Efficient Large-Scale Multi-Modal Classification

Douwe Kiela, Edouard Grave, Armand Joulin, Tomas Mikolov

Facebook AI Research

{dkiela,egrave,ajoulin,tmikolov}@fb.com

Abstract

While the incipient internet was largely text-based, the modern digital world is becoming increasingly multi-modal. Here, we examine multi-modal classification where one modality is discrete, e.g. text, and the other is continuous, e.g. visual representations transferred from a convolutional neural network. In particular, we focus on scenarios where we have to be able to classify large quantities of data quickly. We investigate various methods for performing multi-modal fusion and analyze their trade-offs in terms of classification accuracy and computational efficiency. Our findings indicate that the inclusion of continuous information improves performance over text-only on a range of multi-modal classification tasks, even with simple fusion methods. In addition, we experiment with discretizing the continuous features in order to speed up and simplify the fusion process even further. Our results show that fusion with discretized features outperforms text-only classification, at a fraction of the computational cost of full multi-modal fusion, with the additional benefit of improved interpretability.

Text classification is one of the core problems in machine learning and natural language processing (Borko and Bernick 1963; Sebastiani 2002). It plays a crucial role in important tasks ranging from document retrieval and categorization to sentiment and topic classification (Deerwester et al. 1990; Joachims 1998; Pang and Lee 2008). However, while the incipient Web was largely text-based, the recent decade has seen a surge in multi-modal content: billions of images and videos are posted and shared online every single day. That is, text is either replaced as the dominant modality, as is the case with Instagram posts or YouTube videos, or it is augmented with non-textual content, as with most of today's web pages. This makes multi-modal classification an important problem.

Here, we examine the task of multi-modal classification using neural networks. We are primarily interested in two questions: what is the best way to combine (i.e., fuse) data from different modalities, and how can we do so in the most efficient manner? We examine various efficient multi-modal fusion methods and investigate ways to speed up the fusion process. In particular, we explore discretizing the continuous features, which leads to much faster training and requires

less storage, yet is still able to benefit from the inclusion of multi-modal information. To the best of our knowledge, this work constitutes the first attempt to examine the accuracy/speed trade-off in multi-modal classification; and the first to directly show the value of discretized features in this particular task.

If current trends continue, the Web will become increasingly multi-modal, making the question of multi-modal classification ever more pertinent. At the same time, as the Web keeps growing, we have to be able to efficiently handle ever larger quantities of data, making it important to focus on machine learning methods that can be applied to large-scale scenarios. This work aims to examine these two questions together.

Our contributions are as follows. First, we compare various multi-modal fusion methods, examine their trade-offs, and show that simpler models are often desirable. Second, we experiment with discretizing continuous features in order to speed up and simplify the fusion process even further. Third, we examine learned representations for discretized features and show that they yield interpretability as a beneficial side effect. The work reported here constitutes a solid and scalable baseline for other approaches to follow; our investigation of discretized features shows how multi-modal classification does not necessarily imply a large performance penalty and is feasible in large-scale scenarios.

Related work

Text classification. Neural network-based methods have become increasingly popular for text classification (Socher et al. 2011; Wang and Manning 2012). Recent work has used neural networks for text classification either at a sentence (Kim 2014; Hill, Cho, and Korhonen 2016) or full document (Le and Mikolov 2014; Baker, Kiela, and Korhonen 2016; Joulin et al. 2016) level. Many core NLP tasks are essentially text classification, from tweets (Sriram et al. 2010) to reviews to spam. Even though there has been extensive work on feature engineering for text classification (Chen et al. 2009), modern approaches often make use of word embeddings (Mikolov et al. 2013) or sentence representations (Kiros et al. 2015) learned from a large corpus in an unsupervised fashion.

Fusion strategies. Multi-modal fusion, or the integration of input from various modalities, is an important topic in the field of multimedia analysis (Wu et al. 2004; Atrey et al. 2010). The question of fusion has been explored in a variety of tasks, from audio-visual speech recognition (Potamianos et al. 2003) to multi-sensor management (Zhao et al. 2003) and face recognition (Xiong and Svensson 2002). Much of this research has focused on the combination of two or more continuous modalities. Here, we are specifically interested in the fusion of discrete textual input with another, continuous, modality.

