Learning Image Embeddings using Convolutional Neural Networks for Improved Multi-Modal Semantics

Douwe Kiela
University of Cambridge
Computer Laboratory
douwe.kiela@cl.cam.ac.uk

Léon Bottou
Microsoft Research
641 Ave of the Americas
New York NY 10011
leon@bottou.org

Abstract

We construct multi-modal concept representations by concatenating a skip-gram linguistic representation vector with a visual concept representation vector computed using the feature extraction layers of a deep convolutional neural network (CNN) trained on a large labeled object recognition dataset. This transfer learning approach brings a clear performance gain over features based on the traditional bag-of-visual-word approach. Experimental results are reported on the WordSim353 and MEN semantic relatedness evaluation tasks. We use visual features computed using either ImageNet or ESP Game images.

1 Introduction

Recent works have shown that multi-modal semantic representation models outperform unimodal linguistic models on a variety of tasks, including modeling semantic relatedness and predicting compositionality (Feng and Lapata, 2010; Leong and Mihalcea, 2011; Bruni et al., 2012; Roller and Schulte im Walde, 2013; Kiela et al., 2014). These results were obtained by combining linguistic feature representations with robust visual features extracted from a set of images associated with the concept in question. This extraction of visual features usually follows the popular computer vision approach consisting of computing local features, such as SIFT features (Lowe, 1999), and aggregating them as bags of visual words (Sivic and Zisserman, 2003).

Meanwhile, deep transfer learning techniques have gained considerable attention in the computer vision community. First, a deep convolutional neural network (CNN) is trained on a large labeled dataset (Krizhevsky et al., 2012). The convolutional layers are then used as mid-level feature extractors on a variety of computer vision tasks (Oquab et al., 2014; Girshick et al., 2013; Zeiler and Fergus, 2013; Donahue et al., 2014). Although transferring convolutional network features is not a new idea (Driancourt and Bottou, 1990), the simultaneous availability of large datasets and cheap GPU co-processors has contributed to the achievement of considerable performance gains on a variety computer vision benchmarks: *"SIFT and HOG descriptors produced big performance gains a decade ago, and now deep convolutional features are providing a similar breakthrough"* (Razavian et al., 2014).

This work reports on results obtained by using CNN-extracted features in multi-modal semantic representation models. These results are interesting in several respects. First, these superior features provide the opportunity to increase the performance gap achieved by augmenting linguistic features with multi-modal features. Second, this increased performance confirms that the multimodal performance improvement results from the information contained in the images and not the information used to select which images to use to represent a concept. Third, our evaluation reveals an intriguing property of the CNN-extracted features. Finally, since we use the skip-gram approach of Mikolov et al. (2013) to generate our linguistic features, we believe that this work represents the first approach to multimodal distributional semantics that exclusively relies on deep learning for both its linguistic and visual components.

2 Related work

2.1 Multi-Modal Distributional Semantics

Multi-modal models are motivated by parallels with human concept acquisition. Standard semantic space models extract meanings solely from linguistic data, even though we know that human semantic knowledge relies heavily on perceptual information (Louwerse, 2011). That is, there exists substantial evidence that many concepts are *grounded* in the perceptual system (Barsalou, 2008). One way to do this grounding in the context of distributional semantics is to obtain representations that combine information from linguistic corpora with information from another modality, obtained from e.g. property norming experiments (Silberer and Lapata, 2012; Roller and Schulte im Walde, 2013) or from processing and extracting features from images (Feng and Lapata, 2010; Leong and Mihalcea, 2011; Bruni et al., 2012). This approach has met with quite some success (Bruni et al., 2014).

2.2 Multi-modal Deep Learning

Other examples that apply multi-modal deep learning use restricted Boltzmann machines (Srivastava and Salakhutdinov, 2012; Feng et al., 2013), auto-encoders (Wu et al., 2013) or recursive neural networks (Socher et al., 2014). Multi-modal models with deep learning components have also successfully been employed in cross-modal tasks (Lazaridou et al., 2014). Work that is closely related in spirit to ours is by Silberer and Lapata (2014). They use a stacked auto-encoder to learn combined embeddings of textual and visual input. Their visual inputs consist of vectors of visual attributes obtained from learning SVM classifiers on attribute prediction tasks. In contrast, our work keeps the modalities separate and follows the standard multi-modal approach of concatenating linguistic and visual representations in a single semantic space model. This has the advantage that it allows for separate data sources for the individual modalities. We also learn visual representations directly from the images (i.e., we apply deep learning directly to the images), as opposed to taking a higher-level representation as a starting point. Frome et al. (2013) jointly learn multi-modal representations as well, but apply them to a visual object recognition task instead of concept meaning.

