Gaussian Visual-Linguistic Embedding for Zero-Shot Recognition

Tanmoy Mukherjee and **Timothy Hospedales**
Queen Mary University of London
School of Electronic Engineering and Computer Science
{k.m.tanmoy,t.hospedales}@qmul.ac.uk

Abstract

An exciting outcome of research at the intersection of language and vision is that of zero-shot learning (ZSL). ZSL promises to scale visual recognition by borrowing distributed semantic models learned from linguistic corpora and turning them into visual recognition models. However the popular word-vector DSM embeddings are relatively impoverished in their expressivity as they model each word as a single vector point. In this paper we explore word-*distribution* embeddings for ZSL. We present a visual-linguistic mapping for ZSL in the case where words and visual categories are both represented by distributions. Experiments show improved results on ZSL benchmarks due to this better exploiting of intra-concept variability in each modality

1 Introduction

Learning vector representations of word meaning is a topical area in computational linguistics. Based on the distributional hypothesis (Harris, 1954) – that words in similar context have similar meanings – distributed semantic models (DSM)s build vector representations based on corpus-extracted context. DSM approaches such as topic models (Blei et al., 2003), and more recently neural networks (Collobert et al., 2011; Mikolov et al., 2013) have had great success in a variety of lexical and semantic tasks (Arora et al., 2015; Schwenk, 2007).

However despite their successes, classic DSMs are severely impoverished compared to humans due to learning solely from word cooccurrence without grounding in the outside world. This has motivated a wave of recent research into multi-modal and cross-modal learning that aims to *ground* DSMs in non-linguistic modalities (Bruni et al., 2014; Kiela and Bottou, 2014; Silberer and Lapata, 2014; **?**). Such multi-modal DSMs are attractive because they learn richer representations than language-only models (e.g., that bananas are *yellow* fruit (Bruni et al., 2012b)), and thus often outperform language only models in various lexical tasks (Bruni et al., 2012a).

In this paper, we focus on a key unique and practically valuable capability enabled by cross-modal DSMs: that of zero-shot learning (ZSL). Zero-shot recognition aims to recognise visual categories in the absence of any training examples by cross-modal transfer from language. The idea is to use a limited set of training data to learn a linguistic-visual mapping and then apply the induced function to map images from novel visual categories (unseen during training) to a linguistic embedding: thus enabling recognition in the absence of visual training examples. ZSL has generated big impact (Lampert et al., 2009; Socher et al., 2013; Lazaridou et al., 2014) due to the potential of leveraging language to help visual recognition scale to many categories without labor intensive image annotation.

DSMs typically generate *vector* embeddings of words, and hence ZSL is typically realised by variants of vector-valued cross-modal regression. However, such vector representations have limited expressivity – each word is represented by a point, with no notion of intra-class variability. In this paper, we consider ZSL in the case where both visual and linguistic concepts are represented by *Gaussian distribution* embeddings. Specifically, our Gaussian-

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 912–918,
Austin, Texas, November 1-5, 2016.

embedding approach to ZSL learns concept distributions in both domains: Gaussians representing individual words (as in (Vilnis and McCallum, 2015)) and Gaussians representing visual concepts. Simultaneously, it learns a cross-domain mapping that warps language-domain Gaussian concept representations into alignment with visual-domain concept Gaussians. Some existing vector DSM-based cross-modal ZSL mappings (Akata et al., 2013; Frome et al., 2013) can be seen as special cases of ours where the within-domain model is pre-fixed as vector corresponding to the Gaussian means alone, and only the cross-domain mapping is learned. Our results show that modeling linguistic and visual concepts as Gaussian distributions rather than vectors can significantly improve zero-shot recognition results.

