The Role of Pre-trained Word Embeddings in Recognition of Unseen Classes

Abstract

A notable characteristic of human cognition is its ability to derive reliable hypotheses even in situations characterized by extreme uncertainty. Even in situations marked by the absence of relevant knowledge to make a correct inference, humans are able to draw upon related knowledge to make an approximate inference that is semantically close to the correct inference. In the context of object recognition, this ability amounts to being able to guess the identity of an object in an image without having ever seen any visual training examples of that object. The paradigm of zero-shot and one-shot classification have been traditionally used to address these situations. However, zero-shot and one-shot approaches entail training classifiers in situations where a majority of classes are seen and a minority of classes are unseen, in which case the classifiers for the unseen classes are learned in terms of the classifiers for the seen classes. In this paper, we address the problem of object recognition in situations where a few object classes are seen classes whereas a majority the object classes are unseen. Specifically, we pose the following questions: (a) Is it possible to guess the identity of an object in an image without having seen any visual training examples for that object?, and (b) Could the visual training examples of a few *seen* object classes provide reliable priors for guessing the identities of objects in an image that belong to the majority *unseen* object classes? In this paper, we propose a model for recognition of objects in an image in situations where visual classifiers are available for only a limited number of object classes. To this end, we leverage word embeddings trained on publicly available text corpora and use them as natural language priors for hypothesizing the identities of objects that belong to the *unseen* classes. Experimental results on the *Microsoft Common Objects in Context* (MS-COCO) dataset show that it is possible to come up with reliable hypotheses with regard to object identities by exploiting word embeddings trained on *Wikipedia* even in the absence of explicit visual classifiers for those object classes.

Introduction

One of the characteristics of human intelligence is to hypothesize about unknown situations even in the face of extreme uncertainty. Even in situations where they do not have relevant knowledge to draw an inference about a situation, humans can still retrieve related knowledge to draw an inference about a situation. In the domain of object recognition, such an ability would amount to being able to guess object in an image without having ever seen the training examples of such an image. The paradigm of zero-shot (Socher et al. 2013) and one-shot classification have been traditionally used to address such situations. However, these approaches entail training classifiers in situations where majority of classes are seen and minority of classes are unseen. Additionally, these approaches learn a classifier in terms of other classifiers. However, what if we are ever in situations where we have few seen classes and majority of unseen classes. Could we guess objects in an image in such situations? Could few seen classes provide reliable priors for guessing many unseen classes in an image? In this paper, we propose a model that correctly identifies objects in an image using a limited number of visual classifiers.

In this paper, we propose a model that correctly identifies objects in an image using a limited number of visual classifiers. The proposed model can correctly identify an object in an image even when the object that the visual classifier is trained to detect is not present in the image. For instance, if the visual classifier indicates that a *car* is the most likely object in an image - even when the image does not contain a car - then using a natural language model that exploits word embeddings obtained via training on a publicly available natural language corpus, alternative hypotheses based on other vehicles and/or transport-related entities, such as *train*, *bus*, or even *road* or *pedestrian*, should be offered. Likewise, if the visual classifier indicates that the image contains a *bird*, then natural language-based priors should allow the system to suggest an *airplane* as an alternative hypothesis. In summary, nonspecific visual classifiers could still prove useful in accurately identifying an actual object in an image, as the most likely object deduced by the visual classifier could belong to the same general category as the actual object in the image. Alternatively, the context of the most likely object deduced by the visual classifier could be similar to that of the actual object in an image, or the most likely object deduced by the visual classifier and actual object in an image could coexist or co-occur in the real world with high probability.

The context of an object is any information that aids in the recognition of that object without explicitly using an object classifier/detector. In this paper, we try to guess an object

in an image without using a classifier trained for that particular object. Hence, we do not use the features extracted from an image to recognize a particular object in an image; instead, we try to use a classifier for another object to guess a particular object in an image. Prior to the deep learning era, various types of contexts were used to aid object recognition, the most prominent were scene-based context and object-object context. Regarding object-object context, for some models, the co-occurrence statistics were learned from text/web data (Rabinovich et al. 2007). After the rise of deep learning-based approaches, object recognition systems made a leap in performance and research into improving object recognition using context slowed down. However, several of the ideas developed could still improve deep learning-based approaches. For example, in this paper, we draw upon object-object context to guess objects in an image in situations where the data for training *all* objects is not available to programmers. In fact, data for training most objects are not available. We address situations in which we have data to train *extremely* few classifiers. For this purpose, we take advantage of natural language priors to guess objects in an image. We work under an assumption that real-world natural language could, to a certain extent, provide priors for visual context. Such natural language priors may not always be accurate because co-occurrences in the natural world do not necessarily reflect co-occurrences in the visual world. However, probabilistically, the natural language priors could still yield a reasonable performance in situations when the situation in an image is completely unknown. In other words, we could get something for nothing. For instance, cars and roads may occur together in the visual domain, but in natural language, people do not typically use the word roads in the proximity of the word car often. In addition, it is necessary to mention that we do not require co-occurrences in the stricter sense with the word embedding models because these models are capable of assigning semantically similar words to nearby space in hypothetical word embedding space. This is because word embedding models are inspired by the distributional hypothesis in which words that share common linguistic context tend to have semantic similarity.

