PKWE: Prior-Knowledge Word Embeddings

Yanbo Fang
Department of Computer Science
Rutgers University
yanbo.fang@rutgers.edu

Abstract

We will introduce a multimodal embedding through learning prior knowledge from a casual graph of multimodal representation. Our experiment shows that our multimodal embeddings can be used to improve model performance in other multimodal tasks, such as visual semantic embeddings (VSE). Moreover, the result is competitive with other popular pretrained embeddings.

1 Introduction

Building artificial models containing and having the ability to infer on the real-world knowledge is always the focus of researchers. At first, scientists wrote knowledge facts and stored them in a knowledge graph. However, the knowledge is vast and hard to enumerate them. Recently, Deep Neural Networks (DNNs) (Schmidhuber, 2015; Bengio et al., 2013) emerges and exhibits a significant advance than former methods. DNNs can learn a generalization model from big enough datasets and process other unseen data through learning meaningful representation. But the drawbacks of DNNs are these models can not infer the knowledge they have to other unseen data, even though the unseen data are close to other seen data.

In 2013, (Mikolov et al., 2013) created Word2Vec, a classical word embeddings, which can be used to build correlations between similar or frequent co-occurrence words. This method paved a new way for NLP research, as the embeddings can build relations among different words. However, some abstract knowledge can not be learned only by text data. For example, one agent, like the human, will not have the comprehensive understanding of ”red” if only can learn from texts, even though the agent can know the usage of this word and the similarity with other color words. Nevertheless, when the agent sees a red object, the meaning of ”red” shall be understood immediately. Therefore, some research works about multimodal embeddings were proposed to address this issue. (Calixto and Liu, 2017; Collell et al., 2017; Lazaridou et al., 2015).

However, these multimodal embeddings researches focus more on finding shared joint embeddings for different modalities. An important assumption behind these methods is that different modalities have the same status. However, we assume that different modalities have different priorities. Modalities with higher priority can decide the representation of lower priority modalities. For example, our understanding of the word ”apple” is a general impression of apples in the real world. In simple, Visual perceptions bring prior knowledge into our language. Also, it is worth to mention that the prior knowledge is different from common sense knowledge (Bisk et al., 2020; Bosselut et al., 2019; Bosselut and Choi, 2019; Liu and Singh, 2004). Because common sense knowledge can be learned from experience, prior knowledge exists ahead of experience and is the basis of common sense. Here in our research, the meaning of prior knowledge is a kind of definite knowledge before training. More concretely, in our scenes, the prior knowledge is the relations between one object and its name; for example, we call apple the name ”apple,” and inference on prior knowledge can lead to common sense knowledge.

Different from past methods, we build a multimodal graph according to head-to-head casual graph structure. In our method, we set visual perceptions as the cause, and word embeddings as the collider (Pearl et al., 2016), the link among these two items is a casual relationship. Therefore, word embeddings can infer the prior knowledge from visual perceptions through their connections. Then we apply the `CBOW` algorithm on the multi-modal graph for distinction and interrelations among different words. Because prior knowledge is the cornerstone of common sense knowledge, word embeddings are the cornerstone of language models. Therefore, we call our embedding Prior-Knowledge Word Embedding (PKWE). We prove the effectiveness of prior knowledge through experiments by applying the embeddings on downstream tasks, visual semantic embeddings (VSE). Our results show that our combined model has a 5% improvement compared with random initialized models and has a similar performance with other pre-trained embeddings and outperforms baseline models. We also tested the model's performance of incorporating tree structure into the graph.

2 Related Work

2.1 Word Embeddingss

A large part of the success of modern language models can be attributed to the pre-trained word embeddings. (Bengio et al., 2003) originally coined the term word embeddings. Then, it was (Mikolov et al., 2013) to promote the term word embeddings by creating Word2Vec, by predicting a word's representation through its context. Later, another famous word embeddings Glove proposed by (Pennington et al., 2014), which is based on word occurrences in a textual corpus. Although these embeddings are formed through contextualized learning on large-scale language corpora, only learning from text imposes bias inevitably. Because some abstract words can not be described by text; thus, other modal information is required.

2.2 MultiModal Embeddings

Some recent research works focus on Multimodal embeddings. (Bruni et al., 2014) applied of single value decomposition (SVD) to the matrix of concatenated visual and textual representation. The auto-encoder structure was adopted in the research of (Silberer and Lapata, 2014). Encoders are fed with pre-learned visual and text features, and the hidden representations are then used as multimodal embeddings. (Lazaridou et al., 2015) proposed the multimodal skip-gram (MMSG) model extends the original skip-gram model (Mikolov et al., 2013) through incorporating visual features. (Ailem et al., 2018) proposes a Probabilistic Model for text and images embedding learning. All of these methods mentioned above are considering fusing visual features into textual features. However, our method is motivating the word embeddings to infer a general representation from visual features.