2.3 Deep Convolutional Neural Networks

A flurry of recent results indicates that image descriptors extracted from deep convolutional neural networks (CNNs) are very powerful and consistently outperform highly tuned state-of-the-art systems on a variety of visual recognition tasks (Razavian et al., 2014). Embeddings from state-of-the-art CNNs (such as Krizhevsky et al. (2012)) have been applied successfully to a number of problems in computer vision (Girshick et al., 2013; Zeiler and Fergus, 2013; Donahue et al., 2014). This contribution follows the approach described by Oquab et al. (2014): they train a CNN on 1512 ImageNet synsets (Deng et al., 2009), use the first seven layers of the trained network as feature extractors on the Pascal VOC dataset, and achieve state-of-the-art performance on the Pascal VOC classification task.

3 Improving Multi-Modal Representations

Figure 1 illustrates how our system computes multi-modal semantic representations.

3.1 Perceptual Representations

The perceptual component of standard multi-modal models that rely on visual data is often an instance of the bag-of-visual-words (BOVW) representation (Sivic and Zisserman, 2003). This approach takes a collection of images associated with words or tags representing the concept in question. For each image, keypoints are laid out as a dense grid. Each keypoint is represented by a vector of robust local visual features such as SIFT (Lowe, 1999), SURF (Bay et al., 2008) and HOG (Dalal and Triggs, 2005), as well as pyramidal variants of these descriptors such as PHOW (Bosch et al., 2007). These descriptors are subsequently clustered into a discrete set of "visual words" using a standard clustering algorithm like k-means and quantized into vector representations by comparing the local descriptors with the cluster centroids. Visual representations are obtained by taking the average of the BOVW vectors for the images that correspond to a given word. We use BOVW as a baseline.

Our approach similarly makes use of a collection of images associated with words or tags representing a particular concept. Each image is processed by the first seven layers of the convolutional network defined by Krizhevsky et al. (2012) and adapted by Oquab et al. (2014)[1]. This network takes 224×224 pixel RGB images and applies five successive convolutional layers followed by three fully connected layers. Its eighth and last layer produces a vector of 1512 scores associated