2 Methodology

2.1 Background

Vector Word Embeddings In a typical setup for unsupervised learning of word-vectors, we observe a sequence of tokens $\{w_i\}$ and their context words $\{c(w)_i\}$. The goal is to map each word w to a d-dimensional vector e_w reflecting its distributional properties. Popular skip-gram and CBOW models (Mikolov et al., 2013), learn a matrix $W \in \mathbb{R}^{|V| \times d}$ of word embeddings for each of V vocabulary words ($e_w = W_{(w,:)}$) based on the objective of predicting words given their contexts.

Another way to formalise a word vector representation learning problem is to search for a representation W so that words w have high representational similarity with co-occuring words $c(w)$, and low similarity with representations of non-co-occurring words $\neg c(w)$. This could be expressed as optimisation of max-margin loss J; requiring that each word w's representation e_w is more similar to that of context words e_p than non-context words e_n.

$$J(W) = \sum_{w, w_p \in c(w), w_n \in \neg c(w)} \max(0, \delta - E(e_w, e_{w_p}) + E(e_w, e_{w_n})) \quad (1)$$

where similarity measure $E(\cdot, \cdot)$ is a distance in $\mathbb{R}^d$ space such as cosine or euclidean.

Gaussian Word Embeddings Vector-space models are successful, but have limited expressivity in terms of modelling the variance of a concept, or asymmetric distances between words, etc. This has motivated recent work into *distribution*-based embeddings (Vilnis and McCallum, 2015). Rather than learning word-vectors e_w, the goal here is now to learn a distribution for each word, represented by a per-word mean μ_w and covariance Σ_w.

In order to extend word representation learning approaches such as Eq. (1) to learning Gaussians, we need to replace vector similarity measure $E(\cdot, \cdot)$ with a similarity measure for Gaussians. We follow (Vilnis and McCallum, 2015) in using the inner product between distributions f and g – the probability product kernel (Jebara et al., 2004).

$$E(f, g) = \int_{x \in \mathbb{R}^n} f(x) g(x). \quad (2)$$

The probability product kernel (PPK) has a convenient closed form in the case of Gaussians:

$$\begin{aligned} E(f, g) &= \int_{x \in \mathbb{R}^n} \mathcal{N}(x; \mu_f, \Sigma_f) \mathcal{N}(x; \mu_g, \Sigma_g) dx \\ &= \mathcal{N}(0; \mu_f - \mu_g, \Sigma_f + \Sigma_g) \end{aligned} \quad (3)$$

where μ_f, μ_g are the means and Σ_f, Σ_g are the covariances of the probability distribution f and g.

2.2 Cross-Modal Distribution Mapping

Gaussian models of words can be learned as in the previous section, and that Gaussian models of image categories can be trivially obtained by maximum likelihood. The central task is therefore to establish a mapping between word-and image-Gaussians, which will be of different dimensions d_w and d_x.

We aim to find a projection matrix $A \in \mathbb{R}^{d_x \times d_w}$ such that a word w generates an image vector as $e_x = A e_w$. Working with distributions, this implies that we have $\mu_x = A \mu_w$ and $\Sigma_x = A \Sigma_w A^T$. We can now evaluate the similarity of concept distributions across modalities. The similarity between image-and text-domain Gaussians f and g is:

$$E(f, g) = \mathcal{N}(0; \mu_f - A\mu_g, \Sigma_f + A \Sigma_g A^T) \quad (4)$$

Using this metric, we can train our cross-modal projection A via the cross-domain loss:

$$J(A) = \sum_{f, g \in P,\ h, k \in N} \max(0, \delta - E(f, g) + E(h, k)) \quad (5)$$

where P is the set of matching pairs that should be aligned (e.g., the word Gaussian 'plane' and the Gaussian of plane images) and N is the set of mismatching pairs that should be separated (e.g., 'plane' and images of dogs). This can be optimised with SGD using the gradient:

