MM-GATBT: Enriching Multimodal Representation Using Graph Attention Network

Seung Byum Seo, Hyoungwook Nam, Payam Delgosha
University of Illinois at Urbana-Champaign
{sbseo2, hn5, delgosha}@illinois.edu

Abstract

While there have been advances in Natural Language Processing (NLP), their success is mainly gained by applying a self-attention mechanism into single or multi-modalities. While this approach has brought significant improvements in multiple downstream tasks, it fails to capture the interaction between different entities. Therefore, we propose MM-GATBT, a multimodal graph representation learning model that captures not only the relational semantics within one modality but also the interactions between different modalities. Specifically, the proposed method constructs image-based node embedding which contains relational semantics of entities. Our empirical results show that MM-GATBT achieves state-of-the-art results among all published papers on the MM-IMDb dataset.

1 Introduction

Despite the huge success of learning algorithms for applications involving unimodal data such as text, less is known for applications involving multimodal data, i.e. scenarios where each data entity has data attributes from multiple modes, such as text and image. While the previous works show that models with multimodal representation outperforms unimodal representation in downstream tasks such as classification, VQA, and disambiguation, the benefit of multimodal representation mostly comes from only one mode (such as text), while the other mode only contribute a marginal improvement. That is, the performance difference between text-only representation and multimodal representation is smaller than that of the image-only representation and multimodal representation (Arevalo et al., 2017; Vielzeuf et al., 2018; Moon et al., 2018; Kiela et al., 2020; Singh et al., 2020; Kiela et al., 2021).

We suspect that improper usage of image-modality presents a limitation in creating multimodal representation. Existing multimodal models have been applying a self-attention mechanism or create a graph with a single modality's attribute. However, these approaches ignore the interaction among entities, multi-modalities, or both. In other words, one modality is tied within its space and cannot see beyond its modality space. To overcome this limitation, we propose a novel framework by constructing a multimodal entity graph which simultaneously captures the interconnection between different data entries and data modalities. Our idea is motivated by *homophily*, in which similar nodes tend to be connected and tend to share similar labels (Hamilton, 2020).

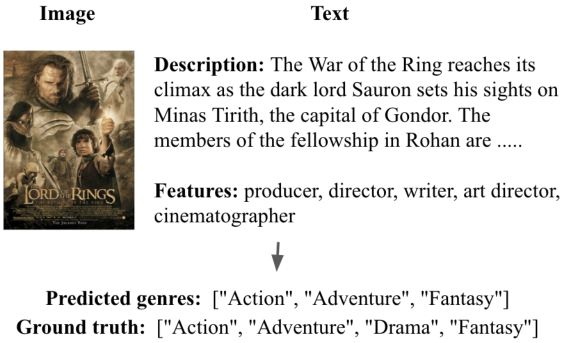
Image
Text
Description: The War of the Ring reaches its climax as the dark lord Sauron sets his sights on Minas Tirith, the capital of Gondor. The members of the fellowship in Rohan are .....
Features: producer, director, writer, art director, cinematographer
Predicted genres: ["Action", "Adventure", "Fantasy"]
Ground truth: ["Action", "Adventure", "Drama", "Fantasy"]

Figure 1: Given movie poster and text information, the problem is to predict the multilabel genres of movies. Our method narrows down this problem into a node classification task by constructing a multimodal entity graph where each node represents a movie entity and edge represents a shared feature between the movie entities.

We demonstrate our claim by considering a multilabel classification task using the MM-IMDb dataset (Arevalo et al., 2017) as in Figure 1. In the MM-IMDb dataset, each movie entity is provided with image and text, and our goal is to predict the movie's genre. Using this data, we construct a graph where each node represents a movie, and is given the movie image as an attribute. Furthermore, we connect two nodes if the corresponding movies share features, i.e. if they have the same producer,

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop, pages 106 - 112
July 10-15, 2022 ©2022 Association for Computational Linguistics

director, etc. By capturing dependency and interaction between the entities generated from Graph Attention Network (GAT) (Veličković et al., 2018), we expect to gain latent information that cannot be extracted from the image encoder solely.