[1] http://www.di.ens.fr/willow/research/cnn/

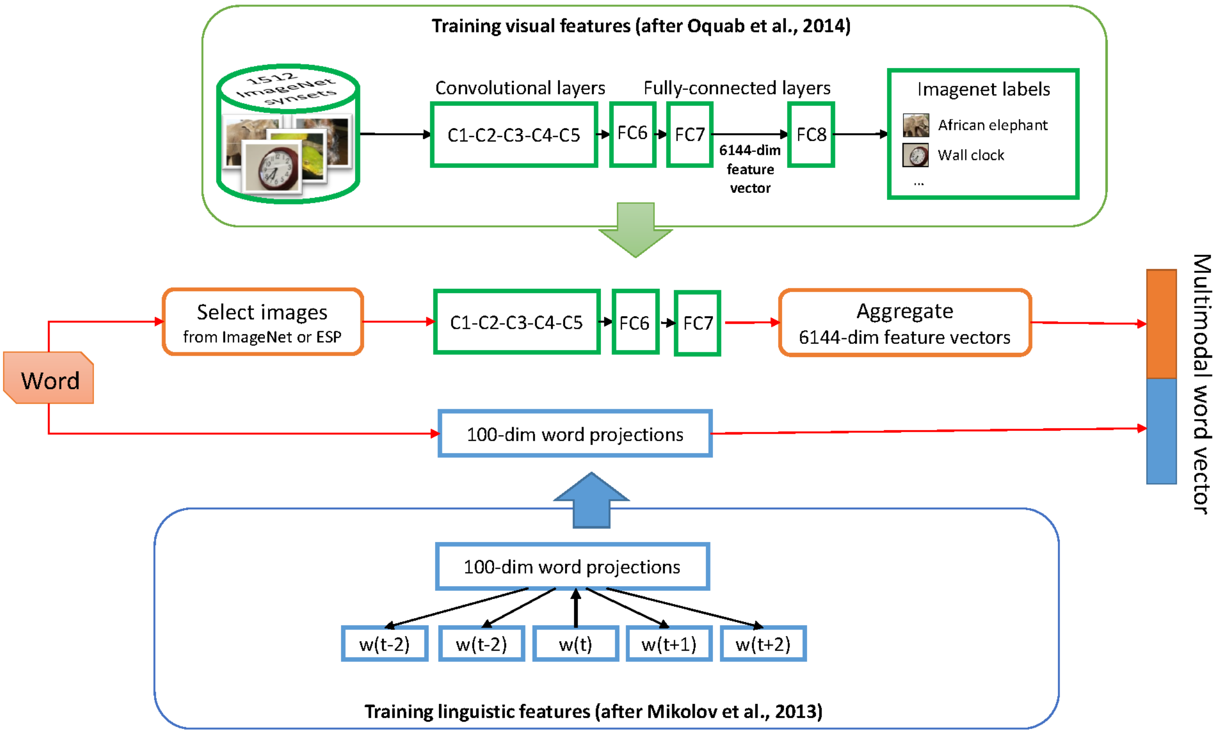

Figure 1: Computing word feature vectors.

with 1000 categories of the ILSVRC-2012 challenge and the 512 additional categories selected by Oquab et al. (2014). This network was trained using about 1.6 million ImageNet images associated with these 1512 categories. We then freeze the trained parameters, chop the last network layer, and use the remaining seventh layer as a filter to compute a 6144-dimensional feature vector on arbitrary 224×224 input images.

We consider two ways to aggregate the feature vectors representing each image.

1. The first method (**CNN-Mean**) simply computes the average of all feature vectors.

2. The second method (**CNN-Max**) computes the component-wise maximum of all feature vectors. This approach makes sense because the feature vectors extracted from this particular network are quite sparse (about 22% non-zero coefficients) and can be interpreted as bags of visual properties.

3.2 Linguistic representations

For our linguistic representations we extract 100-dimensional continuous vector representations using the log-linear skip-gram model of Mikolov et al. (2013) trained on a corpus consisting of the 400M word Text8 corpus of Wikipedia text[2] together with the 100M word British National Corpus (Leech et al., 1994). We also experimented with dependency-based skip-grams (Levy and Goldberg, 2014) but this did not improve results. The skip-gram model learns high quality semantic representations based on the distributional properties of words in text, and outperforms standard distributional models on a variety of semantic similarity and relatedness tasks. However we note that Bruni et al. (2014) have recently reported an even better performance for their linguistic component using a standard distributional model, although this may have been tuned to the task.

3.3 Multi-modal Representations

Following Bruni et al. (2014), we construct multi-modal semantic representations by concatenating the centered and L_2-normalized linguistic and perceptual feature vectors $\vec{v}_{ling}$ and $\vec{v}_{vis}$,

$$\vec{v}_{concept} = \alpha \times \vec{v}_{ling} \,||\, (1-\alpha) \times \vec{v}_{vis} \,, \quad (1)$$

where $||$ denotes the concatenation operator and α is an optional tuning parameter.

[2]http://mattmahoney.net/dc/textdata.html

Figure 2: Examples of *dog* in the ESP Game dataset.

Figure 3: Examples of *golden retriever* in ImageNet.

4 Experimental Setup

We carried out experiments using visual representations computed using two canonical image datasets. The resulting multi-modal concept representations were evaluated using two well-known semantic relatedness datasets.

4.1 Visual Data

We carried out experiments using two distinct sources of images to compute the visual representations.

The **ImageNet** dataset (Deng et al., 2009) is a large-scale ontology of images organized according to the hierarchy of WordNet (Fellbaum, 1999). The dataset was constructed by manually re-labelling candidate images collected using web searches for each WordNet synset. The images tend to be of high quality with the designated object roughly centered in the image. Our copy of ImageNet contains about 12.5 million images organized in 22K synsets. This implies that ImageNet covers only a small fraction of the existing 117K WordNet synsets.

The **ESP Game** dataset (Von Ahn and Dabbish, 2004) was famously collected as a “game with a purpose”, in which two players must independently and rapidly agree on a correct word label for randomly selected images. Once a word label has been used sufficiently frequently for a given image, that word is added to the image’s tags. This dataset contains 100K images, but with every image having on average 14 tags, that amounts to a coverage of 20,515 words. Since players are encouraged to produce as many terms per image, the dataset’s increased coverage is at the expense of accuracy in the word-to-image mapping: a dog in a field with a house in the background might be a *golden retriever* in ImageNet and could have tags *dog, golden retriever, grass, field, house, door* in the ESP Dataset. In other words, images in the ESP dataset do not make a distinction between objects in the foreground and in the background, or between the relative size of the objects (tags for images are provided in a random order, so the top tag is not necessarily the best one).