Multi-modal NLP. The usage of non-textual information in natural language processing (Mooney 2008) has become increasingly popular. On the one hand, there has been a lot of interest in cross-modal applications, such as image annotation (Weston, Bengio, and Usunier 2011), image captioning (Bernardi et al. 2016), mapping images to text or vice versa (Frome et al. 2013; Socher et al. 2013; Lazaridou, Bruni, and Baroni 2014) and visual question answering (Antol et al. 2015; Fukui et al. 2016). On the other hand, multi-modal fusion has been extensively explored in the context of grounded representation learning for lexical semantics (Bruni, Tran, and Baroni 2014; Kiela and Bottou 2014; Lazaridou, Pham, and Baroni 2015). While much of this work has focused on vision (Baroni 2016), other perceptual modalities modalities (Lopopolo and van Miltenburg 2015; Kiela and Clark 2015; Kiela, Bulat, and Clark 2015) have also been explored, as well as robotics (Mei, Bansal, and Walter 2016), videos (Regneri et al. 2013) and games (Branavan, Silver, and Barzilay 2012; Narasimhan, Kulkarni, and Barzilay 2015). This work is similar in spirit to (Bruni, Tran, and Baroni 2014), in that we explore fusion techniques. However, similarly to (Lazaridou, Pham, and Baroni 2015), we learn how to integrate the multi-modal inputs, and use transferred representations as in (Kiela and Bottou 2014).

Multi-modal deep learning. Our work relates to previous work on integrating information from multiple modalities in neural networks (Ngiam et al. 2011; Srivastava and Salakhutdinov 2012; Kiros, Salakhutdinov, and Zemel 2014). Here, we enhance a well-known neural network architecture for efficient text classification (Joulin et al. 2016) with the ability to include continuous information, and explore methods for combining multi-modal features. The works of (Arevalo et al. 2017) and (Fukui et al. 2016), explore complex gating mechanisms and compact bilinear pooling as multi-modal fusion methods. In order to obtain visual representations, we transfer continuous features from neural networks trained on other tasks (in this case ImageNet), as has been shown to work well for a wide variety of tasks (Oquab et al. 2014; Razavian et al. 2014).

Evaluation

Surprisingly, there are not many large-scale multi-modal classification datasets available. We evaluate on three datasets that are large enough to examine accuracy/speed trade-offs in a meaningful way. Two of our datasets (Food101 and MM-IMDB) are medium-sized; while the third dataset (FlickrTag) is very large by today's standards. The quantitative properties of the respective datasets are shown in Table 1 and they are described in more detail in what follows.

Food101

The UPMC Food101 dataset (Wang et al. 2015) contains web pages with textual recipe descriptions for 101 food labels automatically retrieved online. Each page was matched with a single image, where the images were obtained by querying Google Image Search for the given category. Examples of food labels are *Filet Mignon*, *Pad Thai*, *Breakfast Burrito* and *Spaghetti Bolognese*. The web pages were processed with html2text[1] to obtain the raw text.

MM-IMDB

The recently introduced MM-IMDB dataset (Arevalo et al. 2017) contains movie plot outlines and movie posters. The objective is to classify the movie by genre. This is a multi-label prediction problem, i.e., one movie can have multiple genres. The dataset was specifically introduced to address the lack of multi-modal classification datasets.

FlickrTag and FlickrTag-1

We use the FlickrTag dataset based on the massive YFCC100M Flickr dataset of (Thomee et al. 2016) that was used in (Joulin et al. 2016). The dataset consists of Flickr photographs together (in most, but not all cases) with short user-provided captions. The objective is to predict the user-provided tags that belong to the photograph. This is a very large-scale dataset, so we perform the multi-modal fusion operator and speed-versus-accuracy studies on a subset (specifically, the first shard, which corresponds to one-tenth of the full dataset) for those studies, which we denote FlickrTag-1. We show that the inclusion of discretized features yields classification accuracy improvements with respect to text on the whole dataset.