2.3 Visual Semantic Embeddings

Visual semantic embedding (Frome et al., 2013) is a technique for learning a joint representation among two different modalities, vision and language, through mapping into a common embedding space. In some research works, the embedding space was applied to a set of cross-modal tasks such as image captioning (Vinyals et al., 2015; Donahue et al., 2015), and visual question answering (Agrawal et al., 2015). (Frome et al., 2013) proposed a method for using textual data to learn semantic embedding and visual data to learn visual embedding than mapping pair-wised embeddings into the joint embedding space. VSE++ (Faghri et al., 2018) uses the online hard negative mining (OHEM) strategy for data comparison on the base structure of (Frome et al., 2013) and shows the performance gain. In this paper, we choose the image-text retrieval task as our downstream task to test our embedding, and we incorporate our embedding into VSE++ (Faghri et al., 2018) as our tested model.

3 Method

3.1 Prior-Knowledge Word Embeddings

The process of embeddings generation has two steps, the first is to build a image-text graph, the direction is always from image to word, the second step is creating optimized distributed word embedding through `CBOW` algorithm.

We use pre-trained Resnet to extract image features v_i for image i, and set v_i as ground truth, for corresponding sentence $\{w_{i,1}, w_{i,2}, \ldots, w_{i,j}\}$, creating a directional edge from image v_i to each words $w_{i,1:j}$. Fig.1 is an example of our built model, in this figure red nodes denotes images features, $v^{(i)}$ is the image representation of the sentence "`the cat is on the ground`", $v^{(l)}$ is the image representation of the sentence "`the dog is on the ground`", and each blue node in the graph is one unique word, and each edge points from image node to its paired word node, for example, some nodes `a`, `on`, and `ground` are directed from both $v^{(i)}$ and $v^{(l)}$, and some word only appears with one image, thus these words nodes are dominated by their unique parent node, like $v^{(i)}$ points to `cat` and $v^{(l)}$ points to `dog`.

After graph completion, keeping the image feature v_i static, iterate the following steps for each

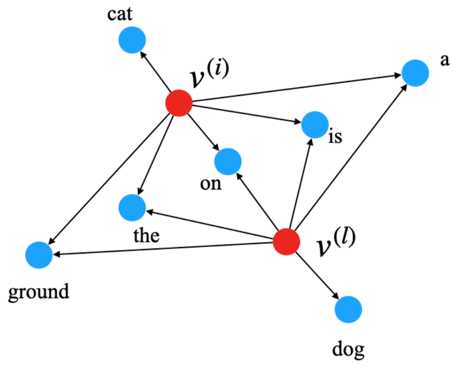

Figure 1: Example of the multi-modal graph

word nodes i until every node stable,

$$fea(w_i) = \frac{\sum_{k \in \mathrm{NER}(w_i)} v_k}{\mathrm{NUM}(\mathrm{NER}(w_i))}$$

Here, $fea(w_i)$ means the feature of node w_i, $\mathrm{NER}(w_i)$ means the parent node of w_i, $\mathrm{NUM}(\mathrm{NER}(w_i))$ represents the number nodes that points to node w_i. The purpose of the above propose is to import prior knowledge into each word node by getting close to their deciding modality.

Then we use `CBOW` algorithm, created by (Mikolov et al., 2013), to bring distinction and relations among words. `CBOW` is one of unsupervised learning algorithm which can be used to bring connection into similar words. In the `CBOW` model, the representations of parent nodes are combined to predict the word as their children. The detail is for node w_i having m parents $v_1, v_2, \ldots, v_m$, and the average vector is

$$\hat{v} = \frac{v_1 + v_2 + \cdots + v_m}{m}$$

and getting a score function by computing the similarity $z = (fea(w_i)^{\mathrm{T}}\hat{v})$, and formulate the optimization objective as the negative log probability by comparing with all other nodes in the graph

$$\begin{aligned} J &= -\log(\mathrm{softmax}(\hat{z})) \\ &= -\log \frac{\exp\left(fea(w_i)^{\mathrm{T}}\hat{v}\right)}{\sum_{j \in |V|} \exp\left(fea(w_i)^{\mathrm{T}} v_j\right)} \end{aligned}$$