The contributions of this work are as follows: (1) We propose a novel Multimodal Graph Attention Network (MM-GATBT) which enables interaction between data modalities. (2) To our best knowledge, this is the first attempt to construct image-based entity graph to enrich image representation by capturing relational semantics between the entities. (3) MM-GATBT achieves state-of-the-art results on the multilabel classification task among all published papers on MM-IMDb dataset.

2 Background

Multimodal Representation Joint representation is one of the most popular methods to combine modality vectors. This method has a strong advantage in implementation because it concatenates the modalities into a single vector. (Guo et al., 2019) explains that it is an intuitive approach to learn a shared semantic subspace from different modalities providing richer and complementary contexts.

(Bayoudh et al., 2021) also explains three different fusion methods depending on the timing when modalities are combined. Early fusion (Sun et al., 2018) method fuses data before the feature extractor or classifier to preserve the richness of original features. The late fusion method fuses data after extracting features from separate modalities. Hybrid method uses both early fusion and late fusion at some point in their architecture to take advantage of both worlds.

Graph Neural Network Graph Neural Network (GNN) is powered by neural message passing and generates node embeddings. A graph $G = (V, E)$ is defined as a tuple such that V is a set of vertices and $E \subseteq V \times V$ is a set of edges. We also employ the node feature matrix $X \in \mathbb{R}^{d\times|V|}$ where d is the feature dimension. Vanilla GNN (Kipf and Welling, 2017) averages neighbor messages for each layer using the mean aggregation function. Formally, it is defined by the following Eq. (1) where l is the layer index, h_i^l is hidden representation of node i at layer l, and U^l is a learnable parameter.

$$h_i^{l+1} = \sigma\left(\sum_{j\in\mathcal{N}_i} \frac{1}{\text{Deg}_i} U^l h_j^l\right). \qquad (1)$$

Here, Deg_i and $\mathcal{N}_i$ denote the degree and the neighbor set of node i, respectively, and $\sigma(.)$ is a non-linear activation function.

Graph Convolution Network (GCN) (Kipf and Welling, 2017) improves vanilla GNN by employing symmetric normalization (Hamilton, 2020). This model runs a spectral-based convolution operation. Because the spectral method assumes fixed graph, it often leads to poor generalization ability (Wu et al., 2021). Therefore, spatial-based models such as GraphSAGE (Hamilton et al., 2017) are often considered to enable inductive generalization.

$$h_i^{l+1} = \sigma(U^l \cdot [h_i^{l-1}; h_j^{l-1}]) \qquad (2)$$

In Eq. (2), $[h_i^{l-1}; h_j^{l-1}]$ denotes a concatenated representation between the node's previous hidden state h_i^{l-1} and an aggregated representation of local neighbor nodes h_j^{l-1} where $j \in \mathcal{N}_i$.

Attention Mechanism Attention mechanism (Luong et al., 2015; Bahdanau et al., 2015) computes a probability distribution $\alpha = (\alpha_{t1}, \alpha_{t2}, ...\alpha_{ts})$ over the encoder's hidden states $h^{(s)}$ that depends on the decoder's current hidden state $h^{(t)}$. (Luong et al., 2015) computes global attention by

$$\alpha_{st} = \frac{exp(h^{(t)} \cdot h^{(s)})}{\sum_{s'} exp(h^{(t)} \cdot h^{(s')})} \qquad (3)$$

where s refers to the index number of source hidden state and t refers to the index number of target hidden state. This method was introduced to assign more importance to more relevant $h^{(s)}$. This method has been developed into self-attention (Vaswani et al., 2017) and GAT (Veličković et al., 2018). Self-attention mechanism computes weighted average of the input vectors. Similarly, GAT performs attention on the neighbor nodes.