$$\begin{aligned}\frac{\partial E}{\partial A} &= \frac{1}{2}((\Sigma_f + A\Sigma_g A^T)^{-1}A(\Sigma_g + \Sigma_g^T)) \\ &+ ((\mu_g^T(\Sigma_f + A\Sigma_g A^T)^{-1}(\mu_f - A\mu_g) \\ &+ (\mu_f - A\mu_g)^T(\Sigma_f + A\Sigma_g A^T)^{-1}\mu_g^T \\ &+ (\mu_f - A\mu_g)^T(\Sigma_f + A\Sigma_g A^T)^{-1} \\ &A^T(\Sigma_g + \Sigma_j^T)(\Sigma_f + A\Sigma_g A^T)^{-1}(\mu_f - A\mu_g))\end{aligned}$$

2.3 Joint Representation and Mapping

The cross-domain mapping A can be learned (Eq. 5) for fixed within-domain representations (word and image Gaussians). It is also possible to simultaneously learn the text and image-domain gaussians ($\{\mu_i, \Sigma_i\}^{text}, \{\mu_j, \Sigma_j\}^{img}$) by optimising the sum of three coupled losses: Eq. 1 with Eq. 3, Eq. 5 and max-margin image-classification using Gaussians. We found jointly learning the image-classification Gaussians did not bring much benefit over the MLE Gaussians, so we only jointly learn the text Gaussians and cross-domain mapping.

2.4 Application to Zero-Shot Recognition

Once the text-domain Gaussians and cross-domain mapping have been trained for a set of known words/classes, we can use the learned model to recognise any novel/unseen but name-able visual category w as follows: 1. Get the word-Gaussians of target categories w, $\mathcal{N}(\mu_w, \Sigma_w)$. 2. Project those Gaussians to image modality, $\mathcal{N}(A\mu_w, A\Sigma_w A^T)$. 3. Classify a test image x by evaluating its likelihood under each Gaussian, and picking the most likely Gaussian: $p(w|x) \propto \mathcal{N}(x|A\mu_w, A\Sigma_w A^T)$.

2.5 Contextual Query

To illustrate our approach, we also experiment with a new variant of the ZSL setting. In conventional ZSL, a novel word can be matched against images by projecting it into image space, and sorting images by their distance to the word (vector), or likelihood under the word (Gaussian). However, results may be unreliable when used with polysemous words, or words with large appearance variability. In this case we may wish to enrich the query with contextual words that disambiguate the visual meaning of the query. With regular vector-based queries, the typical approach is to sum the word-vectors. For example: For contextual disambiguation of polysemy, we may hope that vec('bank')+vec('river') may retrieve a very different set of images than vec('bank')+vec('finance'). For specification of a specific subcategory or variant, we may hope that vec('plane')+vec('military') retrieves a different set of images than vec('plane')+vec('passenger'). By using distributions rather than vectors, our framework provides a richer means to make such queries that accounts for the intra-class variability of each concept. When each word is represented by a Gaussian, a two-word query can be represented by their product, which is the new Gaussian $\mathcal{N}(\frac{\Sigma_1^{-1}\mu_1+\Sigma_2^{-1}\mu_2}{\Sigma_1^{-1}+\Sigma_2^{-1}}, (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1})$.

3 Experiments

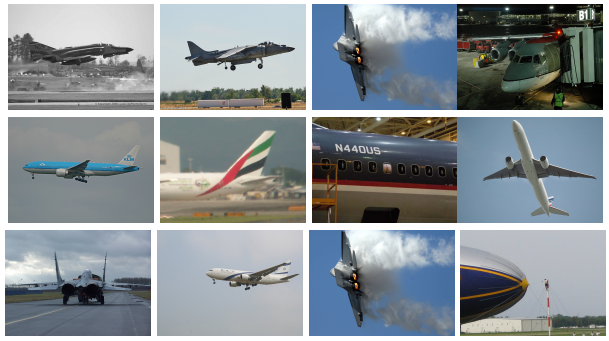
3.1 Datasets and Settings

Datasets: We evaluate our method [1] using the main Animals with Attributes (**AWA**) and **ImageNet1K** benchmarks. To extract visual features we use the VGG-16 CNN (Simonyan and Zisserman, 2015) to extract a $d_x = 4096$ dimensional feature for each image. To train the word Gaussian representation, we use a combination of UkWAC (Ferraresi et al., 2008) and Wikipedia corpus of 25 million tokens, and learn a $d_w = 100$ dimensional Gaussian representation. We set our margin parameter to $\Delta = 1$.