Figures 2 and 3 show typical examples of images belonging to these datasets. Both datasets have attractive properties. On the one hand, ImageNet has higher quality images with better labels. On the other hand, the ESP dataset has an interesting coverage because the MEN task (see section 4.4) was specifically designed to be covered by the ESP dataset.

4.2 Image Selection

Since ImageNet follows the WordNet hierarchy, we would have to include almost all images in the dataset to obtain representations for high-level concepts such as *entity*, *object* and *animal*. Doing so is both computationally expensive and unlikely to improve the results. For this reason, we randomly sample up to N distinct images from the subtree associated with each concept. When this returns less than N images, we attempt to increase coverage by sampling images from the subtree of the concept’s hypernym instead. In order to allow for a fair comparison, we apply the same method of sampling up to N on the ESP Game dataset. In all following experiments, $N = 1.000$. We used the WordNet lemmatizer from NLTK (Bird et al., 2009) to lemmatize tags and concept words so as to further improve the dataset’s coverage.

4.3 Image Processing

The ImageNet images were preprocessed as described by (Krizhevsky et al., 2012). The largest centered square contained in each image is resampled to form a 256×256 image. The CNN input is then formed by cropping 16 pixels off each border and subtracting 128 to the image components. The ESP Game images were preprocessed slightly differently because we do not expect the objects to be centered. Each image was rescaled to fit inside a 224×224 rectangle. The CNN input is then formed by centering this image into the 224×224 input field, subtracting 128 to the image components, and zero padding.

The BOVW features were obtained by computing DSIFT descriptors using VLFeat (Vedaldi and Fulkerson, 2008). These descriptors were subsequently clustered using mini-batch k-means (Sculley, 2010) with 100 clusters. Each image is then represented by a bag of clusters (visual words) quantized as a 100-dimensional feature vector. These vectors were then combined into visual concept representations by taking their mean.

4.4 Evaluation

We evaluate our multi-modal word representations using two semantic relatedness datasets widely used in distributional semantics (Agirre et al., 2009; Feng and Lapata, 2010; Bruni et al., 2012; Kiela and Clark, 2014; Bruni et al., 2014).

WordSim353 (Finkelstein et al., 2001) is a selection of 353 concept pairs with a similarity rating provided by human annotators. Since this is probably the most widely used evaluation dataset for distributional semantics, we include it for comparison with other approaches. WordSim353 has some known idiosyncracies: it includes named entities, such as *OPEC*, *Arafat*, and *Maradona*, as well as abstract words, such as *antecedent* and *credibility*, for which it may be hard to find corresponding images. Multi-modal representations are often evaluated on an unspecified subset of WordSim353 (Feng and Lapata, 2010; Bruni et al., 2012; Bruni et al., 2014), making it impossible to compare the reported scores. In this work, we report scores on the full WordSim353 dataset (**W353**) by setting the visual vector $\vec{v}_{vis}$ to zero for concepts without images. We also report scores on the subset (**W353-Relevant**) of pairs for which both concepts have both ImageNet and ESP Game images using the aforementioned selection procedure.

MEN (Bruni et al., 2012) was in part designed to alleviate the WordSim353 problems. It was constructed in such a way that only frequent words with at least 50 images in the ESP Game dataset were included in the evaluation pairs. The MEN dataset has been found to mirror the aggregate score over a variety of tasks and similarity datasets (Kiela and Clark, 2014). It is also much larger, with 3000 words pairs consisting of 751 individual words. Although MEN was constructed so as to have at least a minimum amount of images available in the ESP Game dataset for each concept, this is not the case for ImageNet. Hence, similarly to WordSim353, we also evaluate on a subset (**MEN-Relevant**) for which images are available in both datasets.

