Learning Words from Images and Speech

Gabriel Synnaeve, Maarten Versteegh, Emmanuel Dupoux
Ecole Normale Supérieure / PSL Research University / EHESS / CNRS, Paris, France
gabriel.synnaeve@gmail.com, maartenversteegh@gmail.com, emmanuel.dupoux@gmail.com

Abstract

This paper explores the possibility to learn a semantically-relevant lexicon from images and speech only. For this, we train a multi-modal neural network working both on image fragments and on speech features, by learning an embedding in which images and content words that co-occur together are close. Making no assumption on the acoustic model, this paper shows promising results on how multi-modality could help word learning.

1 Introduction

Previous research has shown that an efficient acoustic model can be trained from word-level annotation alone [1]. Here, we explore the possibility of learning both an acoustic model and a word/image association from multi-modal co-occurrences between speech and pictures alone. Our work is inspired by the observation that infants achieve spontaneously this kind of correspondence during their first year of life. However, before they know any words, young infants can already categorize geometric shapes [2], as well as pictures of natural kinds like animals [3]. They also organize complex scenes in terms of perceptual objects that exhibit spatio-temporal continuity [4]. Thus, it is developmentally plausible to base word/meaning learning on a pre-segmented and optimized representation of images like it was done in [5, 6].

Indeed in [6], the authors use fragments of images coded through a convolutional neural network trained for detection and classification on ImageNet [7, 8]. Here, we learn the image fragments and words co-occurrences, based on a flattened representation of the sentences (without the compositional tree structure of the sentence), which resembles the task of cross-situational learning studied in psycholinguistics [9].

The idea of "Siamese" (identical sub-networks) neural networks [10] is to compute the distance between two inputs and use it to train an embedding. In our case, the network is not "Siamese" as it is multi-modal. Previous work have shown that same-different side information can benefit metric learning [11]. Hadsell et al. [12] also used an asymmetric loss for same-different pairs to learn invariant-properties manifolds (on images). Our architecture uses such a same-different based asymmetric loss function, but does not share weights (as it is multi-modal).

2 Data and Model Architecture

The dataset was constructed by merging two existing datasets, one for speech, and one for images-sentences mapping. For the images, we used the Pascal1K dataset [13], which contains 1,000 images, each annotated with 5 written sentences using Amazon Mechanical Turk. We used the Regional Convolution Neural Network (R-CNN) last layer ("fc7") [8] as image features (a vector of 4096 floats per image fragment), exactly as in [6]. These image features were extracted using Caffe [14].

The content words in the sentences in Pascal1K were matched with words from the LUCID speech corpus [15]. Other types of words (function words, articles etc.) were excluded, as were words occurring less than twice in the corpus. The LUCID corpus is a corpus of conversational British English by 40 speakers, and is annotated at the word level. Words that occur in Pascal1K but not in LUCID were replaced by spoken words matched for part-of-speech type, frequency, and length, such that each sentence in Pascal1K was replaced by a set of words spoken by the same speaker. The spoken words are represented as 71 stacked frames (strided by 10ms each, thus 710ms) of 40 log-compressed Mel filterbanks centered on the middle of the word, yielding a vector of 2840 floats per word token.

We designed a multi-modal neural network that takes image features and speech features as inputs. The architecture for the results presented in this paper is detailed in Fig. 1. The network is presented with pairs of "co-occurring" image fragments and spoken tokens (of content words) and these are considered "same", while if we sample randomly image fragments and spoken tokens, we consider such pairs "different" (even though there is a low probability that this is a "same" pair). We use the same number of "same" and "different" pairs during training.

For the nonlinearity, we used rectified linear units ($h(v) = \max(0, v)$) because they are used widely in state of the art deep neural networks nowadays (speech [16], vision [8]) and they performed as well (or better) as sigmoid units for our task, while being faster to train.