Approach

As a starting point, we take the highly efficient text classification approach of FastText (Joulin et al. 2016). To ensure a fair comparison, we enhance that model with the capability to handle continuous or discretized features. Specifically, we use 2048-dimensional continuous features that were obtained by transferring the pre-softmax layer of a 152-layer ResNet (He et al. 2016) trained on the ImageNet classification task. In the case of the large-scale FlickrTag datasets, we use ResNet-34 features (of 512 dimensions). It has been shown that convolutional network features can be transferred successfully to a variety of tasks (Razavian et al. 2014) and we take the same approach here. We explore a variety of models and experiment with discretization.

The scenario of multi-modal classification certainly admits, or even invites, highly sophisticated models. In our

---

[1]https://pypi.python.org/pypi/html2text

Dataset	#Train	#Words	#Valid	#Words	#Test	#Words
Food101	58,131	98,365,392	6,452	10,893,597	21,519	36,955,182
MM-IMDB	15,552	2,564,734	2608	425,863	7799	1,266,681
FlickrTag	70,243,104	1,134,118,808	656,687	10,100,945	621,444	9,913,566
FlickrTag-1	7,166,110	92,651,036	48,048	682,663	48,471	672,900

Table 1: Evaluation datasets with their quantitative properties.

case, however, we also have to take into account efficiency, so we want to focus on models that are simple and efficient enough to handle large-scale datasets, while obtaining improved performance over our baselines. We experiment with a comprehensive set of models, listed below in increasing order of complexity.

In all cases, given a set of N documents, the objective is to minimize the negative log likelihood over the classes:

$$-\frac{1}{N}\sum_{n=1}^{N}\log(softmax(o(x_n), y_n)), \qquad (1)$$

where o is the network's output, x_n is the multi-modal input and y_n is the label.

Baselines

Text   The first baseline consists of FastText (Joulin et al. 2016), a library for highly efficient word representation learning and sentence classification. FastText is trained asynchronously on multiple CPUs using stochastic gradient descent and a learning rate that linearly decays with the amount of words. It yields competitive performance with more sophisticated text classification approaches, while being much more efficient. That is, we ignore the visual signal altogether and only use textual information, i.e.,

$$o(x_n) = WUx_n^t,$$

where W and U are weight matrices and x_n^t is the normalized bag of textual features representation.

Continuous   The second baseline consists of training a classifier only on top of the transferred ResNet features (He et al. 2016). That is, we ignore the textual information and only use the visual input, i.e.,

$$o(x_n) = WVx_n^v,$$

where W and V are weight matrices and x_n^v consists of the ResNet features, normalized to unit length.

Continuous Multi-Modal Models

Additive   We combine the information from both modalities using component-wise addition, i.e.,

$$o(x_n) = W(Ux_n^t + Vx_n^v).$$

Max-pooling   We combine the information from both modalities using the component-wise maximum, i.e.,

$$o(x_n) = W\max(Ux_n^t, Vx_n^v).$$

Gated   We allow one modality to "gate" or "attend" over the other modality, via a sigmoid non-linearity, i.e.,

$$o(x_n) = W(\sigma(Ux_n^t) * Vx_n^v),$$

or alternatively,

$$o(x_n) = W(Ux_n^t * \sigma(Vx_n^v)).$$

One can think of this approach as performing attention from one modality over the other. It is a conceptually similar simplification of multi-modal gated units, introduced in (Arevalo et al. 2017). The modality to be gated is a hyperparameter (see below).

Bilinear   Finally, to fully capture any associations between the two different modalities, we examine a bilinear model, i.e.,

$$o(x_n) = W(Ux_n^t \otimes Vx_n^v).$$

This approach can be thought of as a simpler version of the more complex multi-modal bilinear pooling introduced by (Fukui et al. 2016). We also experiment with a method where we introduce a gating non-linearity into the bilinear model, which we call **Bilinear-Gated**.

Discretized Multi-Modal Models

A downside of continuous models is that they require an expensive matrix-vector multiplication Vx_i^v and storing large matrices of floating point numbers requires a lot of space. While the ResNet features used in these experiments consist of a relatively small number of components, these can easily run into the tens of thousands: consider e.g. combinations of SIFT and Fisher vectors used in state-of-the-art computer vision applications (Perronnin and Larlus 2015). Hence, we experiment with discretizing the continuous features, where we convert the continuous features to a discrete sequence of tokens, which can be treated as if they are special tokens, which we normalize separately, and used in the standard FastText setup. This is a simple, computationally less intensive solution. Discretized features also obviously require less storage.