In this paper, to help guess objects in an image, we rely on pretrained word embedding models that were trained on Wikipedia text (Bojanowski et al. 2016). These pretrained embeddings provide a natural language context. Word embedding models such as word2vec (Mikolov 2013), GloVe (Pennington, Socher, and Manning 2014) and FastText (Bojanowski et al. 2016) have gained popularity owing to their flexibility in solving challenging natural language processing problems. These models convert each word to a vector in low dimensional space, and these vectors are shown to have some interesting properties, among them, the tendency for vectors to be in close proximity when they are semantically similar.

In the proposed approach, as a first step, k-means clustering is performed on natural language word embeddings obtained via training on a publicly available natural language text corpus. This results in the formation of object clusters that share some degree of similarity. From each object cluster, a representative object is selected to train the corresponding visual classifier. Typically, the object within the cluster with the greatest number of training instances available is deemed to be the representative object for that cluster. After the visual classifiers for all of the representative objects have been trained, they can be employed on real test images. The visual classifiers for all the representative objects are executed against the test image to determine the object(s) in the test image. Using the most probable object hypothesis and natural language word embeddings obtained via training on a public text corpus, it is possible to refine the initial object hypothesis to propose alternative object hypotheses for the test image. Previously, Sharma et al. (Sharma, Kumar, and Bhandarkar 2016) presented a preliminary object detection and classification model based on the above approach with encouraging results. In their approach, Sharma et al. (Sharma, Kumar, and Bhandarkar 2016) used word embeddings trained on natural language captions that accompany the test images. In this paper, we extend the approach of Sharma et al. (Sharma, Kumar, and Bhandarkar 2016) by exploiting natural language word embeddings (such as FastText) obtained via training on a general text corpus (such as Wikipedia) and perform a more comprehensive analysis of the results.

Motivation

To accurately classify or guess an object in an image, the proposed approach embodies the following observations about the real world.

1. The first observation is that similarity in the visual deep learning feature space for object categories that belong to a more general category often translates to proximity of the corresponding natural language word embeddings in the hypothetical space. For instance, the object categories *orange* and *apple* belong to the general *fruit* category and would be in close proximity in the hypothetical space, as would the objects *car*, *truck* and *train*. Similarly, in the visual world, many object categories that belong to a single general category tend to share the same (or similar) visual deep learning features. Toy Experiment—We conduct a toy experiment on the category *car*. We select a random car image as well as one random image of each annotated object in the MS-COCO dataset (Lin et al. 2014), ensuring that each image has only one annotated object. Using these images, we compute the similarity between the car image and all the other objects in the visual deep learning feature space. For textual similarity, we extract word embedding features from FastText pretrained embeddings pertaining to car and all the other annotated objects in the MS-COCO dataset. Using these extracted vectors, we compare the similarity between the car and other objects. Next, we calculate the correlations between similarities in the visual deep learning space and the word embedding space. The correlation turned out to be 0.4, which can be considered moderate, thus suggesting that there is a moderate relation between the visual and textual data.
2. The second observation is that the general semantic context in the form of object co-occurrences in natural language is effectively captured by word embed-

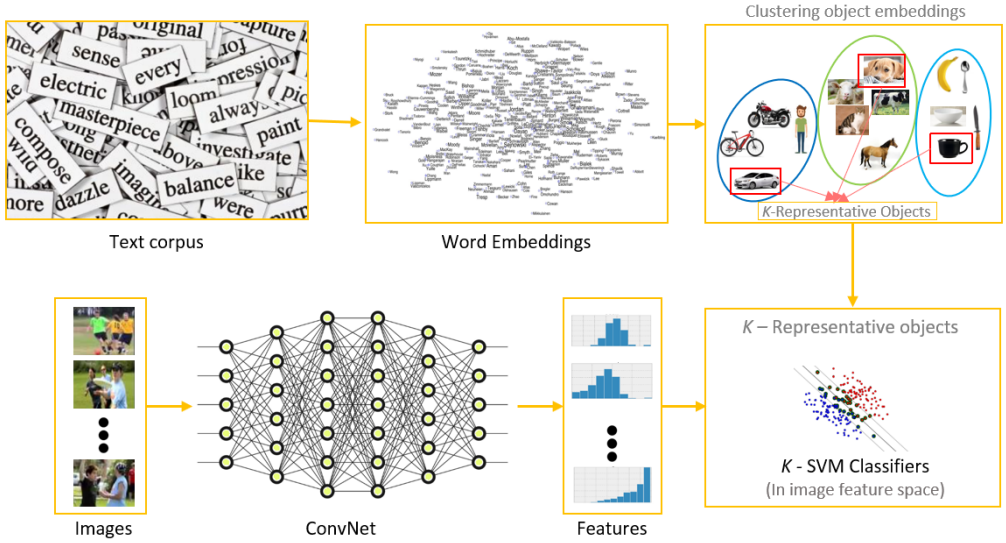

Figure 1: Overview of the proposed approach.