Settings: Our zero-shot setting involves training a visual recogniser (i.e., our mapping A) on a subset of classes, and evaluating it on a disjoint subset. For AWA, we use the standard 40/10 class split (Lampert et al., 2009), and for ImageNet we use a standard 800/200 class split (Mensink et al., 2012).

Competitors: We implement a set of representative alternatives for direct comparison with ours on the same visual features and text corpus. These include: cross-modal linear regression (LinReg, (Dinu et al., 2015)), non-linear regression (NLinReg, (Lazaridou et al., 2014; Socher et al., 2013)),

[1] Code and datasets kept at `http://bit.ly/2cI64Zf`



(a) Top: 'Military'+'Plane' (Gaussian), Middle: 'Passenger'+'Plane' (Gaussian), Bottom: 'Passenger'+'Plane' (Vector)

(b) Top: 'White'+'Horse' (Gaussian), Middle: 'Black'+'Horse' (Gaussian), Bottom: 'Black'+'Horse' (Vector)

Figure 1: Qualitative visualisation of zero-shot query with context words.

	Vector space models				Ours
Dataset	LinReg	NLinReg	CME	ES-ZSL	Gaussian
AWA	44.0	48.4	43.1	58.2	65.4

Table 1: Zero-shot recognition results on AWA (% accuracy).

ImageNet	
ConSE (Norouzi et al., 2014)	28.5%
DeVISE (Frome et al., 2013)	31.8%
Large Scale Metric. (Mensink et al., 2012)	35.7%
Semantic Manifold. (Fu et al., 2015)	41.0%
Gaussian Embedding	45.7%
AwA	
DAP (CNN feat) (Lampert et al., 2009)	53.2%
ALE (Akata et al., 2013)	43.5%
TMV-BLP (Fu et al., 2014)	47.1%
ES-ZSL (Romera-Paredes and Torr, 2015)	49.3%
Gaussian Embedding	65.4%

Table 2: Comparison of our ZSL results with state of the art.

ES-ZSL (Romera-Paredes and Torr, 2015), and a max-margin cross-modal energy function method (CME, (Akata et al., 2013; Frome et al., 2013)). Note that the CME strategy is the most closely related to ours in that it also trains a $d_x \times d_w$ matrix with max-margin loss, but uses it in a bilinear energy function with vectors $E(x, y) = x^T Ay$; while our energy function operates on Gaussians.

3.2 Results

Table 1 compares our results on the AWA benchmark against alternatives using the same visual features, and word vectors trained on the same corpus. We observe that: (i) Our Gaussian-embedding obtains the best performance overall. (ii) Our method outperforms CME which shares an objective function and optimisation strategy with ours, but operates on vectors rather than Gaussians. This suggests that our new distribution rather than vector-embedding does indeed bring significant benefit.

A comparison to published results obtained by other studies on the same ZSL splits is given in Table 2, where we see that our results are competitive despite exploitation of supervised embeddings such as attributes (Fu et al., 2014), or combinations of embeddings (Akata et al., 2013) by other methods.

We next demonstrate our approach qualitatively by means of the contextual query idea introduced in Sec 2.5. Fig. 1 shows examples of how the top retrieved images differ intuitively when querying ImageNet for zero-shot categories 'plane' and 'horse' with different context words. To ease interpretation, we constrain the retrieval to the true target class, and focus on the effect of the context word. Our learned Gaussian method retrieves more relevant images than the word-vector sum baseline. E.g., with the Gaussian model all of the top-4 retrieved images for Passenger+Plane are relevant, while only two are relevant with the vector model. Similarly, the retrieved black horses are more clearly black.