In particular, we investigate product quantization (PQ) (Jegou, Douze, and Schmid 2011), where we divide the continuous vector into subvectors of equal size, and then perform k-means clustering on each of the subvectors. For each image, we subsequently determine the closest centroid for each of its subwords, which is combined with the subvector index in order to obtain a discretized vector. For example, a

	Model	Food101	MM-IMDB	FlickrTag-1
Previous work	(Wang et al. 2015)	85.1	—	—
	(Arevalo et al. 2017)-GMU	—	**63.0**	—
	(Arevalo et al. 2017)-AVG	—	61.5	—
Baselines	FastText	88.0 ± 0.1	58.8 ± 0.1	23.0 ± 0.0
	Continuous	56.7 ± 0.2	49.3 ± 0.0	12.4 ± 0.0
Continuous	Additive	90.4 ± 0.1	61.0 ± 0.0	26.8 ± 0.0
	Max-pooling	90.5 ± 0.1	62.2 ± 0.1	26.9 ± 0.0
	Gated	90.1 ± 0.1	61.8 ± 0.1	27.7 ± 0.0
	Bilinear	88.1 ± 0.3	61.5 ± 0.1	27.8 ± 0.0
	Bilinear-gated	$\mathbf{90.8 \pm 0.1}$	$\mathbf{62.3 \pm 0.2}$	$\mathbf{28.6 \pm 0.0}$
Discretized	PQ (n=4, k=256)	89.5 ± 0.1	60.5 ± 0.1	25.6 ± 0.1
	RSPQ (n=4, k=256, r=4)	89.8 ± 0.0	60.7 ± 0.1	26.1 ± 0.1

Table 2: Accuracy (averaged over 5 runs) of continuous and discretized multi-modal models, compared to baselines.

100-dim continuous vector x_i^v may be divided into ten 10-dimensional subvectors s_i. Let $N(s_i)$ denote the index of the nearest centroid for s_i. The discretized representation of v is then given as $\langle (1, N(s_1)), (2, N(s_2)), .., (10, N(s_{10})) \rangle$. We include these tokens in the text and treat them as if they were special tokens, in the standard fastText model, i.e.,

$$o(x_n) = W(Ux_n^t + \alpha Ux_n^d).$$

where x_n^d are the discretized features and α is a reweighting hyperparameter. We normalize x_n^t and x_n^d independently. As we can see, the discretized models are closely related to the additive model, except that they use the same weight matrix U with the discretized features used as "words" in the text.

While PQ is great for compressing information into a discretized sequence, it does impose hard boundaries on subvectors, which means that overlapping semantic content that is shared between subvectors may be lost. Hence, we introduce a novel quantization method, called random sample product quantization (RSPQ), in order to maintain (at least some) overlapping semantic information. In RSPQ, the process is the same as in PQ, except we perform PQ over r repetitions of random permutations of x_i^v. In both cases, we treat the discretized features as if they are reweighted special tokens included in the textual data and run standard fastText.

Model complexity

There are various trade-offs at stake between these models. The additive, max-pooling and gated models are simplest and result in a hidden layer of the same size as with the normal FastText. The computational complexity of the linear classifier is thus $O(HK)$, where K is the number of classes and H is the size of Ux^t and Vx^v. The max-pooling and gating models are slightly more complicated than the additive one, requiring an extra operation. For the bilinear model the complexity amounts to $O(H^2K)$. Thus, the bilinear model is by far the most expensive to compute. The additive model has the benefit that it does not strictly require a continuous input at all times.

Hyperparameters and training

In all experiments, the model is tuned on the validation set. We tried the following hyperparameters: a learning rate in $\{0.1, 0.25, 0.5, 1.0, 2.0\}$, a number of epochs in $\{5, 10, 20\}$, a reweighting parameter in $\{0.01, 0.02, 0.05, 0.1, 0.2, 0.5\}$ and an embedding dimensionality of either 20 or 100. These hyperparameters were swept using grid search and we used a softmax loss. For other hyperparameters, such as the number of threads in the parallel optimization and the minimum word count, we fixed their values to standard values in FastText (4 threads, minimum count of 1, respectively), since we found that these did not impact classification accuracy. In the case of the gated and bilinear-gated models, the modality used to serve as a gate over the other modality is treated as a hyperparameter as well.