dings in the hypothetical space. Moreover, object co-occurrences in natural language often correspond to object co-occurrences in the visual world. To revisit the previous example, a local search in the hypothetical space will yield alternative object hypotheses, such as *truck* along with other vehicle categories that are in semantic proximity to the word embedding of the object *car*.

3. The third observation is that objects that share the same general context in the visual world often share the same general semantic context in the natural language world. For instance, *bird* and *airplane* both fly in the sky and hence will tend to share similar visual features because of their shared semantic context, such as wings for flying, and these similarities will be reflected in their common visual deep learning features. Toy Experiment - We consider the two categories bird and airplane for the purpose of this toy experiment. Again, we select random images from the MS-COCO dataset for each category while ensuring that only one object occurs in an image. We extract one random image for each category. In the visual deep learning world, the mean similarity of *bird* to all categories is 0.22, and the mean similarity of *airplane* to all categories is 0.26. However, the similarity between airplane and bird is only 0.31. In the textual world (FastText embedding space), the mean similarity between bird and all categories is 0.13, and between airplane and all categories it is 0.16. However, the similarity between bird and airplane is 0.19. Thus, in both the visual and textual worlds, the similarity between airplane and bird is greater than the mean similarities of bird and airplane with other categories. This suggests that one could aid in the detection of another. The effect is not particularly strong, yet it is sufficiently substantial to matter in many real-world situations.

Related Work

In recent years, a wide variety of models for context-based object recognition have been proposed. Divvala et al. (Divvala et al. 2009) identified and categorized various context sources: pixel-level interactions, semantic context, GIST, geographic context, illumination and weather, cultural context, and photogrammetric context, among others. Divvala et al. (Divvala et al. 2009) further demonstrated that incorporation of each type of context leads to moderate improvements in the recognition accuracy. However, two classes of contextual models have gained prominence in recent years, i.e., scene-based contextual models and object-based contextual models (Rabinovich and Belongie 2009). In a scene-based contextual model, the statistics pertaining to the entire scene are used to detect and locate the scene objects, whereas in an object-based contextual model, objects in the spatial vicinity of the target object are used to recognize the target object.

In the Co-occurrence Location and Appearance (CoLA) model, a widely used object-based contextual model proposed by Rabinovich et al. (Rabinovich et al. 2007), bottom-up image segmentation is followed by a bag-of-words-based object recognition system. Additionally, a conditional random field (CRF) is used to capture the inter-object interactions in the dataset. Although capable of capturing obvious reoccurring patterns in the real world, the CRF-based contextual model cannot identify certain subtle patterns that may characterize similar objects. For example, a rear view of car is often encountered in the spatial vicinity of an oblique view of a building, yet this subtlety not captured by a CRF-based contextual model (Malisiewicz and Efros 2009). To address these shortcomings, Malisiewicz and Efros (Malisiewicz and Efros 2009) introduce a visual memex model, which is based on the premise *"Ask not 'what this is, but what this is like'."* In their visual memex model, Malisiewicz and Efros (Malisiewicz and Efros 2009) use a graph-theoretic approach to model real-world images, where similar objects are connected by similarity edges while objects that are contextually related are connected by context edges. Consequently, using the earlier example, different types of buildings are connected by similarity edges, whereas a building and a car are linked via a context edge in the visual memex model (Malisiewicz and Efros 2009). The graph-theoretic model automatically learns the visual memex graph from the input images and is shown to successfully outperform the CoLA model (Rabinovich et al. 2007) on Torralba's context challenge dataset (Torralba).