We evaluate the models in terms of their Spearman ρ correlation with the human relatedness ratings. The similarity between the representations associated with a pair of words is calculated using the cosine similarity:

$$\cos(v_1, v_2) = \frac{v_1 \cdot v_2}{\|v_1\| \, \|v_2\|} \tag{2}$$

5 Results

We evaluate on the two semantic relatedness datasets using solely linguistic, solely visual and multi-modal representations. In the case of MEN-Relevant and W353-Relevant, we report scores for BOVW, CNN-Mean and CNN-Max visual representations. For all datasets we report the scores obtained by BOVW, CNN-Mean and CNN-Max multi-modal representations. Since we have full coverage with the ESP Game dataset on MEN, we are able to report visual representation scores for the entire dataset as well. The results can be seen in Table 1.

There are a number of questions to ask. First of all, do CNNs yield better visual representations? Second, do CNNs yield better multi-modal representations? And third, is there a difference between the high-quality low-coverage ImageNet and the low-quality higher-coverage ESP Game dataset representations?

5.1 Visual Representations

In all cases, CNN-generated visual representations perform better or as good as BOVW representations (we report results for BOVW-Mean, which performs slightly better than taking the element-wise maximum). This confirms the motivation outlined in the introduction: by applying state-of-the-art approaches from computer vision to multi-modal semantics, we obtain a signficant perfor-

Dataset	Linguistic	Visual			Multi-modal		
		BOVW	CNN-Mean	CNN-Max	BOVW	CNN-Mean	CNN-Max
ImageNet visual features							
MEN	0.64	-	-	-	0.64	**0.70**	0.67
MEN-Relevant	0.62	0.40	0.64	0.63	0.64	**0.72**	0.71
W353	0.57	-	-	-	0.58	0.59	**0.60**
W353-Relevant	0.51	0.30	0.32	0.30	0.55	0.56	**0.57**
ESP game visual features							
MEN	0.64	0.17	0.51	0.20	0.64	**0.71**	0.65
MEN-Relevant	0.62	0.35	0.58	0.57	0.63	0.69	**0.70**
W353	0.57	-	-	-	0.58	0.59	**0.60**
W353-Relevant	0.51	0.38	0.44	0.56	0.52	0.55	**0.61**

Table 1: Results (see sections 4 and 5).

mance increase over standard multi-modal models.

5.2 Multi-modal Representations

Higher-quality perceptual input leads to better-performing multi-modal representations. In all cases multi-modal models with CNNs outperform multi-modal models with BOVW, occasionally by quite a margin. In all cases, multi-modal representations outperform purely linguistic vectors that were obtained using a state-of-the-art system. This re-affirms the importance of multi-modal representations for distributional semantics.

5.3 The Contribution of Images

Since the ESP Game images come with a multitude of word labels, one could question whether a performance increase of multi-modal models based on that dataset comes from the images themselves, or from overlapping word labels. It might also be possible that similar concepts are more likely to occur in the same image, which encodes relatedness information without necessarily taking the image data itself into account. In short, it is a natural question to ask whether the performance gain is due to image data or due to word label associations? We conclusively show that the image data matters in two ways: (a) using a different dataset (ImageNet) also results in a performance boost, and (b) using higher-quality image features on the ESP game images increases the performance boost without changing the association between word labels.

5.4 Image Datasets

It is important to ask whether the source image dataset has a large impact on performance. Although the scores for the visual representation in some cases differ, performance of multi-modal representations remains close for both image datasets. This implies that our method is robust over different datasets. It also suggests that it is beneficial to train on high-quality datasets like ImageNet and to subsequently generate embeddings for other sets of images like the ESP Game dataset that are more noisy but have better coverage. The results show the benefit of transfering convolutional network features, corroborating recent results in computer vision.

5.5 Semantic Similarity/Relatedness Datasets

There is an interesting discrepancy between the two types of network with respect to dataset performance: CNN-Mean multi-modal models tend to perform best on MEN and MEN-Relevant, while CNN-Max multi-modal models perform better on W353 and W353-Relevant. There also appears to be some interplay between the source corpus, the evaluation dataset and the best performing CNN: the performance leap on W353-