Results

The results of the comparison may be found in Table 2. We compare the continuous and discretized multi-modal models against the text-only FastText model and to the continuous features-only model. We also include results from (Wang et al. 2015) on Food-101, where they used TF-IDF features for text and a deep convolutional neural network features for images, as well as results from (Arevalo et al. 2017) for Gated Multimodal Units (GMU) and their AVG_Probs model. GMUs are a substantially more complicated model architecture than any of our relatively simple fusion methods, so this study is a good test of their capability. We note that in the case of Food101, our methods work considerably better than previously reported results. For MM-IMDB, the continuous multi-modal models perform very close to the GMU model and outperform the AVG_Probs method, while being simpler and computationally more efficient.

We observe that multi-modal models always outperform standard FastText and the continuous-only approach, disregarding the particular type of fusion. This shows that the inclusion of multi-modal information (at least in these types of classification tasks) always helps and that making use of

Model	Train time (FlickrTag-1)
FastText	0h01m
Additive	0h39m
Max-pooling	0h39m
Gated	0h40m
Bilinear	1h04m
Bilinear-Gated	1h06m
PQ	0h01m
RSPQ	0h02m

Table 3: Training time on FlickrTag-1

Model	FlickrTag
FastText	36.7
PQ (n=4, k=256)	38.9
RSPQ (n=4, k=256, r=4)	39.4

Table 4: Performance on full FlickrTag.

$\_\_q\_\_0\_253$	$\_\_q\_\_0\_13$	$\_\_q\_\_1\_253$
donuts 0.987	crème 0.933	oishii 0.905
doughnuts 0.981	ramekins 0.928	shoga 0.885
donut 0.980	brulee 0.925	tenkasu 0.884
doughnut 0.979	brûlée 0.916	octopus 0.883
donuts? 0.939	custards 0.916	aonori 0.881

Table 5: Examples of nearest neighbors for quantized features in Food101.

multi-modal information, where available, will lead to increased performance. FastText outperforms the continuous-only method on all datasets, which indicates that text plays a big role in these tasks, and that it is relatively more important than the visual information.

If we examine the continuous multi-modal models, we see that the bilinear-gated model is the clear winner: it outperforms all other methods on all three tasks. It is however also the most complicated model, and as a result is less efficient. We found that placing the gating non-linearity on the text led to the best performance on Food101 and MM-IMDB, while placing it on the visual modality led to the best performance on FlickrTag. It is interesting to observe that the more complicated gated model, as well as the non-gated bilinear model, do not necessarily outperform the simpler additive and max-pooling models. In fact, the performance of these much simpler models is not too far removed from the best scores. The take-away message appears to be: if you care more about accuracy, use the bilinear method with gating; if you care more about speed, use a simple model like the additive or max-pooling one, which have the additional potential benefit that they do not necessarily require the presence of continuous information if none is available.

Speed

We can draw inspiration from the fact that the additive model performed reasonably well: if speed is essential—if necessary at the expense of some accuracy—the discretized models are an obvious choice to further simplify and speed up the model. Even though they outperform standard FastText by a large margin, as shown in Table 2, they only come with a minor performance impact. Table 3 shows the training times for the various models on the FlickrTag-1 dataset: while the bilinear models take around one hour to train (recall that this constitutes only the first shard of the full dataset); the discretized methods, similarly to FastText, only take around

one minute. If we scale up to the full FlickrTag dataset, Table 4 shows that the discretized models substantially outperform standard FastText. An increase of 2.7% in accuracy, as seen from FastText to RSPQ, represents having an additional 16778 test set documents correctly classified using that model, which is a non-negligible amount.

Interpretability

An interesting side effect of the discretized multi-modal methods is that they allow us to examine the nearest neighbors of the quantized features: if a particular feature corresponds to something that looks like a donut, for example, then its embedding should be close to words related to *donut*. Indeed, as Table 5 shows, we can find clearly identifiable clusters, e.g. for donuts, crème brûlée and certain types of Japanese food. Interpretability is an important but often overlooked aspect of classification models: we show that a simple and efficient method, that outperforms text-only methods by a large margin, yields the additional benefit that it allows for the interpretation of the visual features that a classifier picks up on—something that is difficult to achieve with standard convolutional features.