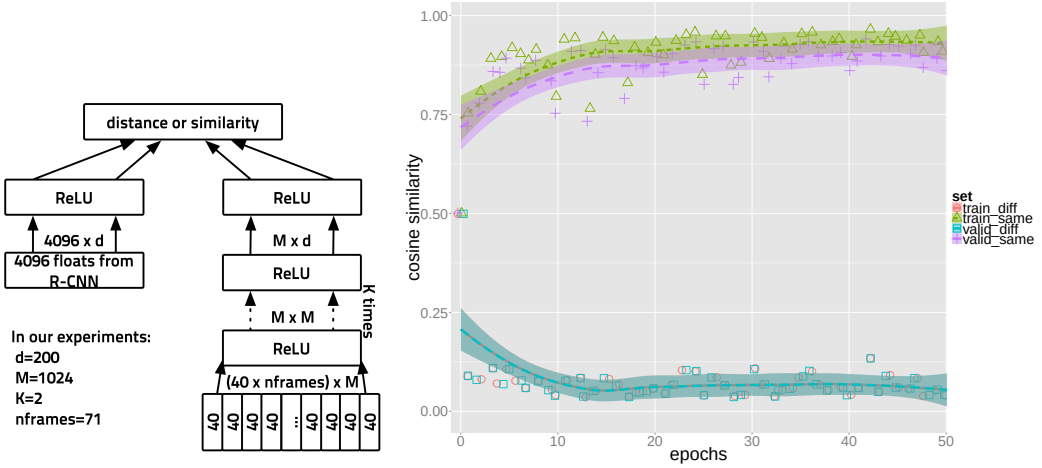

Figure 1: Left: model of our cross-situational learning neural net, the left part comes from images (R-CNN last layer features), the right part from speech sounds (stacked log-compressed Mel filterbanks). Right: training and validation costs.

For the cost function, we used the "cosine squared cosine" ($\text{Loss}_{\text{COSCOS}^2}$) as explained below. Consider Y_I and Y_S being the output representations for input image fragment I, and input spoken word token S:

$$\text{Loss}_{\text{COSCOS}^2}(I, S) = \begin{cases} (1 - \cos(Y_I, Y_S))/2 & \text{if same} \\ \cos^2(Y_I, Y_S) & \text{if different} \end{cases}$$

with

$$\cos(x, y) = \frac{\langle x, y \rangle}{\|x\| \|y\|}$$

The intuition behind this cost function is that pairs of "different" vectors should be orthogonal instead of anti-collinear (if we minimized $1 - \cos(Y_I, Y_S)$ for pairs of "different"), because anti-collinearity in a high-dimensional space (we used d=200 for the embedding) is more of a "feat" that orthogonality. On the other hand, if we used $\cos^2(Y_I, Y_S)$ also for pairs of "same" vectors, we would consider 0 and π to be equivalent. There are more details

about this cost function in our previous paper [1]. We did not try other cost functions (e.g. based on the Euclidean distance) yet.

3 Results

All the code (along with the datasets pairings from content words to spoken word tokens) to reproduce these results is free (MIT license) and available on Github[1]. The Pascal1K and the LUCID datasets are also available for free. We trained our model on the 800/100/100 images train/validation/test sets split of [6]. For the training, we added an L2 regularization term ($\lambda\|\theta\|_2^2$) with $\lambda = 10^{-4}$. We use the Adadelta [17] variant of stochastic gradient descent (with $\rho = 0.95$ and $\epsilon = 10^{-6}$), with batches of 1 image (20 fragments and 5 sentences), and a max-norm regularization of 6 (see details in [18]). We performed early stopping on the validation set. The cosine similarities are shown separately for pairs of "same" and "different", on the training and validation set, in Fig. 1.

We analyse the results from our model using a number of procedures. The first set calculates the rank of relevant neighbors in four ways. **Word Similarity:** following Socher et al. [5], we map the spoken words into the embedding space. Next, for each spoken word, we record the rank of the nearest word (in terms of cosine similarity) that occurs in the description of the same image. **Word Token Search by Image:** given an image query, find the rank of the closest word neighbor that is a token in the description of the image. **Word Type Search by Image:** given an image query, find the rank of the closest word neighbor whose transcription is in the description of the image. **Image Token Search by Word:** given a word query, find the rank of the closest image neighbor that has the word query in its description. Table 1 shows these results:

	Random		Our Model	
	Median Rank	Mean Rank	Median Rank	Mean Rank
Word Similarity	**4.00**	22.96	5.00	**11.53**
Word Token Search by Image	93.00	293.39	**40.00**	**72.52**
Word Type Search by Image	19.00	293.40	**15.00**	**30.11**
Image Token Search by Word	117.00	178.44	**81.00**	**141.29**

Table 1: Comparison of the model to a random projection of the input data. Ranks are out of 21,000 word tokens (by 40 speakers), 20,000 image fragments. Lower numbers are better.

The second set of evaluation procedures calculates the recall of relevant items in three ways. Recall is calculated at the first 1, 5, 10 and 25 neighbors. We report recall of word tokens in response to an image query, word types in response to an image query, and image tokens in response to a word query. Table 2 shows the recall scores.