3 Methods

3.1 Problem Statement

We address the multilabel classification task. We assume that n data sample are given, where each data sample corresponds to a movie entity that has a text and an image attribute. The goal is to classify the movie genre. Note that this is a multilabel classification task, as each movie can belong to more than one genre. Therefore, given text data $X_{\text{txt}} = \{T^1, T^2, \ldots, T^n\}$ and image data $X_{\text{img}} =$

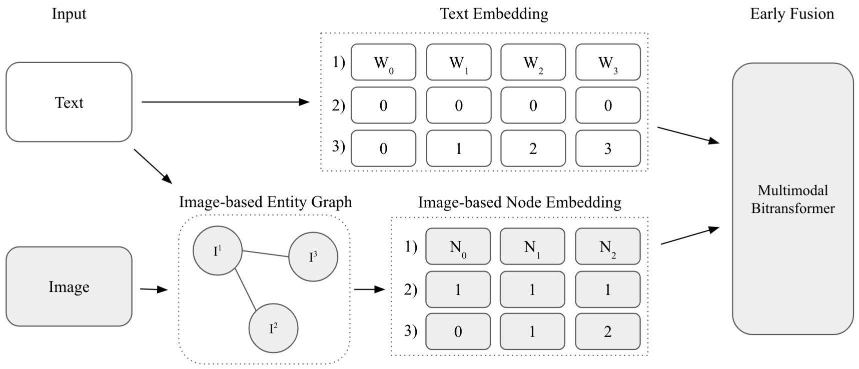

Figure 2: Model architecture of MM-GATBT. The top side of the architecture encodes text descriptions. The bottom side captures the interaction between entities by aggregating the neighbor images connected via text features. Then, MM-GATBT concatenates text embedding and image-based node embedding to generate a joint multimodal representation used for classifier. 1), 2), and 3) denotes token embedding, segment embedding, and positional embedding respectively, following BERT-like tokenization method.

$\{I^1, I^2, \ldots, I^n\}$, we train function f that predicts binary label y^i_j for all j where i is an index number of an entity and j is an index number of classes. Binary label y^i_j is only accessible from training set.

Our approach towards this problem is to construct a graph and use graph neural networks. The details are discussed in Section 3.3 below.

3.2 Model Overview

MM-GATBT consists of three main components: text encoder, image encoder, and GNN. We chose BERT (Devlin et al., 2019) as text encoder, EfficientNet (Tan and Le, 2019) as image encoder, and GAT (Veličković et al., 2018) as GNN. The encoded images are used as node features in GAT to learn the relational semantics of entities. Then we fuse text embedding and image-based node embedding using MMBT (Kiela et al., 2020). We chose this architecture because unlike VilBERT (Lu et al., 2019) and VisualBERT (Li et al., 2019), encoders can be trained independently as opposed to be trained jointly. That is, we can easily upgrade any of these three main components in the future. Thanks to this simple but powerful architecture, MM-GATBT leaves considerable room to increase its performance in the future.

3.3 Graph Construction

To represent relational semantics, we first construct an undirected graph $G = (V, E)$ where a vertex represents an entity (i.e. a movie) and an edge denotes the presence of shared feature between the corresponding entities (such as sharing a director).

More precisely, if $A = (A_{i,j} : 1 \leq i \leq n)$ denotes the adjacency matrix of G, we have

$$A_{ij} = \begin{cases} 1 & \text{if } \{T^i_{feat} \cap T^j_{feat}\} \neq \emptyset. \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Here, T^i_{feat} denotes the feature set corresponding to entity i. Since there can be multiple combinations to create these feature set, we carefully chose five features that shows the best performance empirically: *director*, *producer*, *writer*, *cinematographer*, and *art director*.

For implementation purposes, we add a self loops to isolated vertices, i.e. those vertices with degree zero. The constructed graph G is on the whole train and test dataset. While train vertices are accessible to labels, we mask the labels for test vertices to prevent the model from seeing the ground truth during training phase.

3.4 Image-based Node Embedding (GAT)

Graphs representing relations within a single image is a well-studied problem as in (Guo et al., 2020; Johnson et al., 2015). However, no attempts have been made to represent image-objects as nodes input to a GNN. We define this novel graph as *image-based entity graph* as visualized in Figure 2.