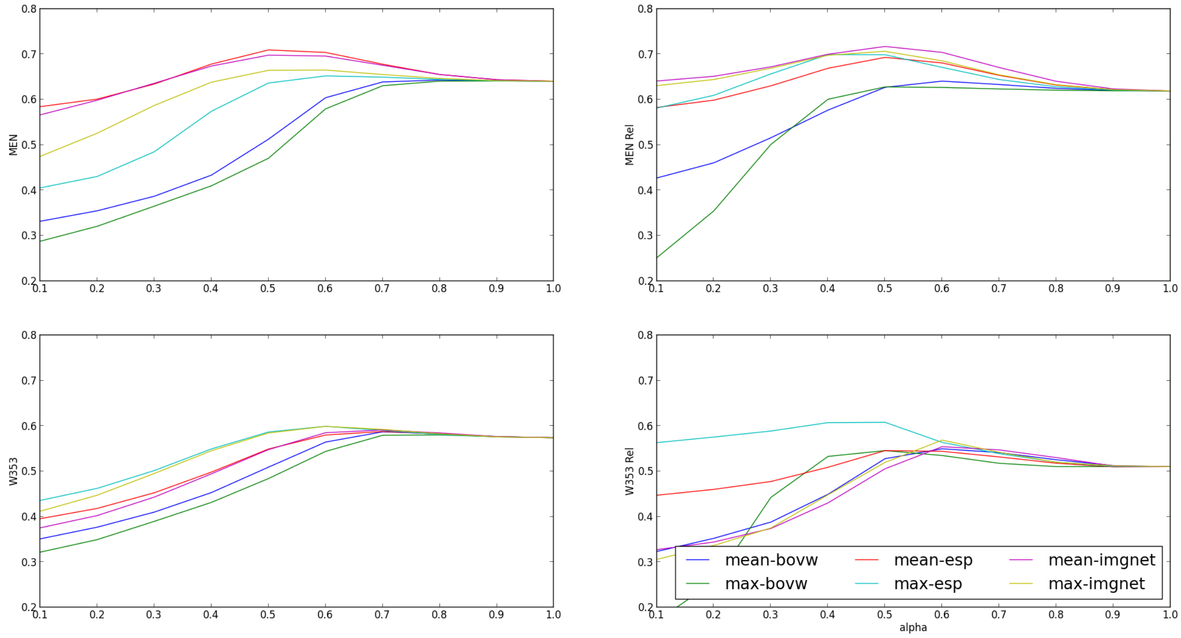

Figure 4: Varying the α parameter for MEN, MEN-Relevant, WordSim353 and WordSim353-Relevant, respectively.

Relevant for CNN-Max is much larger using ESP Game images than with ImageNet images.

We speculate that this is because CNN-Max performs better than CNN-Mean on a somewhat different type of similarity. It has been noted (Agirre et al., 2009) that WordSim353 captures both similarity (as in *tiger-cat*, with a score of 7.35) as well as relatedness (as in *Maradona-football*, with a score of 8.62). MEN, however, is explicitly designed to capture semantic relatedness only (Bruni et al., 2012). CNN-Max using sparse feature vectors means that we treat the dominant components as definitive of the concept class, which is more suited to similarity. CNN-Mean averages over all the feature components, and as such might be more suited to relatedness. We conjecture that the performance increase on WordSim353 is due to increased performance on the similarity subset of that dataset.

5.6 Tuning

The concatenation scheme in Equation 1 allows for a tuning parameter α to weight the relative contribution of the respective modalities. Previous work on MEN has found that the optimal parameter for that dataset is close to 0.5 (Bruni et al., 2014). We have found that this is indeed the case. On WordSim353, however, we have found the parameter for optimal performance to be shifted to the right, meaning that optimal performance is achieved when we include less of the visual input compared to the linguistic input. Figure 4 shows what happens when we vary alpha over the four datasets. There are a number of observations to be made here.

First of all, we can see that the performance peak for the MEN datastes is much higher than for the WordSim353 ones, and that its peak is relatively higher as well. This indicates that MEN is in a sense a more balanced dataset. There are two possible explanations: as indicated earlier, WordSim353 contains slightly idiosyncratic word pairs which may have a detrimental effect on performance; or, WordSim353 was not constructed with multi-modal semantics in mind, and contains a substantial amount of abstract words that would not benefit at all from including visual information.

Due to the nature of the datasets and the tasks at hand, it is arguably much more important that CNNs beat standard bag-of-visual-words representations on MEN than on W353, and indeed we see that there exists no α for which BOVW would beat any of the CNN networks.

