massive “open” image datasets of this form will dramatically improve the quality of visual object categorization systems.

Second, we believe that the 1-of-N (and nearly balanced) visual object classification problem is soon to be outmoded by practical visual object categorization systems that can handle very large numbers of labels [5] and the re-definition of valid label sets at test time. For example, our model can be trained once on all available data, and simultaneously used in one application requiring only coarse object categorization (e.g. house, car, pedestrian) and another application requiring fine categorization in a very specialized subset (e.g. Honda Civic, Ferrari F355, Tesla Model-S). Moreover, in a practical object recognition setting, it is more useful to have a system that can capture
that the labels “great white”, “shark”, “bitten”, “fish”, “man-eater”, “Jaws”, “ocean”, and “blue” actually all might be valid labels for a single image, without having had to be explicitly trained on all possible reasonable labelings of every image. Because test time computation can be sub-linear in the number of labels contained in the training set, our model can be used in exactly such systems with much larger numbers of labels, including overlapping or never-observed categories.

References


A Appendix

A.1 Validation/test Data Methodology

For all experiments we train our visual and joint models on the ImageNet 2012 1K data set. Experiments in Section 4.1 use only the 1K set for testing, and the zero-shot experiments in Section 4.2 make use of the 1K as well as the ImageNet 2011 21K data set. The same subset of the ImageNet
2012 1K data used to train the visual model (Section 3.2) is also used to train DeViSE and the random embedding-based models, and when training all models, we randomly flip, crop and translate images to artificially expand the training set several-fold. The remaining 50K images from the ImageNet 2012 1K set that are not used for training are split randomly 10/90 into validation and held-out test sets of 5K and 45K images, respectively. The validation set is used to choose hyperparameters, and results are reported on the held-out set. The 1,000 classes are roughly balanced in the validation and held-out sets. The same validation/held-out split is also applied to the ImageNet 2011 21K data set for zero-shot experiments. At test time, images are center-cropped to $225 \times 225$ for input to the visual model, and no other distortions are applied.

A.2 Mapping Text Terms to ImageNet Synset

The language model represents terms and phrases gathered from unannotated text as embedding vectors, whereas the ImageNet data set represents each class as a synset, a set of synonymous terms, where each term is a word or phrase. When training the joint model, a method is needed for mapping from an ImageNet synset to the target embedding vector, and at prediction time, label predictions from the embedding space must be translated back into ImageNet synsets from the test set for scoring. There are two complications: (1) the same term occurs in multiple ImageNet synsets, often representing different concepts, for example the two synsets consisting only of “crane” in ImageNet 2012 1K; (2) the language model as we have trained it only has one embedding for each word or phrase, so there is only one embedding vector representing both senses of “crane”.

When training, we choose the target embedding vector by mapping the first term of the synset to its embedding vector through the text of the synset term. We found this to work well in practice; other possible approaches are to choose randomly from among the synset terms or take an average of the embedding vectors for the synset terms.

When making a prediction, the mapping from embedded text vectors back to ImageNet synsets is more difficult: each embedding vector can correspond to several ImageNet synsets due to the repetition of terms between synsets, up to 9 different synsets in the case of “head”. Additionally, multiple of our predicted embedding vectors can correspond to terms from the same synset, e.g. “ballroom”, “dance hall”, “dance palace”. In practice, this happens frequently since synonymous terms are embedded close to one another by the skip-gram language model.

For a given visual output vector, our model first finds the $N$ nearest embedding label vectors using cosine similarity, where $N > k$. The first step in converting these to ImageNet synsets is to expand every embedded term to all of its corresponding ImageNet synsets. For example, “crane” would be expanded to the two synsets which contain the term “crane” (the order of the two “crane” synsets in the final list is arbitrary). After expansion, duplicate synsets are collapsed so that each synset only occurs once in the prediction list. Finally, the list is truncated to the top $k$ predictions. We experimented with choosing randomly from among all the possible synsets instead of expanding to all of them and found this to slightly reduce performance in the ImageNet 2012 1K experiments.

A.3 Hierarchical Precision-at-k Metric

We defined the following hierarchical precision-at-k metric, $hp@k$, to assess the accuracy of model predictions with respect to the ImageNet object category hierarchy. For each image in the test set, the model in question emits its top $k$ predicted ImageNet object category labels (synsets). We calculate $hp@k$ as the fraction of these $k$ predictions which are in $hCorrectSet$, averaged across the test examples:

$$hp@k = \frac{1}{N} \sum_{i=1}^{N} \frac{\text{number of model’s top } k \text{ predictions in } hCorrectSet \text{ for image } i}{k}$$

(2)

The $hCorrectSet$ for a true label is constructed by iteratively adding nodes from the ImageNet hierarchy in a widening radius around the true label until $hCorrectSet$ has a size $\geq k$:

```python
hCorrectSet = {}
R = 0
while NumberElements(hCorrectSet < k):
```
Table 4: Mean sizes of $h_{CorrectSet}$ lists used for hierarchical evaluation, averaged across the test examples, shown for various label sets and values of $k$. At $k = 1$, $h_{CorrectSet}$ always contains only the true label. Note that for the zero-shot data sets, $k_{CorrectSet}$ includes the test set labels as well as the ImageNet 2012 1K labels. List sizes vary among test examples depending upon the local topology of the graph around the true label as well as how many labels from the graph are in the ground truth set.

<table>
<thead>
<tr>
<th>Label set</th>
<th># $k_{CorrectSet}$ Labels</th>
<th>$k$</th>
<th>2</th>
<th>5</th>
<th>10</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>ImageNet 2012 1K</td>
<td>1000</td>
<td>6.5</td>
<td>12.5</td>
<td>22.5</td>
<td>41.7</td>
<td></td>
</tr>
<tr>
<td>Zero-shot 2-hop</td>
<td>2589</td>
<td>3.2</td>
<td>16.8</td>
<td>25.5</td>
<td>45.2</td>
<td></td>
</tr>
<tr>
<td>Zero-shot 3-hop</td>
<td>8860</td>
<td>4.4</td>
<td>57.4</td>
<td>635.4</td>
<td>668.0</td>
<td></td>
</tr>
<tr>
<td>Zero-shot ImageNet 2011 21K</td>
<td>21900</td>
<td>5.4</td>
<td>273.3</td>
<td>317.4</td>
<td>350.6</td>
<td></td>
</tr>
</tbody>
</table>

radiusSet = all nodes in the ImageNet hierarchy which are $R$ hops from the true label
validRadiusSet = ValidLabelNodes(radiusSet)
hCorrectSet = Union(hCorrectSet, validRadiusSet)
$R = R + 1$
return $h_{CorrectSet}$

The size of $h_{CorrectSet}$ for a given test example depends on the combination of the structure of the hierarchy around a given label and which classes are included in the test set. It is exactly 1 when $k = 1$ and is rarely equal to $k$ when $k > 1$. The ValidLabelNodes() function allows us to restrict $h_{CorrectSet}$ to any subset of labels in the ImageNet (or larger WordNet) hierarchy. For example, in generating the results in Table 2 we restrict the nodes in the $h_{CorrectSet}$ to be drawn from only those nodes which are both members the ImageNet 2011 21K label set and are three-hops or less from at least one of the ImageNet 2012 1K labels.

Note that this hierarchical metric differs from the hierarchical metric used in some of the earlier ImageNet Challenge competitions. That metric was generally considered to be rather insensitive, and was withdrawn from more recent years of the competition. Our DeViSE model does perform better than the baseline softmax model on that metric as well, but effect sizes are generally much smaller.