Instead of using a complex image encoder, we use EfficientNet b4 (Tan and Le, 2019) to maximize efficiency. Then each encoded image is fed as node feature of an entity. Note that entire images represent nodes, not segments of images. Related works such as MMBT-Region (Kiela et al., 2021), VilBERT (Lu et al., 2019) and VisualBERT (Li et al., 2019) employs pretrained ResNet (He et al., 2015) based Faster-R-CNN, but they are overly expensive for GNN. That is because one single channel image is sufficient to enable an effective message passing.

While GraphSAGE (Hamilton, 2020) assigns the equal importance to neighbor nodes, in our application, depending on the context, different features can have different importance. Therefore, instead of using GraphSAGE, we employ GAT (Veličković et al., 2018) where it assigns different importance to different neighbor edges. This is done by

$$e_{ij} = a([U^l h_i^l ; U^l h_j^l]) \tag{5}$$

$$\alpha_{ij} = \frac{exp(e_{ij})}{\sum_{k \in \mathcal{N}_i} exp(e_{ik})} \tag{6}$$

$$h_i^{l+1} = \sigma(\sum_{j \in \mathcal{N}_i} \alpha_{ij} U^l h_j^l) \tag{7}$$

where a is a learnable weight vector for linear transformation. For non-linear activation function $\sigma(.)$, we use *LeakyReLU* function.

3.5 Contextualized Text Embedding

BERT (Devlin et al., 2019) achieved remarkable success in various downstream tasks with its unique tokenizing method and its self-attention mechanism. As visualized in Figure 2, we apply the same BERT tokenizer to textual data by tokenizing into 1) token embedding, 2) segment embedding, and 3) positional embedding. Their aggregated result is fed into a transformer and the final hidden state of this classification token is used for classification task. In figure 2, W_i denotes tokenized word given text data where i is sequence index.

3.6 Multimodal Bitransformer

MMBT (Kiela et al., 2020) is used as an early fusion method. This model originally extends BERT (Devlin et al., 2019) by applying BERT style tokenizing method into image modality as in Figure 2. For MM-GATBT, because we use image-based node embedding, we consider each node feature I^n as a token.

After applying BERT-like tokenization method in both Section 3.4 and Section 3.5, we concatenate them. Note that the original MMBT (Kiela et al., 2020) pools the image and uses multiple separate image embeddings. However, we only use one single output vector of image-based node embedding per each image.

3.7 Training

To solve multi-label classification task, we optimize binary cross-entropy loss defined as

$$\begin{aligned} \mathcal{L}_{bce} = -\frac{1}{M} \sum_{m=1}^{M} -\omega_m [y_m \log \hat{y}_m + \\ (1 - y_m) \log(1 - \hat{y}_m)] \end{aligned} \tag{8}$$

where M is the number of classes, ω_m is the fraction of samples of class m, y_m is true label, and $\hat{y}_m$ is predicted label. Because the MM-IMDb dataset is an imbalanced dataset, we assign different ω for different classes.

4 Experiment

System Configuration During the training phase, we used a single Nvidia RTX 3090 with a batch size of 12. We implemented our model using PyTorch (Paszke et al., 2019) and DGL (Wang et al., 2020) on top of MMBT code available on the public repository.[1] For every encoder, we used pre-trained models to reduce the computational cost and maximize their performance. In the case of the text encoder, we used the BERT uncased base model available from Hugginface (Wolf et al., 2020). For the image encoder, we used pre-trained EfficientNet b4 (Tan and Le, 2019). For GNN, we chose GAT (Veličković et al., 2018) available from DGL. We pre-trained GAT before employing to MM-GATBT. We used five features to construct our graph, as was explained in Section 3.3 and Eq. (4) therein. The average degree of the resulting graph is 59 and it has 554 isolated nodes.

Experiment Setup We used Multimodal IMDb (MM-IMDb) dataset from (Arevalo et al., 2017). This dataset consists of 23351 movie entities. Each movie in the dataset has a title, description, movie poster, producer, and related genres. Note that each movie can have multiple genres, making this task a multi-label classification task.

Empirical results from previous works imply that text modality carries more significant importance than image modality (Jin et al., 2021). The dataset is provided in a splitted format where the number of training set and testing set are 15552 and 7799 respectively.