6 Error Analysis

Table 2 shows the top 5 best and top 5 worst scoring word pairs for the two datasets using CNN-

W353-Relevant

ImageNet				ESP Game			
word1	word2	system score	gold standard	word1	word2	system score	gold standard
tiger	tiger	1.00	1.00	tiger	tiger	1.00	1.00
man	governor	0.53	0.53	man	governor	0.53	0.53
stock	phone	0.15	0.16	stock	phone	0.15	0.16
football	tennis	0.68	0.66	football	tennis	0.68	0.66
man	woman	0.85	0.83	man	woman	0.85	0.83
cell	phone	0.27	0.78	law	lawyer	0.33	0.84
discovery	space	0.10	0.63	monk	slave	0.58	0.09
closet	clothes	0.22	0.80	gem	jewel	0.41	0.90
king	queen	0.26	0.86	stock	market	0.33	0.81
wood	forest	0.13	0.77	planet	space	0.32	0.79

MEN-Relevant

ImageNet				ESP Game			
word1	word2	system score	gold standard	word1	word2	system score	gold standard
beef	potatoes	0.35	0.35	beef	potatoes	0.35	0.35
art	work	0.35	0.35	art	work	0.35	0.35
grass	stop	0.06	0.06	grass	stop	0.06	0.06
shade	tree	0.45	0.45	shade	tree	0.45	0.45
blonde	rock	0.07	0.07	blonde	rock	0.07	0.07
bread	potatoes	0.88	0.34	bread	dessert	0.78	0.24
fruit	potatoes	0.80	0.26	jacket	shirt	0.89	0.34
dessert	sandwich	0.76	0.23	fruit	nuts	0.88	0.33
pepper	tomato	0.79	0.27	dinner	lunch	0.93	0.37
dessert	tomato	0.66	0.14	dessert	soup	0.81	0.23

Table 2: The top 5 best and top 5 worst scoring pairs with respect to the gold standard.

Mean multi-modal vectors. The most accurate pairs are consistently the same across the two image datasets. There are some clear differences between the least accurate pairs, however. The MEN words *potatoes* and *tomato* probably have low quality ImageNet-derived representations, because they occur often in the bottom pairs for that dataset. The MEN words *dessert*, *bread* and *fruit* occur in the bottom 5 for both image datasets, which implies that their linguistic representations are probably not very good. For WordSim353, the bottom pairs on ImageNet could be said to be similarity mistakes; while the ESP Game dataset contains more relatedness mistakes (*king* and *queen* would evaluate similarity, while *stock* and *market* would evaluate relatedness). It is difficult to say anything conclusive about this discrepancy, but it is clearly a direction for future research.

7 Image embeddings

To facilitate further research on image embeddings and multi-modal semantics, we publicly release embeddings for all the image labels occurring in the ESP Game dataset. Please see the following web page: `http://www.cl.cam.ac.uk/~dk427/imgembed.html`

8 Conclusion

We presented a novel approach to improving multi-modal representations using deep convolutional neural network-extracted features. We reported high results on two well-known and widely-used semantic relatedness benchmarks, with increased performance both in the separate visual representations and in the combined multi-modal representations. Our results indicate that such multi-modal representations outperform both linguistic and standard bag-of-visual-words multi-modal representations. We have shown that our approach is robust and that CNN-extracted features from separate image datasets can succesfully be applied to semantic relatedness.

In addition to improving multi-modal representations, we have shown that the source of this improvement is due to image data and is not simply a result of word label associations. We have shown this by obtaining performance improvements on two different image datasets, and by obtaining higher performance with higher-quality image features on the ESP game images, without changing the association between word labels.

In future work, we will investigate whether our system can be further improved by including concreteness information or a substitute metric such as image dispersion, as has been suggested by other work on multi-modal semantics (Kiela et al., 2014). Furthermore, a logical next step to increase performance would be to jointly learn multi-modal representations or to learn weighting parameters. Another interesting possibility would be to examine multi-modal distributional compositional semantics, where multi-modal representations are composed to obtain phrasal representations.

Acknowledgments

We would like to thank Maxime Oquab for providing the feature extraction code.

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 19–27, Boulder, Colorado.

Lawrence W. Barsalou. 2008. Grounded cognition. *Annual Review of Psychology*, 59:617—845.

Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. 2008. SURF: Speeded Up Robust Features. In *Computer Vision and Image Understanding (CVIU)*, volume 110, pages 346–359.

Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Anna Bosch, Andrew Zisserman, and Xavier Munoz. 2007. Image classification using random forests and ferns. In *Proceedings of ICCV*.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of ACL 2012*, pages 136–145. Association for Computational Linguistics.

Elia Bruni, Nam Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, pages 886–893.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. IEEE.

Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. 2014. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. In *International Conference on Machine Learning (ICML 2014)*.