In order to improve the efficiency, we adopt Negative Sampling, replacing selecting all words as randomly choosing K words, the random selected vector is represented by $\tilde{v}$. Thus the loss functions becomes

$$\begin{aligned} J &= -\log(\mathrm{softmax}(\hat{z})) \\ &= -\log \frac{\exp\left(fea(w_i)^{\mathrm{T}}\hat{v}\right)}{\sum_{k \in K} \exp\left(fea(w_i)^{\mathrm{T}} \tilde{v}_k\right)} \end{aligned}$$

3.2 Visual Semantic Embeddings

We select cross-modal tasks as the evaluation of our embdding. Specifically, we repalce the randomly initialzed embeddings in model VSE++ as PKWE. It can jointly learn the common embedding spaces of two modalities: vision and language, and aligns them using parallel image-text pairs and 25K sentences.

Let $\mathrm{v} \in \mathbf{R}^d$ be the representation of images by an image encoder and $\mathrm{u} \in \mathbf{R}$ as the representation of pair sentences through a language encoder, and we define our hard negative sampling strategy following the instruction of []. Given pairs of image-sentence pairs v and u, the hardest negatives are $\mathrm{v}' = \arg\max_{j \neq i} s(\mathrm{v}_j, \mathrm{u})$ and $\mathrm{u}' = \arg\max_{j \neq i} s(\mathrm{v}, \mathrm{u}_j)$, here $s(\cdot,\cdot)$ is similarity measurement, and the Max-Hinge loss is defined as

$$\begin{aligned} l(\mathrm{v}, \mathrm{u}) = \sum_{\mathrm{v}} \max_{\mathrm{v}'}(\alpha + |s(\mathrm{v}_j, \mathrm{u}) - (\mathrm{v}, \mathrm{u})|_+) \\ + \sum_{\mathrm{u}} \max_{\mathrm{u}'}(\alpha + |s(\mathrm{v}, \mathrm{u}_j) - (\mathrm{v}, \mathrm{u})|_+) \end{aligned}$$

, where α serves as a margin parameter, and $|\cdot|_+ = \max(0, \cdot)$ is the traditional ranking loss,

4 Experiments

The dataset we used to produce our word-embedding and evaluate our combined model is MS-COCO datasets, which contains 11536 images for training, 1K images for evaluation and test, every image has 5 paired sentences. We also tested our model on a more challenging 5K Images.

4.1 Implementation Detail

4.1.1 Embedding Generation

We used DGL open source library (Wang et al., 2019) and all training data to build graph, and adopt ResNet as the image encoder. These extracted image features are 4096 dimensions, then using SVD algorithm to reduce the image dimension to word embedding space with 300 dimensions. Models are trained for at most 50 epochs or the changes of the whole graph smaller than 0.001, the learning rate is a constant number 1. The number of neagtive sample is 10.

Task	Image to Text				Text to Image				
Metric	R@1	R@5	R@10	Med.r	R@1	R@5	R@10	Med.r	rsum
1K testing split (5000 captions)									
VSE++	63.5	89.4	96.2	1	47.3	80.6	89.5	2	466.4
Glove+VSE++	64.8	90.6	96.3	1	**49.7**	82.1	90.9	2	474.4
Fasttext+VSE++	64.3	**91.5**	**96.7**	1	48.7	81.0	90.4	2	472.8
Word2Vec+VSE++	**65.5**	90.6	96.1	1	49.1	**82.8**	**91.0**	2	**475.1**
GWE+VSE++(Ours)	64.0	91.2	96.6	1	49.5	82.2	90.8	2	474.3
5K testing split (25000 captions)									
VSE++	35.2	65.4	77.6	3	23.8	51.7	65.2	5	318.8
Glove+VSE++	**37.4**	66.8	79.2	**2**	25.1	54.4	67.5	5	330.4
Fasttext+VSE++	35.8	66.7	78.8	3	25.0	53.0	66.7	5	326.0
Word2Vec+VSE++	37.3	67.9	79.2	**2**	**25.5**	54.0	67.6	5	331.6
GWE+VSE++(Ours)	36.2	**68.7**	**80.0**	3	25.2	**54.5**	**67.8**	**4**	**332.5**

Table 1: Results of cross-modal retrieval task on MS-COCO dataset among our word embeddings with other pre-trained word embeddings and randomly initialized embeddings. And one thing worth to mention is the gap of our experiment of VSE++ and in its original paper probability due to we used pre-extracted image features by ResNet for computing efficiency