	Random				Our Model			
	R@1	R@5	R@10	R@25	R@1	R@5	R@10	R@25
Word Token Search by Image	1.42	6.21	11.47	24.16	**2.45**	**10.25**	**20.15**	**39.60**
Word Type Search By Image	4.48	19.84	34.68	58.05	**6.80**	**27.30**	**43.45**	**72.20**
Image Token Search by Word	**2.04**	7.42	13.61	22.76	1.86	**9.41**	**15.99**	**42.91**

Table 2: Recall scores for the model and a random projection. Higher numbers are better.

4 Discussion

We introduced a novel task of mapping speech and images into the same space and learning the associations between them. This task was approached using a neural network trained

[1] `https://github.com/bootphon/crossitlearn`

on isolated spoken words and image segments. The results of this approach, while simple, are promising for future improvements.

As can be seen from the results, our model has more trouble finding images from words than vice-versa. In addition, the finding that the average rank in the image search task is much higher than in the word search task, leads us to conclude that the acoustic model is open for many improvements. Chosen for its simplicity, stacked log-compressed Mel filterbanks will have trouble with the multi-speaker setting of the tasks, as well as with handling the acoustic variation introduced by co-articulation with the (excluded) words in the context of the token. A more sophisticated speech representation may well improve the results significantly. Also, the word similarity score could be improved by explicitly modeling their semantics, by adding a topic modeling or word embedding representation.

Current work in progress involves improvements in the model training procedure, such as improving the cross-validation scheme. Our "cosine squared cosine" loss function, while having shown previous success, may be sub-optimal; so an exploration of different loss functions is in order, such as margin-inducing Euclidean costs as in [6]. In addition, experiments with larger datasets (Flickr8K and Flickr32K) will provide much needed insight into the generalization characteristics of the current approach. In summary, the current work is a first step in a new and challenging task, one that may provide insights useful both for developing machine learning algorithms for multi-modal learning, as well as providing an object of study for research in human acquisition of words.

5 Acknowledgements

This project is funded in part by the European Research Council (ERC-2011-AdG-295810 BOOTPHON), the Agence Nationale pour la Recherche (ANR-10-LABX-0087 IEC, ANR-10-IDEX-0001-02 PSL*), the Fondation de France, the Ecole de Neurosciences de Paris, and the Region Ile de France (DIM cerveau et pensée).

References

[1] Gabriel Synnaeve, Thomas Schatz, and Emmanuel Dupoux. Phonetics embedding learning with side information. In IEEE Speech and Language Technologies, 2014.

[2] Paul C Bomba and Einar R Siqueland. The nature and structure of infant form categories. Journal of Experimental Child Psychology, 35(2):294–328, 1983.

[3] Peter D Eimas and Paul C Quinn. Studies on the formation of perceptually based basic-level categories in young infants. Child development, 65(3):903–917, 1994.

[4] Elizabeth S Spelke. Principles of object perception. Cognitive science, 14(1):29–56, 1990.

[5] Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. Grounded compositional semantics for finding and describing images with sentences. Transactions of the ACL, 2014.

[6] Andrej Karpathy, Armand Joulin, and Li Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In Advances in Neural Information Processing Systems (NIPS), 2014.

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems (NIPS), pages 1097–1105, 2012.

[8] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

[9] Chen Yu and Linda B Smith. Rapid word learning under uncertainty via cross-situational statistics. Psychological Science, 18(5):414–420, 2007.

[10] Jane Bromley, Isabelle Guyon, Yann Lecun, Eduard Säckinger, and Roopak Shah. Signature verification using a "siamese" time delay neural network. In Advances in Neural Information Processing Systems (NIPS), 1994.

[11] E.P. Xing, I. Jordan, Michael, R.M. Karp, and S. Russell. Distance metric learning, with application to clustering with side-information. In Advances in Neural Information Processing Systems (NIPS), 2003.

[12] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In Computer vision and pattern recognition, 2006 IEEE computer society conference on, volume 2, pages 1735–1742. IEEE, 2006.

[13] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. Collecting image annotations using Amazon's Mechanical Turk. In Proceedings of the NAACL HLT2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pages 139–147, 2010.

[14] Yangqing Jia. Caffe: An open source convolutional architecture for fast feature embedding. `http://caffe.berkeleyvision.org/`, 2013.

[15] Rachel Baker and Valerie Hazan. LUCID: a corpus of spontaneous and read clear speech in british english. In DiSS-LPSS, pages 3–6, 2010.

[16] MD Zeiler, M Ranzato, R Monga, M Mao, K Yang, QV Le, P Nguyen, A Senior, V Vanhoucke, J Dean, and G.E. Hinton. On rectified linear units for speech processing. In ICASSP, 2013.

[17] Matthew D Zeiler. Adadelta: An adaptive learning rate method. arXiv preprint arXiv:1212.5701, 2012.

[18] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 15(1):1929–1958, 2014.