Heitz and Koller (Heitz and Koller 2008) introduced the Things-and-Stuff (TAS) model, a category-free model that relies on unsupervised learning. In the TAS model, a *thing* is an object that has a concrete shape and size, whereas *stuff* is malleable but has a repetitive pattern and typically contains *things.* For instance, a *car* (thing) is most likely to be found on a *road* (stuff), and likewise a *cow* (thing) on *grass*(stuff). The TAS model is shown to capture regularities not inherent in other contextual models. Another category of contextual models that has gained prominence in recent times is scene-based models. Choi et al. (Choi, Torralba, and Willsky 2012) proposed a scene-based contextual model developed using pre-labeled images by optimizing information derived from GIST features, the relative locations of objects, and a co-occurrence tree. The co-occurrence tree is generated using a hierarchical CRF, where a positive edge denotes object co-occurrence, and a negative edge indicates that the corresponding objects do not occur together. Choi et al. (Choi, Torralba, and Willsky 2012) used a deformable parts model as their baseline detector on the SUN dataset introduced by Xiao et al. (Xiao et al. 2010) and applied it to the output of the baseline detector. They showed that it outperforms the baseline detector when deformable object parts are present in the input image.

More recently, due to the advent of deep learning, several effective attempts have been made to blend context with deep learning methods (Gkioxari 2016). For instance, Sun and Jacobs (Sun and Jacobs 2017) proposed an architecture that can be employed to predict a missing object in an image by exploiting contextual information. Similarly, Zhang et al. (Zhang et al. 2016) integrated 3D context into deep learning for 3D holistic scene understanding, whereas Gonzalez et al. (Gonzalez-Garcia, Modolo, and Ferrari 2017) employed a deep learning model to detect object parts by using object context. One disadvantage of all of the aforementioned contextual models is that a very constrained environment is assumed in the model learning phase, i.e., that pre-labeled images are readily available and that the labeled images reliably capture the co-occurrence patterns. The approach proposed in this work is unique in that - instead of using context to improve object recognition performance - the context itself is guessed based on the output from a small number of object classifiers.

Proposed Approach

Our approach is explained step by step in this section. Briefly, in our approach, we train very few classifiers and guess objects in an image using natural language embeddings. The major advantage of this approach is that it eliminates the need to generate training datasets, which is an expensive process in real-world applications. In addition, in situations when reliable datasets are not available for rare categories, our approach could yield some performance improvements. In passing, an additional advantage could be the need to execute only a few classifiers at test time. The proposed approach is composed of the following steps::

1. Extract the word embeddings corresponding to objects in MS-COCO from FastText (Bojanowski et al. 2016). These word embeddings are low-dimensional vectors for each word corresponding to an object. The FastText word embeddings were trained on Wikipedia using the skip-gram approach; we used the embeddings provided by (Bojanowski et al. 2016). FastText is an enhancement of word2vec (Bojanowski et al. 2016). In our research, we used FastText embeddings pretrained on Wikipedia.

2. Cluster the embeddings obtained in the preceding step using k-means clustering, where the optimal value of k is determined by the performance of the validation set. After k-means clustering, each cluster will encompass a certain number of objects that tend to have particular relationships –they are co-occurring, belong to the same general category, or share the same general context –as explained in the previous section. The optimal value of k is determined by the algorithm's performance on the validation set. We emphasize that all the k-means clustering was performed on word embeddings obtained from natural language public datasets; no visual information was used.

3. From each of the clusters for a particular value of k, an object category that is representative of this cluster is selected. The representative object that is selected for training is the one with the greatest number of instances in the training set in MS-COCO. An analogy could be drawn to the real world where we have unbalanced training sets, with some categories outnumbering others by a significant margin. Thus, our model resembles and is applicable to real-world situations as a proof of concept.

4. Given the test data, run representative object classifiers of all representative object classifiers (aka cluster centers) on an image, and select the most probable object/objects they contain. While the most probable object may not exist, this procedure can still provide useful clues concerning objects that could be present, as explained in Motivation Section.

5. Using these most probable object/objects as the starting point, other objects in an image could be identified using their cosine similarities in the hypothetical word embedding space. In the current implementation, the most probable, the second most likely, and the third most probable detections are utilized for guessing other objects in the image.

In our approach, we use 1 to 3 top-object classifications to guess other objects. We reiterate that the guesses for the objects in an image are made from representative objects selected using the steps explained above. It could be argued that even in the test set of MS-COCO, the distribution of most representative objects remains the same as the training set. However, even in many real-world applications, the distribution of training and testing sets tends to remain the same. Even when the distribution of training and testing sets is different in real-world applications, we believe there is sufficient regularity in the world so that the system inspired by our proof of concept will yield reasonable results.