4.1.2 Visual Semantic Embedding

We followed the Details of implementation in VSE++ (Faghri et al., 2018), setting ResNet (He et al., 2016) as image encoder, GRU (Cho et al., 2014) as sentence encoder, using Adam optimizer, at most 30 epochs training, especially with lr 0.0002 for 15 epochs, and then 0.00002 lr for the next 15 epochs. Batch size is 128. The test checkpoint is selected based on the performance on the validation set. Word embeddings were fine-tuned during training. And for efficiency, we used precomputed image features by ResNet in our experiments

4.2 Evaluation of Cross-Modal Retrieval

We select the R@1(recall), R@5, R@10, and the median retrieval rank as our evaluation metric. Also we compute `rsum` as the summation of R@1, R@5 and R@10 as the overall standard. The baseline model is (Wang et al., 2016; Eisenschtat and Wolf, 2017; Karpathy and Li, 2015; Niu et al., 2017; Vendrov et al., 2015; Faghri et al., 2018)

The result is shown in Tab.1, we can see the our combined VSE++ significally outperforms the original VSE++ no matter in 1K or 5K test, approximately 21.6 `rsum` points improvements, and has a close performance to other VSE++ models combined with widely used pre-trained word embeddings like Word2Vec (Mikolov et al., 2013), Glove (Pennington et al., 2014), and Fasttext (Bojanowski et al., 2017; Joulin et al., 2017). From Tab.2, we can see our combined model is better than baseline models. This validates the effectiveness of learning prior knowledge from images. Especially considering the data size we used to produce our word embeddings, only 110k images and 550K sentences and 26383 unique tokens, while other popular pre-trained embeddings were trained on Billion size sentences.

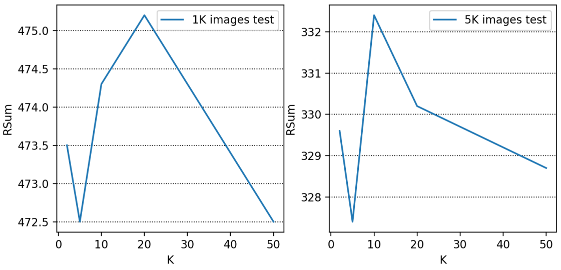

Figure 2: The performance on cross-modal retrieval with different Negative Samples numbers K, The **left** is the result on 1000 images test dataset, the **right** is the result on 5000 images test dataset.

4.3 Choice of the Number of Negative Samples

We study the effect of number K on the performance of prior-knowledge embeddings on retrieval tasks. Fig.2 shows the `rsum` performance on the image-text retrieval task corresponding to different choices of K. As show in Fig.2, choosing K from the range $[10, 20]$ has the best performance both on 1K and 5K test.

Task	Image to Text				Text to Image				
Metric	R@1	R@5	R@10	Med.r	R@1	R@5	R@10	Med.r	rsum
1K testing split (5000 captions)									
DVSA	38.4	69.9	80.5	1	27.4	60.2	74.8	3	351.2
HM-LSTM	43.9	-	87.8	2	36.1	-	86.7	3	
Order-Embedding	46.7	-	88.9	2	37.9	-	85.9	2	-
DeepSP	50.1	79.7	89.2	-	39.6	75.2	86.9	-	420.7
2WayNet	55.8	75.2	-	-	39.7	63.3	-	-	-
VSE++	63.5	89.4	96.2	1	47.3	80.6	89.5	2	466.4
GWE+VSE++(Ours)	**64.0**	**91.2**	**96.6**	1	**49.5**	**82.2**	**90.8**	2	**474.3**
5K testing split (25000 captions)									
Order-embedding	23.3	-	65.0	5	18.0	-	57.6	7	-
VSE++	35.2	65.4	77.6	3	23.8	51.7	65.2	5	318.8
GWE+VSE++(Ours)	**36.2**	**68.7**	**80.0**	3	**25.2**	**54.5**	**67.8**	**4**	**332.5**

Table 2: Results of cross-modal retrieval task on MS-COCO dataset among different models. For fairness, we didn't include models trained on less data, like (Shi et al., 2018) and (Wu et al., 2019)

4.4 Tree Structure Graph

In addition to only pointing from the image node to the word node, we tried to introduce tree hierarchies among nodes by applying dependency grammar without a drastic increase in the graph's size. We used Spacy [1] for the production of the dependency tree and pointing from the corresponding image node to the root. The result is shown in Tab.3. Although the result does not have improvement, we guess because of the imprecise tree structures, and we believe it can make progress if there are improvements in the precision of dependency tree extraction.