3.3 Further Analysis

To provide insight into our contribution, we repeat the analysis of the AwA dataset and evaluate several variants of our full method. These use our features, and train the same cross-domain max-margin loss in Eq 5, but vary in the energy function and representa-

AwA	
Bilinear-WordVec	43.1%
Bilinear-MeanVec	52.2%
PPK-MeanVec	52.6%
PPK-Gaussian	65.4%

Table 3: Impact of training and testing with distribution rather than vector-based representations

tions used. Variants include: (i) Bilinear-WordVec: Max-margin training on word vector representations of words and images with a bilinear energy function. (ii) Bilinear-MeanVec: As before, but using our Gaussian means as vector representations in image and text domains. (iii) PPK-MeanVec: Train the max-margin model with Gaussian representation and PPK energy function as in our full model, but treat the resulting means as point estimates for conventional vector-based ZSL matching at testing-time. (v) PPK-Gaussian: Our full model with Gaussian PPK training and testing by Gaussian matching.

From the results in Table 3, we make the observations: (i) Bilinear-MeanVec outperforming Bilinear-WordVec shows that cross-modal (Sec 2.3) training of word Gaussians learns better point estimates of words than conventional word-vector training, since these only differ in the choice of vector representation of class names. (ii) PPK-Gaussian outperforming PPK-MeanVec shows that having a model of intra-class variability (as provided by the word-Gaussians) allows better zero-shot recognition, since these differ only in whether covariance is used at testing time.

3.4 Related Work and Discussion

Our approach models intra-class variability in both images and text. For example, the variability in visual appearance of military versus passenger 'plane's, and the variability in context according to whether a the word 'plane' is being used in a military or civilian sense. Given distribution-based representations in each domain, we find a cross-modal map that warps the two distributions into alignment.

Concurrently with our work, Ren et al (2016) present a related study on distribution-based visual-text embeddings. Methodologically, they benefit from end-to-end learning of deep features as well as cross-modal mapping, but they only discriminatively train word covariances, rather than jointly training both means and covariances as we do.

With regards to efficiency, our model is fast to train if fixing pre-trained word-Gaussians and optimising only the cross-modal mapping A. However, training the mapping jointly with the word-Gaussians comes at the cost of updating the representations of all words in the dictionary, and is thus much slower.

In terms of future work, an immediate improvement would be to generalise our of Gaussian embeddings to model concepts as mixtures of Gaussians or other exponential family distributions (Rudolph et al., 2016; Chen et al., 2015). This would for example, allow polysemy to be represented more cleanly as a mixture, rather than as a wide-covariance Gaussian as happens now. We would also like to explore distribution-based embeddings of sentences/paragraphs for class description (rather than class name) based zero-shot recognition (Reed et al., 2016). Finally, besides end-to-end deep learning of visual features, training non-linear cross-modal mappings is also of interest.

4 Conclusion

In this paper, we advocate using distribution-based embeddings of text and images when bridging the gap between vision and text modalities. This is in contrast to the common practice of point vector-based embeddings. Our distribution-based approach provides a representation of intra-class variability that improves zero-shot recognition, allows more meaningful retrieval by multiple keywords, and also produces better point-estimates of word vectors.

References

[Akata et al.2013] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. 2013. Label-embedding for attribute-based classification. In *Computer Vision and Pattern Recognition*.

[Arora et al.2015] Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2015. Random walks on context spaces: Towards an explanation of the mysteries of semantic word embeddings. *CoRR*, abs/1502.03520.

[Blei et al.2003] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *JMLR*, 3:993–1022.

[Bruni et al.2012a] Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012a. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, pages 136–145.