Conclusion & Outlook

The internet is becoming increasingly multi-modal, which makes the task of multi-modal classification ever more pertinent. In order to be able to handle large quantities of data, we need efficient models for large-scale multi-modal classification. In this work, we examined these two questions together. First, we compared various multi-modal fusion methods and found a bilinear-gated model to achieve the highest accuracy, while the simpler additive and max-pooling models yielded reasonably high accuracy at higher speed. Second, we showed that the model can be speeded up even further by introducing discretized multi-modal features. Lastly, we showed that this method yields the additional benefit of interpretability, where we can examine what the multi-modal model picks up on when making its classification decision. We hope that this work can serve as a useful baseline for further work in multi-modal classification.

Acknowledgments

References

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; and Parikh, D. 2015. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, 2425–2433.

Arevalo, J.; Solorio, T.; Montes-y Gómez, M.; and González, F. A. 2017. Gated multimodal units for information fusion.

Atrey, P. K.; Hossain, M. A.; El Saddik, A.; and Kankanhalli, M. S. 2010. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems* 16(6):345–379.

Baker, S.; Kiela, D.; and Korhonen, A. 2016. Robust text classification for sparsely labelled data using multi-level embeddings. In *Proceedings of COLING*, 2333–2343.

Baroni, M. 2016. Grounding distributional semantics in the visual world. *Language and Linguistics Compass* 10(1):3–13.

Bernardi, R.; Cakici, R.; Elliott, D.; Erdem, A.; Erdem, E.; Ikizler-Cinbis, N.; Keller, F.; Muscat, A.; and Plank, B. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *J. Artif. Intell. Res.(JAIR)* 55:409–442.

Borko, H., and Bernick, M. 1963. Automatic document classification. *Journal of the ACM* 10(2):151–162.

Branavan, S.; Silver, D.; and Barzilay, R. 2012. Learning to win by reading manuals in a monte-carlo framework. *Journal of Artificial Intelligence Research* 43:661–704.

Bruni, E.; Tran, N.-K.; and Baroni, M. 2014. Multimodal distributional semantics. *J. Artif. Intell. Res.(JAIR)* 49(2014):1–47.

Chen, J.; Huang, H.; Tian, S.; and Qu, Y. 2009. Feature selection for text classification with naïve bayes. *Expert Systems with Applications* 36(3):5432–5435.

Deerwester, S.; Dumais, S. T.; Furnas, G. W.; Landauer, T. K.; and Harshman, R. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science* 41(6):391.

Frome, A.; Corrado, G. S.; Shlens, J.; Bengio, S.; Dean, J.; Mikolov, T.; et al. 2013. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, 2121–2129.

Fukui, A.; Park, D. H.; Yang, D.; Rohrbach, A.; Darrell, T.; and Rohrbach, M. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. *CoRR* abs/1606.01847.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hill, F.; Cho, K.; and Korhonen, A. 2016. Learning distributed representations of sentences from unlabelled data. *arXiv preprint arXiv:1602.03483*.

Jegou, H.; Douze, M.; and Schmid, C. 2011. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(1):117–128.

Joachims, T. 1998. Text categorization with support vector machines: Learning with many relevant features. In *European Conference on Machine Learning*, 137–142.

Joulin, A.; Grave, E.; Bojanowski, P.; and Mikolov, T. 2016. Bag of tricks for efficient text classification. *CoRR* abs/1607.01759.

Kiela, D., and Bottou, L. 2014. Learning image embeddings using convolutional neural networks for improved multimodal semantics. In *EMNLP*, 36–45.

Kiela, D., and Clark, S. 2015. Multi-and cross-modal semantics beyond vision: Grounding in auditory perception. In *EMNLP*, 2461–2470.

Kiela, D.; Bulat, L.; and Clark, S. 2015. Grounding semantics in olfactory perception. In *ACL (2)*, 231–236.

Kim, Y. 2014. Convolutional neural networks for sentence classification. In *Proceedings of EMNLP*.

Kiros, R.; Zhu, Y.; Salakhutdinov, R.; Zemel, R. S.; Torralba, A.; Urtasun, R.; and Fidler, S. 2015. Skip-thought vectors. In *Proceedings of NIPS*.