Data Preprocessing We followed the data preprocessing scheme from (Kiela et al., 2020). The

[1] https://github.com/facebookresearch/mmbt

Type	Model	Micro F1	Macro F1	Weighted F1	Samples F1
Unimodal	EfficientNet (Tan and Le, 2019)	0.395	0.314	0.457	0.394
	BERT (Devlin et al., 2019)	0.645	0.587	0.645	0.647
Multimodal	GMU(Arevalo et al., 2017)	0.630	0.541	0.617	0.630
	CentralNet (Vielzeuf et al., 2018)	0.639	0.561	0.631	0.639
	MMBT (Kiela et al., 2020)	0.669	0.618	-	-
	MFM (Braz et al., 2021)	0.675	0.616	0.675	0.673
	ReFNet (Sankaran et al., 2022)	0.680	0.587	-	-
Graphical	*GAT w/ EfficientNet*	0.500	0.394	0.506	0.496
	MM-GATBT (ours)	**0.685**	**0.645**	**0.683**	**0.686**

Table 1: Experimental result shows that the proposed model outperforms against its unimodal submodels and popular multimodal models. For GMU (Arevalo et al., 2017), CentralNet (Vielzeuf et al., 2018), MMBT (Kiela et al., 2020), MFM (Braz et al., 2021), and RefNet (Sankaran et al., 2022), we brought the best numbers from their papers. Missing numbers mean that the results are not shared in their papers.

raw dataset (Arevalo et al., 2017) includes a total of 27 distinct labels from the training and testing set. However, the literature drops entities with News and Adult labels, leaving the training and the testing set with 15513 and 7779 entities respectively. Additionally, while labels with Reality-TV and Talk-Show are included in the training set, they do not appear in the testing set. Therefore, we test with 23 distinct labels as in the literature.

Baseline Models We compare MM-GATBT with two different types of models: unimodal models and multimodal models. For BERT (Devlin et al., 2019) and EfficientNet (Tan and Le, 2019) we use the same size of models used in the main model and compare their performance. For graphical model, we implement *GAT w/ EfficientNet* which outputs image-based node embedding used for the main model. Then we compare it with a single EfficientNet to examine the information gain from this structural difference. Our implementation is publicly available on GitHub.[2]

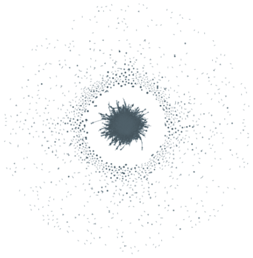

Figure 3: Example of constructed graph visualized using Pyvis (Perrone et al., 2020). Only 1 movie feature is used for visualization purposes.

5 Result

We validated our model using the following metrics: micro f1, macro f1, weighted f1, and samples f1. The results are rounded to 3 decimal places. We report our results as well as the state of the art in Table 1. Table 1 shows that MM-GATBT significantly outperforms baseline models in all metrics. Specifically, MM-GATBT significantly outperforms its unimodal submodels (i.e. considering text / image only) when ran separately. This performance increase can be explained from two perspectives. First, (Singh et al., 2020) addressed that the performance of pretraining models plays a critical role before fusion. As we suspected in Section 1, using image modality solely performs the worst, as it does not leverage the benefits of multimodal fusion. From this perspective, image-only embedding is upgraded into image-based node embedding as shown in *GAT w/ EfficientNet*. Therefore, as we observe, the main model performs better when its submodel performs better. This also indicates that our approach successfully captures the interaction between the entities through message passing.

Secondly, MM-GATBT reflects the connectivity structure of the constructed graph. As visualized in Figure 3, the constructed graph consists of both connected and isolated nodes. Therefore, it is crucial for the architecture to address the graph's density and sparsity. Indeed, the text encoder on the top of Figure 2 generates the word embedding neglecting the graph structure, which justifies its high performance on isolated nodes. In contrast,

[2] https://github.com/sbseo/mm-gatbt

the GAT on the bottom of Figure 2 takes into account the connectivity of nodes. This well justifies why MM-GATBT also performs well on non-isolated nodes. By fusing these two embeddings, MM-GATBT leverages both connected and isolated nodes effectively. Note that neither BERT nor image-based node embedding could achieve the accuracy of 0.685 before they were fused.