[Bruni et al.2012b] Elia Bruni, Jasper Uijlings, Marco Baroni, and Nicu Sebe. 2012b. Distributional semantics with eyes: Using image analysis to improve computational representations of word meaning. In *ACM Multimedia*.

[Bruni et al.2014] Elia Bruni, Nam Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *J. Artif. Int. Res.*, 49(1):1–47, January.

[Chen et al.2015] Xinchi Chen, Xipeng Qiu, Jingxiang Jiang, and Xuanjing Huang. 2015. Gaussian mixture embeddings for multiple word prototypes. *arXiv preprint arXiv:1511.06246*.

[Collobert et al.2011] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537, November.

[Dinu et al.2015] Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. In *ICLR Workshop Paper*.

[Ferraresi et al.2008] Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and evaluating ukwac, a very large web-derived corpus of english. In *In Proceedings of the 4th Web as Corpus Workshop (WAC-4*.

[Frome et al.2013] Andrea Frome, Greg Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, MarcAurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. In *Neural Information Processing Systems (NIPS)*.

[Fu et al.2014] Yanwei Fu, Timothy Hospedales, Tony Xiang, Zhenyong Fu, and Shaogang Gong. 2014. Transductive multi-view embedding for zero-shot recognition and annotation. In *European Conference on Computer Vision*.

[Fu et al.2015] Z. Fu, T. A. Xiang, E. Kodirov, and S. Gong. 2015. Zero-shot object recognition by semantic manifold distance. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2635–2644, June.

[Harris1954] Zellig Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

[Jebara et al.2004] T. Jebara, R. Kondor, and A. Howard. 2004. Probability product kernels. *Journal of Machine Learning Research*, 5:819–844.

[Kiela and Bottou2014] Douwe Kiela and Léon Bottou. 2014. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-14)*.

[Lampert et al.2009] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *Computer Vision and Pattern Recognition*.

[Lazaridou et al.2014] 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 the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, June.

[Mensink et al.2012] Thomas Mensink, Jakob Verbeek, Florent Perronnin, and Gabriela Csurka. 2012. Metric learning for large scale image classification: Generalizing to new classes at near-zero cost. In *European Conference on Computer Vision*.

[Mikolov et al.2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

[Norouzi et al.2014] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg Corrado, and Jeffrey Dean. 2014. Zero-shot learning by convex combination of semantic embeddings. In *ICLR*.

[Reed et al.2016] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. 2016. Learning deep representations of fine-grained visual descriptions. In *IEEE Computer Vision and Pattern Recognition (CVPR)*.

[Ren et al.2016] Zhou Ren, Hailin Jin, Zhe Lin, Chen Fang, and Alan Yuille. 2016. Joint image-text representation by gaussian visual semantic embedding. In *Proceeding of ACM International Conference on Multimedia (ACM MM)*.

[Romera-Paredes and Torr2015] Bernardino Romera-Paredes and Philip H. S. Torr. 2015. An embarrassingly simple approach to zero-shot learning. In *ICML*.

[Rudolph et al.2016] Maja R. Rudolph, Francisco J. R. Ruiz, Stephan Mandt, and David M. Blei. 2016. Exponential Family Embeddings, August.

[Schwenk2007] Holger Schwenk. 2007. Continuous space language models. *Computer Speech and Language*, 21.

[Silberer and Lapata2014] Carina Silberer and Mirella Lapata. 2014. Learning grounded meaning representations with autoencoders. In *ACL*.

[Simonyan and Zisserman2015] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.

[Socher et al.2013] Richard Socher, Milind Ganjoo, Christopher D. Manning, and Andrew Y. Ng. 2013. Zero Shot Learning Through Cross-Modal Transfer. In *Advances in Neural Information Processing Systems 26*.

[Vilnis and McCallum2015] Luke Vilnis and Andrew McCallum. 2015. Word representations via gaussian embedding. In *ICLR*.