At the testing stage, object detectors for all clusters centers (i.e., representative objects) are executed on the image. For 3 top object classifications, given a set of nouns N and top object classifications n_1, n_2 and n_3, the closest guesses from set N are given by:

$$\arg\max_i \ \{SIM(n_i, n_1) + SIM(n_i, n_2) + SIM(n_i, n_3)\} \tag{1}$$

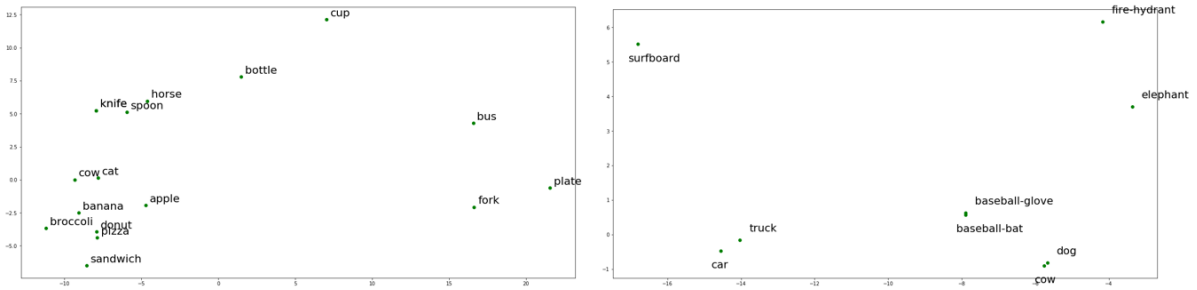

Figure 2: *t-sne* Diagram of word embeddings of MS-COCO annotated objects obtained from Fastext trained on wikipedia.

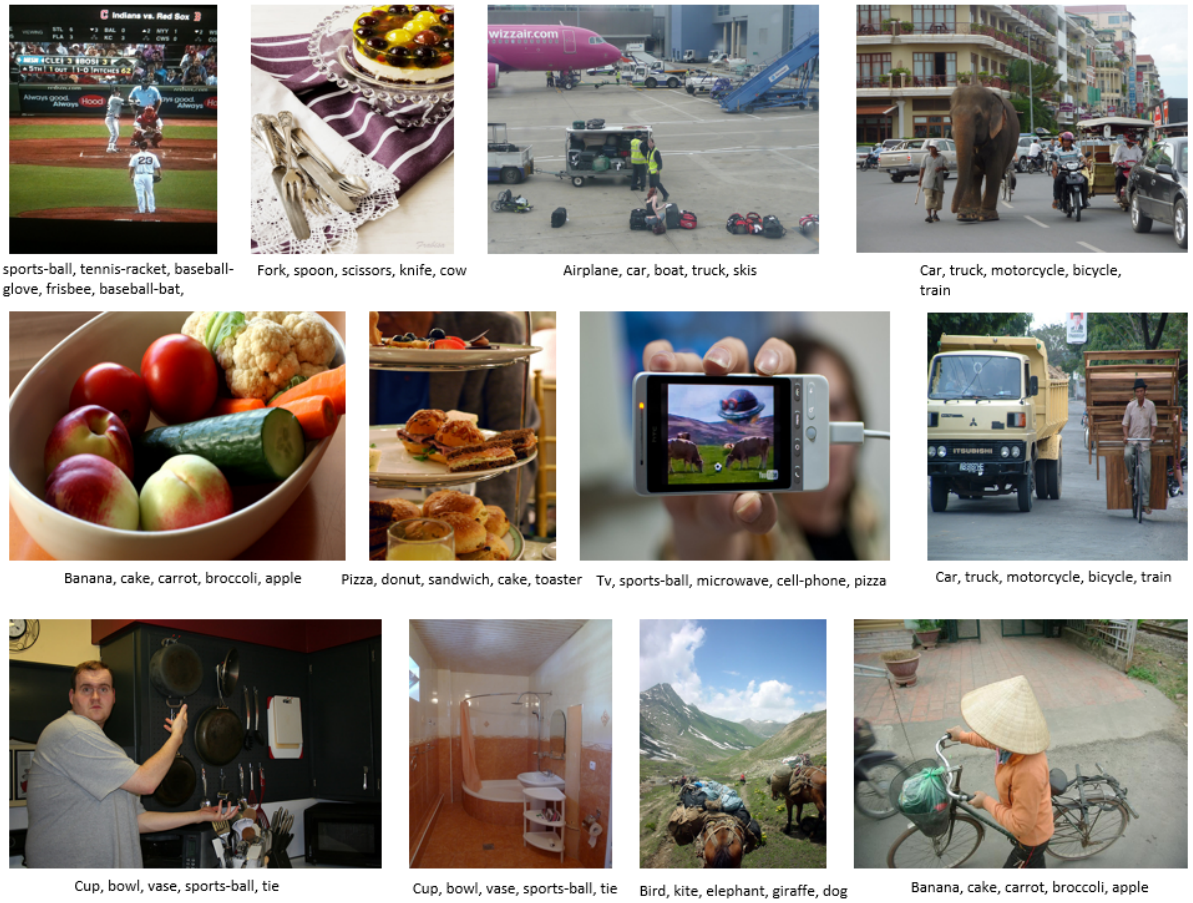

Figure 3: Qualitative Results: Top Row: When the correct top-object classification is able to guess at least one more object in an image. Middle Row: When the incorrect top object classification is able to guess at least one more object in an image. Bottom Row: When the incorrect top object classification is NOT able to guess at least one more object in an image.

where $SIM(n_i, n_1)$ and $SIM(n_i, n_2)$, $SIM(n_i, n_3)$ are the cosine similarities of noun n_i to nouns n_1, n_2 and n_3 respectively.