6 Conclusion

We proposed MM-GATBT, a novel graph-based multimodal architecture, to address the multilabel classification task on the MM-IMDb dataset. MM-GATBT leverages image-based node embedding and attention mechanism during the early fusion phase. The results show that the proposed model successfully captures the latent information generated from the interaction between the entities and achieves state-of-the-art results among all published works on the MM-IMDb dataset.

Acknowledgments

The authors would like to thank the reviewers for providing valuable feedback on this work.

References

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A. González. 2017. Gated Multimodal Units for Information Fusion. *arXiv:1702.01992 [cs, stat]*. ArXiv: 1702.01992 version: 1.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Khaled Bayoudh, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa. 2021. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer*.

Leodécio Braz, Vinícius Teixeira, Helio Pedrini, and Zanoni Dias. 2021. Image-text integration using a multimodal fusion network module for movie genre classification. In *11th International Conference of Pattern Recognition Systems (ICPRS 2021)*, volume 2021, pages 200–205.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Wenzhong Guo, Jianwen Wang, and Shiping Wang. 2019. Deep Multimodal Representation Learning: A Survey. *IEEE Access*, 7:63373–63394.

Xin Guo, Luisa Polania, Bin Zhu, Charles Boncelet, and Kenneth Barner. 2020. Graph neural networks for image understanding based on multiple cues: Group emotion recognition and event recognition as use cases. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive Representation Learning on Large Graphs. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

William L. Hamilton. 2020. *Graph Representation Learning*. Morgan & Claypool.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *arXiv:1512.03385 [cs]*. ArXiv: 1512.03385.

Woojeong Jin, Maziar Sanjabi, Shaoliang Nie, Liang Tan, Xiang Ren, and Hamed Firooz. 2021. MSD: Saliency-aware Knowledge Distillation for Multimodal Understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3557–3569, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2015. Image retrieval using scene graphs. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3668–3678.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. 2020. Supervised Multimodal Bitransformers for Classifying Images and Text. *arXiv:1909.02950 [cs, stat]*. ArXiv: 1909.02950.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2021. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. *arXiv:2005.04790 [cs]*. ArXiv: 2005.04790.

Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv:1609.02907 [cs, stat]*. ArXiv: 1609.02907.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. *arXiv:1908.03557 [cs]*. ArXiv: 1908.03557.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. *arXiv:1908.02265 [cs]*. ArXiv: 1908.02265.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. Multimodal Named Entity Disambiguation for Noisy Social Media Posts. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2000–2008, Melbourne, Australia. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Giancarlo Perrone, Jose Unpingco, and Haw-minn Lu. 2020. Network visualizations with Pyvis and VisJS. *arXiv:2006.04951 [cs]*. ArXiv: 2006.04951.

Sethuraman Sankaran, David Yang, and Ser-Nam Lim. 2022. Refining multimodal representations using a modality-centric self-supervised module.

Amanpreet Singh, Vedanuj Goswami, and Devi Parikh. 2020. Are we pretraining it right? Digging deeper into visio-linguistic pretraining. *arXiv:2004.08744 [cs]*. ArXiv: 2004.08744.

Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Kathryn Mazaitis, Ruslan Salakhutdinov, and William Cohen. 2018. Open domain question answering using early fusion of knowledge bases and text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4231–4242, Brussels, Belgium. Association for Computational Linguistics.

Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. *arXiv:1710.10903 [cs, stat]*. ArXiv: 1710.10903.

Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. 2018. CentralNet: a Multilayer Approach for Multimodal Fusion. *arXiv:1808.07275 [cs]*. ArXiv: 1808.07275.

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. 2020. Deep Graph Library: A Graph-Centric, Highly-Performant Package for Graph Neural Networks. *arXiv:1909.01315 [cs, stat]*. ArXiv: 1909.01315.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. 2021. A Comprehensive Survey on Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24. ArXiv: 1901.00596.