5 Conclusions

We present a causal graph approach for creating multi-modal word embeddings by setting images as the ground truth, then motivating words to learn and infer prior knowledge from images. The downstream task experiment result shows that our combined model (PKWE+VSE++) has an advancement over the original model. Through incorporating PKWE, the VSE++ model has a similar effect with adopting other popular pre-trained word embeddings. And our combined VSE++ model outperforms baseline models.

References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, M. Mitchell, C. L. Zitnick, D. Parikh, and Dhruv Batra. 2015. Vqa: Visual question answering. *International Journal of Computer Vision*, 123:4–31.

Melissa Ailem, Bowen Zhang, Aurélien Bellet, P. Denis, and F. Sha. 2018. A probabilistic model for joint learning of word embeddings from texts and images. In *EMNLP*.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828. Cite arxiv:1206.5538.

Yoshua Bengio, R. Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. *ArXiv*, abs/1911.11641.

P. Bojanowski, E. Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Antoine Bosselut and Yejin Choi. 2019. Dynamic knowledge graph construction for zero-shot commonsense question answering. *ArXiv*, abs/1911.03876.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, A. Çelikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *ACL*.

Elia Bruni, Nam-Khanh Tran, and M. Baroni. 2014. Multimodal distributional semantics. *J. Artif. Intell. Res.*, 49:1–47.

[1] https://spacy.io/

Task	Image to Text				Text to Image				
Metric	R@1	R@5	R@10	Med.r	R@1	R@5	R@10	Med.r	rsum
1K testing split (5000 captions)									
GWE+VSE++(Tree)	63.7	91.6	96.8	1	49.0	81.9	90.3	2	473.3
5K testing split (25000 captions)									
GWE+VSE++(Tree)	35.9	67.7	79.0	3	25.0	53.6	67.1	5	328.3

Table 3: Results of cross-modal retrieval task on MS-COCO dataset for PKWE from tree-structure graph

Iacer Calixto and Qun Liu. 2017. Incorporating global visual features into attention-based neural machine translation. In *EMNLP*.

Kyunghyun Cho, B. V. Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*.

Guillem Collell, Ted Zhang, and Marie-Francine Moens. 2017. Imagined visual representations as multimodal embeddings. In *AAAI*.

J. Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, S. Guadarrama, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2625–2634.

Aviv Eisenschtat and Lior Wolf. 2017. Linking image and text with 2-way nets. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1855–1865.

Fartash Faghri, David J. Fleet, J. Kiros, and S. Fidler. 2018. Vse++: Improving visual-semantic embeddings with hard negatives. In *BMVC*.

Andrea Frome, G. S. Corrado, Jonathon Shlens, S. Bengio, J. Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. In *NIPS*.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Armand Joulin, E. Grave, P. Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. *ArXiv*, abs/1607.01759.

Andrej Karpathy and Fei-Fei Li. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137. IEEE Computer Society.

A. Lazaridou, N. Pham, and M. Baroni. 2015. Combining language and vision with a multimodal skip-gram model. In *HLT-NAACL*.

H. Liu and Push Singh. 2004. Conceptnet — a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22:211–226.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. 2017. Hierarchical multimodal lstm for dense visual-semantic embedding. In *ICCV*, pages 1899–1907. IEEE Computer Society.

J. Pearl, Madelyn Glymour, and N. Jewell. 2016. Causal inference in statistics: A primer.

Jeffrey Pennington, R. Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*.

J. Schmidhuber. 2015. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117. Published online 2014; based on TR arXiv:1404.7828 [cs.NE].

Haoyue Shi, Jiayuan Mao, Tete Xiao, Yuning Jiang, and Jian Sun. 2018. Learning visually-grounded semantics from contrastive adversarial samples. *ArXiv*, abs/1806.10348.

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

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2015. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*.

Oriol Vinyals, A. Toshev, S. Bengio, and D. Erhan. 2015. Show and tell: A neural image caption generator. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164.

Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5005–5013.

Minjie Wang, Da Zheng, Zihao Ye, Quan Gan, Mufei Li, Xiang Song, Jinjing Zhou, Chao Ma, Lingfan Yu, Yu Gai, Tianjun Xiao, Tong He, George Karypis, Jinyang Li, and Zheng Zhang. 2019. Deep graph library: A graph-centric, highly-performant package for graph neural networks. *arXiv preprint arXiv:1909.01315*.

Hao Wu, Jiayuan Mao, Y. Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and W. Ma. 2019. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6602–6611.