In contrast, if we use only 1 top object classification for guessing other objects, then, Given a set of nouns N and top object classifications n_1, n_2 and n_3, the closest guesses from set N are given by:

$$\arg\max_i \ \{SIM(n_i, n_1))\} \tag{2}$$

where $SIM(n_i, n_1)$ is the cosine similarities of noun n_i to nouns n_1 respectively.

Experiments

Training: We used k-means clustering to cluster objects using FastText embeddings trained on Wikipedia. We experimented with several k values, namely, {3, 5, 7, 9, 11, 13, 15, 17}, which were applied to a validation set of 20,000 images obtained by splitting the 40,000 validation images in MS-COCO into 20,000 images each for validation and testing. The optimal value for k was determined by running experiments on the validation set. The k value that yielded the best results was adopted for subsequent processing. We found that among the k values in the above set, k=17 yielded the best results on validation set. Because k-means clustering is non-deterministic, we used ten iterations of each value of k and calculated the average of the obtained results.

We trained object classifiers using features extracted from the fc-7 layer of the VGG architecture (Simonyan and Zisserman 2014), followed by support vector machines, on the MS-COCO training set images. We trained classifiers for 79 objects while excluding the *person* category from our experiments, as a person could co-occur with any type of object.

For each experiment pertaining to a particular iteration for a particular k value, we selected representative objects from each cluster for training. For example, - for a k value of 3, we selected 3 representative objects for training. Hence, k objects were trained for each experiment. In each experiment pertaining to a particular k value, we assume that other objects in the MS-COCO dataset are not available. Finally, we chose k=17 for reporting our results because on validation set, for values of k from 3 to 17, it yielded the best performance. In addition, k=17 implies about twenty percent of the classes are seen while eighty percent are unseen. This is in contrast to previous work on zero-shot object recognition where most classes tend to be seen and fewer are unseen.

During testing, we ran the classifiers for all the representative objects in a given image. The most likely representative objects that occurred in an image were then chosen to determine the other objects in an image. These other objects were predicted from the FastText embeddings using equation 1 (for a guess made with the 3 most probable objects) and equation 2 (for a guess made with 1 top-most probable object only).

We evaluated our results on the test set using ground truth object annotations, as well as the aforementioned predictions (guesses), using the following metrics:

Top-1 accuracy: Achieved when the top 1 object determined by our model intersected with at least one object in the ground truth annotations of the test image when all the objects except the representative objects are considered for classification.

Top-1c accuracy: Attained when the top 1 object determined by our model intersected with at least one object in the ground truth of the test image when only the representative objects were considered for classification. For the top one object, it would be interesting to see how simply knowing the representative object alone affects the accuracy vs another object that is predicted by the representative object. It would be important to know how far we can go when executing classifiers for representative objects only.

Top-3 accuracy: Achieved when any of the top 3 objects predicted by our model intersected with at least one object in the ground truth annotations of the test image. This included all objects (the representative object as well as others).

Top-5 accuracy: Attained when any of the top 5 objects determined by our model intersected with at least one object in the ground truth annotations of the test image. This could include all objects (the representative object as well as others).

Most-frequent baseline –The most frequent object/objects were determined by the most frequently occurring objects in the training set. The most frequent objects in MS-COCO (excluding persons) in the training dataset (in descending order) were *chair, car, dining table, cup, bottle*, and *bowl*. For top-n accuracy, the top-n most frequent objects were used for evaluation on the test set. The most frequent baseline is a fundamental baseline widely used by machine learning researchers. This baseline has been shown by several authors to be sometimes difficult to beat (Preiss et al. 2009).

Table 1: Comparison of our approach with most frequent baseline for guesses made with only one most probable objects in an image for k=17.

	Top-1c	Top-1	Top-3	Top-5
Most-Freq	11%	11%	25%	31%
Ours	35%	13%	47%	55%

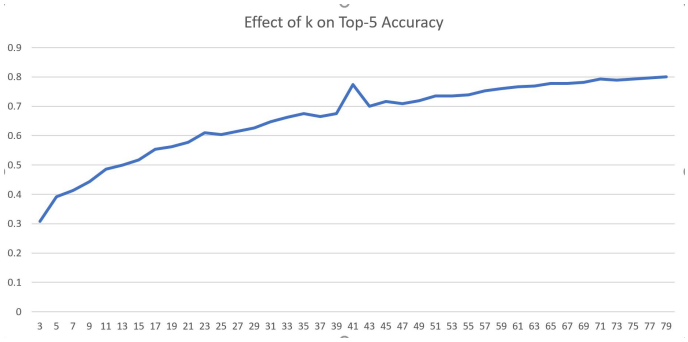

Figure 4: Effect of value of k used in k-means clustering on Top-5 Accuracy.