Kiros, R.; Salakhutdinov, R.; and Zemel, R. S. 2014. Multimodal neural language models. In *Proceedings of ICML*, volume 14, 595–603.

Lazaridou, A.; Bruni, E.; and Baroni, M. 2014. Is this a wampimuk? cross-modal mapping between distributional semantics and the visual world. In *ACL (1)*, 1403–1414.

Lazaridou, A.; Pham, N. T.; and Baroni, M. 2015. Combining language and vision with a multimodal skip-gram model. *arXiv preprint arXiv:1501.02598*.

Le, Q. V., and Mikolov, T. 2014. Distributed representations of sentences and documents. In *ICML*, volume 14, 1188–1196.

Lopopolo, A., and van Miltenburg, E. 2015. Sound-based distributional models. *IWCS 2015* 70.

Mei, H.; Bansal, M.; and Walter, M. R. 2016. Listen, attend, and walk: Neural mapping of navigational instructions to action sequences. In *Proceedings of AAAI*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Mooney, R. J. 2008. Learning to connect language and perception. In *Proceedings of AAAI*.

Narasimhan, K.; Kulkarni, T.; and Barzilay, R. 2015. Language understanding for text-based games using deep reinforcement learning. In *Proceedings of EMNLP*.

Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; and Ng, A. Y. 2011. Multimodal deep learning. In *Proceedings of ICML*, 689–696.

Oquab, M.; Bottou, L.; Laptev, I.; and Sivic, J. 2014. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1717–1724.

Pang, B., and Lee, L. 2008. Opinion mining and sentiment

analysis. *Foundations and trends in information retrieval* 2(1-2):1–135.

Perronnin, F., and Larlus, D. 2015. Fisher vectors meet neural networks: A hybrid classification architecture. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3743–3752.

Potamianos, G.; Neti, C.; Gravier, G.; Garg, A.; and Senior, A. W. 2003. Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE* 91(9):1306–1326.

Razavian, A. S.; Azizpour, H.; Sullivan, J.; and Carlsson, S. 2014. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 806–813.

Regneri, M.; Rohrbach, M.; Wetzel, D.; Thater, S.; Schiele, B.; and Pinkal, M. 2013. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics* 1:25–36.

Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM Computing Surveys* 34(1):1–47.

Socher, R.; Pennington, J.; Huang, E. H.; Ng, A. Y.; and Manning, C. D. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the conference on empirical methods in natural language processing*, 151–161. Association for Computational Linguistics.

Socher, R.; Ganjoo, M.; Manning, C. D.; and Ng, A. 2013. Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*, 935–943.

Sriram, B.; Fuhry, D.; Demir, E.; Ferhatosmanoglu, H.; and Demirbas, M. 2010. Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, 841–842. New York, NY, USA: ACM.

Srivastava, N., and Salakhutdinov, R. R. 2012. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, 2222–2230.

Thomee, B.; Shamma, D.; Friedland, G.; Elizalde, B.; Ni, K.; Poland, D.; Borth, D.; and Li, L. 2016. YFCC100M: The New Data in Multimedia Research. *Communications of the ACM* 59(2):64–73.

Wang, S., and Manning, C. D. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, 90–94. Association for Computational Linguistics.

Wang, X.; Kumar, D.; Thome, N.; Cord, M.; and Precioso, F. 2015. Recipe recognition with large multimodal food dataset. In *Workshop CEA of the IEEE International Conference on Multimedia & Exposition (ICME)*, 1–6. IEEE.

Weston, J.; Bengio, S.; and Usunier, N. 2011. Wsabie: Scaling up to large vocabulary image annotation. In *IJCAI*, volume 11, 2764–2770.

Wu, Y.; Chang, E. Y.; Chang, K. C.-C.; and Smith, J. R. 2004. Optimal multimodal fusion for multimedia data analysis. In *Proceedings of the 12th annual ACM international conference on Multimedia*, 572–579. ACM.

Xiong, N., and Svensson, P. 2002. Multi-sensor management for information fusion: issues and approaches. *Information fusion* 3(2):163–186.

Zhao, W.; Chellappa, R.; Phillips, P. J.; and Rosenfeld, A. 2003. Face recognition: A literature survey. *ACM computing surveys (CSUR)* 35(4):399–458.