Xavier Driancourt and Léon Bottou. 1990. TDNN-extracted features. In *Proceedings of Neuro Nimes 90*, Nimes, France. EC2.

Christiane Fellbaum. 1999. *WordNet*. Wiley Online Library.

Yansong Feng and Mirella Lapata. 2010. Visual information in semantic representation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 91–99. Association for Computational Linguistics.

Fangxiang Feng, Ruifan Li, and Xiaojie Wang. 2013. Constructing hierarchical image-tags bimodal representations for word tags alternative choice. *CoRR*.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414. ACM.

Andrea Frome, Greg Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, MarcÁurelio Ranzato, and Tomas Mikolov. 2013. DeViSE: A Deep Visual-Semantic Embedding Model. In *NIPS*.

R. Girshick, J. Donahue, T. Darrell, and J. Malik. 2013. Rich feature hierarchies for accurate object detection and semantic segmentation. *arXiv preprint:1311.2524*, November.

Douwe Kiela and Stephen Clark. 2014. A Systematic Study of Semantic Vector Space Model Parameters. In *Proceedings of EACL 2014, Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*.

Douwe Kiela, Felix Hill, Anna Korhonen, and Stephen Clark. 2014. Improving Multi-Modal Representations Using Image Dispersion: Why Less is Sometimes More. In *Proceedings of ACL 2014*.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1106–1114.

Angeliki Lazaridou, Elia Bruni, and Marco Baroni. 2014. Is this a wampimuk? cross-modal mapping between distributional semantics and the visual world. In *Proceedings of ACL 2014*.

Geoffrey Leech, Roger Garside, and Michael Bryant. 1994. Claws4: the tagging of the British National Corpus. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, pages 622–628. Association for Computational Linguistics.

Ben Leong and Rada Mihalcea. 2011. Going Beyond Text: A Hybrid Image-Text Approach for Measuring Word Relatedness. In *Proceedings of Joint International Conference on Natural Language Processing (IJCNLP)*, Chiang Mai, Thailand.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of ACL 2014*.

M. M. Louwerse. 2011. Symbol interdependency in symbolic and embodied cognition. *TopiCS in Cognitive Science*, 3:273—302.

David G. Lowe. 1999. Object recognition from local scale-invariant features. In *Proceedings of the International Conference on Computer Vision-Volume 2 - Volume 2*, ICCV '99.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of International Conference of Learning Representations*, Scottsdale, Arizona, USA.

M. Oquab, L. Bottou, I. Laptev, and J. Sivic. 2014. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

A.S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. 2014. CNN features off-the-shelf: an astounding baseline for recognition. *arXiv preprint:1403.6382*.

Stephen Roller and Sabine Schulte im Walde. 2013. A multimodal LDA model integrating textual, cognitive and visual modalities. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1146–1157, Seattle, Washington, USA, October. Association for Computational Linguistics.

D Sculley. 2010. Web-scale k-means clustering. In *Proceedings of the 19th international conference on World wide web*, pages 1177–1178. ACM.

Carina Silberer and Mirella Lapata. 2012. Grounded models of semantic representation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1423–1433. Association for Computational Linguistics.

Carina Silberer and Mirella Lapata. 2014. Learning Grounded Meaning Representations with Autoencoders. In *Proceedings of ACL 2014*, Baltimore, MD.

J. Sivic and A. Zisserman. 2003. Video Google: a text retrieval approach to object matching in videos. In *Proceedings of the Ninth IEEE International Conference on Computer Vision*, volume 2, pages 1470–1477, Oct.

Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Andrew Y. Ng. 2014. Grounded Compositional Semantics for Finding and Describing Images with Sentences. *Transactions of the Association for Computational Linguistics (TACL 2014)*.

Nitish Srivastava and Ruslan Salakhutdinov. 2012. Multimodal learning with deep boltzmann machines. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2222–2230.

A. Vedaldi and B. Fulkerson. 2008. VLFeat: An open and portable library of computer vision algorithms. `http://www.vlfeat.org/`.

Luis Von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326. ACM.

Pengcheng Wu, Steven C.H. Hoi, Hao Xia, Peilin Zhao, Dayong Wang, and Chunyan Miao. 2013. Online multimodal deep similarity learning with application to image retrieval. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, pages 153–162.

Matthew D. Zeiler and Rob Fergus. 2013. Visualizing and understanding convolutional networks. *CoRR*.