Results and Discussion

1. The results reported here show that our approach is superior to the most frequent baseline when objects are predicted using the top-most likely, second most likely, and third most likely objects. This finding lends support to the idea that using only a few classifiers to guess various objects from language models learned from public datasets is an effective technique.

2. To see how the increase in k effects accuracy, we conducted additional experiments with the k values from k=3 to k=79. Our results further show that the object prediction accuracy increases with the value of k in k-means clustering, as shown in figures 4. This increase could be attributed to one or more of the following factors: (1) either performance is improved by increasing the number of classifiers or increasing the number of clusters improves the performance; and (2) a greater number of clusters tends to assign objects possessing some type of relationship into the same cluster and tends to correctly separate unrelated objects.

3. The present investigation also revealed that performance does not improve due to using the most likely object in an image: the same results can be attained with the top-2 or top-3 most likely objects. This implies that the most confident classification is sufficient for making accurate guesses concerning other objects in an image. Importantly, at least for FastText embeddings, including additional information beyond one object classification does not seem to be beneficial when identifying/guessing other objects in an image; hence, this knowledge removes an extra layer of complexity.

4. Because the results obtained using top-5 accuracy were superior to those related to top-3 and top-1 accuracy, the likelihood of obtaining an accurate classification for at least one object increases when the classifier attempts to guess a

Table 2: Comparison of our approach with most frequent baseline for guesses made with three most probable objects in an image for k=17.

	Top-1	Top-3	Top-5
Most-Freq	11%	25%	31%
Ours	22%	46%	53%

Table 3: Comparison of our approach with most frequent baseline for guesses made with two most probable objects in an image for k=17.

	Top-1	Top-3	Top-5
Most-Freq	11%	25%	31%
Ours	26%	47%	55%

Cluster	Nouns
1	cake, donut, pizza, toaster, oven, carrot, sandwich, refrigerator, banana, spoon, broccoli
2	skis, surfboard, snowboard, **backpack,** boat, kite, hair-dryer
3	**dining-table**, wine-glass, toilet, potted-plant, bench, book
4	laptop, keyboard, **cell-phone**, apple, mouse
5	**bowl**, orange
6	**car**, truck, motor-cycle, bicycle, airplane, suitcase
7	**clock**
8	train, **bus**, stop-sign,
9	**cup**, vase
10	parking-meter, **traffic-light**, fire-hydrant,
11	firsbee, skateboard, **umbrella**
12	**remote**, microwave,
13	**dog**, cow, hot-dog, elephant, sheep, giraffe, cat, hoarse, zebra, bird
14	**sink,**
15	scissors, tooth-brush, knife, handbag, bed, bottle, couch, **chair**
16	baseball-glove, sports-ball, baseball-bat, tennis-racket, tie, **TV**
17	bear, teddy-bear, **fork**

Figure 5: Clusters for k=17. The chosen centers are in bold.

number of objects in an image.

Failure Analysis

Even though natural language provides good priors for images, it is well known that the natural language world does not necessarily map to the visual world in all situations. In addition, the reasons why word vectors have such interesting properties is currently not well understood. Hence, such limitations are inevitable, and it is necessary to address those situations. To conduct a failure analysis, we selected one iteration of k=17; the clusters for k=17 are listed in figure 5. We analyze the false positive rates for the categories guessed by each most probable-object (also cluster center) for the selected cluster, and then we analyze the categories with the highest false negative rates for the most likely object. For the purpose of this analysis, the false positive category is the category rated as the most-probable object in an image, as one of four guesses made for this object. In the original results, we used top-5 accuracy, which included the most probable object. However, for the purpose of this analysis, we exclude most probable object.

For each most probable object (cluster center), we calculate the false discovery rate FDR for each of the categories in the COCO dataset as follows: $FDR = \frac{F_p}{F_p+T_p}$

where F_p represents the false positives pertaining to a particular category and T_p represents the true positives.

From the false discovery rate table, we can observe certain situations. First, rather non-obviously, categories with the highest FDR turned out to be ones that are highly unlikely to co-occur, such as *bowl* and *baseball-bat*, *fork* and *cow*, and *dining-table* and *toilet*. This could be attributed to the FastText word vectors because the vectors of these objects turned out to be in close proximity, even though they are unlikely to occur together in images. In the second situation, interestingly, the categories with the highest FDR turned out to be ones which that could plausibly co-occur, yet do not. For example, - *backpack* and *laptop*, and *spoon* and *toaster*. One reason could be that these categories, although related in the textual world, tend not to visibly co-occur in an image. This could be attributed to the fact that the natural language context is sometimes unable to augment the visual context. Another situation with the highest FDR was when another category from the same general category was retrieved, such as *train* by *car* and *bicycle* by *bus*. Cars and trains, although they belong to the same general category, do not tend to co-occur in the real world; hence, this is another example of the natural language context being unable to assist with visual recognition. Nevertheless, language priors provide reasonable guesses for many, if not all, real-world situations.

Table 4: Table reflecting categories with highest FDR for each cluster center when that particular cluster center was the most probable object.

Cluster Center	**Category with Highest FDR**	**False Discovery Rate**
Spoon	Toaster	1
Fork	Cow	1
Bowl	Baseball-bat	1
Bus	Bicycle	0.93
Sink	Boat	1
Dining-Table	Toilet	1
Cup	Sports-ball	1
Car	Train	0.94
Remote	Microwave	0.98
Backpack	Laptop	1

Conclusion

Our models and results lend support to the concept that - even using a limited number of object classifiers - other objects in an image could be successfully guessed, even from language priors learned on unrelated datasets such as Wikipedia. Moreover, even the ability to classify incorrect objects in an image can yield useful cues that could help in guessing correct objects in an image. However, such natural language priors are not always helpful, as described by the results of our failure analysis. Nevertheless, word embeddings trained on unrelated public datasets can still yield effective priors - at least good enough to matter in uncertain situations. Future work will involve incorporating feedback mechanisms to improve classification. In addition, many weakly supervised object detection and image captioning algorithms could benefit from our approach. Additionally, this approach could be used by AI practitioners to help conceptualize human intelligence because humans tend to recognize numerous categories without ever having been exposed to them.

References

Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2016. Enriching word vectors with subword information. arxiv preprint arxiv:1607.04606. *arXiv preprint* 1607:04606.

Choi, M.; Torralba, A.; and Willsky, A. 2012. A tree- based context model for object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34(2):240–252.

Divvala, S.; Hoiem, D.; Hays, J.; Efros, A.; and Hebert, M. 2009. An empirical study of context in object detection. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*.

Gkioxari, G. 2016. Contextual visual recognition from images and videos. *University of California, Berkeley, (Technical Report)*.

Gonzalez-Garcia, A.; Modolo, D.; and Ferrari, V. 2017. Objects as context for part detection. *arXiv preprint* 1703:09529.

Heitz, G., and Koller, D. 2008. Learning spatial context: Using stuff to find things. *Proc. European Conference on Computer Vision*.

Lin, T.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; P Dollár, P.; and Zitnick, C. 2014. Microsoft coco: Common objects in context. *In Proc. ECCV* 740–755.

Malisiewicz, T., and Efros, A. 2009. Beyond categories: The visual memex model for reasoning about object relationships. *Proc. Neural Information Processing Systems*.

Mikolov, T. 2013. Distributed representations of words and phrases and their compositionality. *In Proc. NIPS* 3111–3119.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*.

Preiss, J.; Dehdari, J.; King, J.; and Mehay, D. 2009. Refining the most frequent sense baseline. *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*.

Rabinovich, A., and Belongie, S. 2009. Scenes vs. objects: a comparative study of two approaches to context-based recognition. *International Workshop on Visual Scene Understanding, Miami, FL.*

Rabinovich, A.; Vedaldi, A.; Galleguillos, C.; Wiewiora, E.; and Belongie, S. 2007. Objects in context. *Proc. International Conference on Computer Vision*.

Sharma, K.; Kumar, A.; and Bhandarkar, S. 2016. Guessing objects in context. *In ACM SIGGRAPH 2016 Posters*.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *Proc. Intl. Conf. Learn. Rep.*

Socher, R.; Ganjoo, M.; Manning, C.; and Ng, A. 2013. Zero-shot learning through cross-modal transfer. *In Proc. NIPS*.

Sun, J., and Jacobs, D. W. 2017. Seeing what is not there: Learning context to determine where objects are missing. *arXiv preprint arXiv:1702.07971*.

Torralba, A. The context challenge. *http://web.mit.edu/torralba/www/carsAndFacesInContext.html*.

Xiao, J.; Hays, J.; Ehinger, K.; Oliva, A.; and Torralba, A. 2010. Sun database: Large-scale scene recognition from abbey to zoo. *IEEE Conference on Computer Vision and Pattern Recognition*.

Zhang, Y.; Bai, M.; Kohli, P.; Izadi, S.; and Xiao, J. 2016. Deepcontext: context-encoding neural pathways for 3d holistic scene understanding. *arXiv preprint arXiv:1603.04922*.