ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096188

MRML: MULTIMODAL RUMOR DETECTION BY DEEP METRIC LEARNING

Liwen Peng, Songlei Jian, Dongsheng Li**

National University of Defence Technology
College of Computer
{pengliwen13, jiansonglei,dsli}@nudt.edu.cn

Siqi Shen

Xiamen University
College of Computer
siqishen@xmu.edu.cn

ABSTRACT

Multimodal rumor detection aims at detecting rumors using information from textual and visual modalities. The most critical difficulty in multimodal rumor detection lies in capturing both the intra-modal and inter-modal relationships from multimodal data. However, existing methods mainly focus on the multimodal fusion process while paying little attention to the intra-modal relationships. To address these limitations, we propose a multimodal rumor detection method with deep metric learning (MRML) to effectively extract multimodal relationships of news for detecting rumors. Specifically, we design the metric-based triplet learning to extract the intra-modal relationships between rumors and non-rumors in every modality and the contrastive pairwise learning to capture the inter-modal relationships across multimodal. Extensive experiments on two real-world multimodal datasets show the superior performance of our rumor detection method.

***Index Terms*—** Rumor detection, deep metric learning, social media, multimodal learning

1. INTRODUCTION

As more and more people tend to seek out news and express their opinions through social media platforms, rumors spread more rapidly and widely. Rumors containing intentionally false information will mislead people into biased or fake information, break official organizations' credibility, and even cause riots [1]. Therefore, effectively detecting rumors from social media is in critical need.

Previous methods count more on the text content of news to judge whether it is a rumor or not [2, 3]. Hand-crafted textual statistical characteristics [4] or textual features learned by neural networks [5, 6] are used to represent the news. However, the information contained in plain text is limited for detecting rumors [7]. Researchers attempt to analyze the impact of image contents of news for detecting rumors [8, 9].

Recently, many works have been explored to detect rumors concerning textual and visual modalities [10, 11, 12]. However, detecting rumors from multimodal data is not a trivial task. The challenges of multimodal rumor detection mainly lie in capturing the intra-modal relationships between rumors and non-rumors in every single modality and exploring the inter-modal relationship of multimedia data across multimodal. Unfortunately, existing methods mainly focus on extracting the inter-modal relationship across multimodal but pay little attention to the intra-modal relationship. Methods [13, 14] use pre-trained language and vision models to extract the unimodal features, which are then concatenated as the multimodal representation of news. Some methods design multimodal fusion networks based on attention mechanism, Transformer, GAN, or VAE [15, 16, 17, 18]. However, these methods fail to extract the intra-model relationships between rumors and non-rumors in single modalities. Besides, authors in [19, 20] extract extra information from the real-world knowledge base to help detect rumors. Method [21] introduces the social context to assist the multimodal feature learning of the news. In contrast, our method attempts to capture both the intra- and inter-modal relationships of multimodal news without introducing extra information.

To address the challenges in multimodal rumor detection, we propose a novel Multimodal Rumor detection method based on deep Metric learning (MRML), which can not only discover the intra-modal relationships between rumors and non-rumors in every modality but also extract the inter-modal relationship across multimodal. Specifically, we design the metric-based triplet learning to extract the intra-modal relationships in both textual and visual modalities by measuring the distance between sampled rumor or non-rumor triplets. Then in the designed contrastive pairwise learning, we compare the similarities of sampled news pairs to extract the inter-modal relationship across multimodal. Finally, a rumor detection module is applied to the learned multimodal representation of news to detect rumors.

In summary, this work makes the following contributions: (1) We propose a method MRML based on language and vision for multimodal rumor detection, which can capture effective multimodal representations from the textual and visual modalities for detecting rumors. (2) We design the multimodal learning scheme based on deep metric learning for rumor detection, efficient at extracting intra-modal relation-

*Corresponding authors.

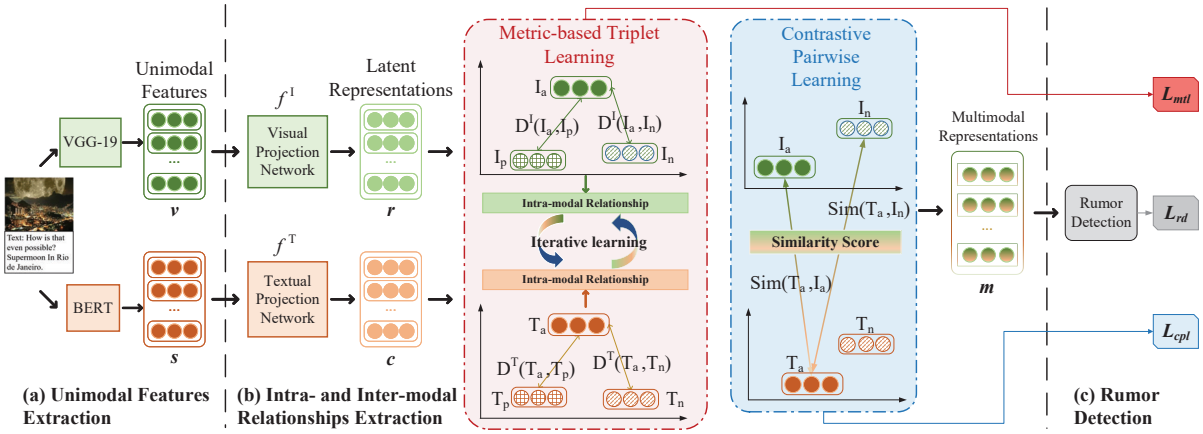

Fig. 1. The overall architecture of MRML. **(a)** extracts the modal-specific features of visual modality (green) and textual modality (orange); **(b)** captures the intra- and inter-modal relationships from multimodal data; **(c)** detects the rumors.

ships in every modality and inter-modal relationship across multiple modalities.

Extensive experiments on two real-world datasets demonstrate the effectiveness of MRML in distinguishing rumors compared with other rumor detection methods.

2. METHODOLOGY

Given the news containing text content T and image content I, the objective of rumor detection is to distinguish whether the news is a rumor ($y = 1$) or non-rumor ($y = 0$), i.e., to find a function $F(T, I) \rightarrow \hat{y} \in \{0, 1\}$. The overall architecture of our proposed MRML is illustrated in Fig. 1.

2.1. Unimodal Features Extraction

As different modalities contain unique characteristics, before learning the multimodal features of news, we apply unimodal feature extraction networks to capture the modality-specific unimodal features of text and image content. In recent years, the development of large-scale pre-trained models, language model BERT [22] and vision model VGG-19 [23] have been proven effective in capturing textual and visual semantic features [24, 25]. Thus in this work, we use BERT to extract the unimodal textual feature $\boldsymbol{s}$. Given the text content T, consisting of l tokens $T = \{t_1, t_2, ..., t_l\}$ (l represents the number of words), $\boldsymbol{s}$ is the concatenation of the l token embeddings calculated by BERT. Moreover, we use VGG-19 to extract the unimodal visual feature $\boldsymbol{v}$.

$$\boldsymbol{s} = \text{CONCAT}\big\{\text{BERT}(t_i)_{i=1}^{l}\big\}, \boldsymbol{v} = \text{VGG-19}(I). \quad (1)$$

2.2. Metric-based Triplet Learning

Typically, the features coming from different modalities reflect different data characteristics and have information gaps, resulting in that unimodal textual and visual features can not be compared and fused directly. Thus, we project unimodal features into a common latent vector space before capturing the multimodal relationships by the textual and visual projection networks f^T and f^I. Then we get the latent textual and visual representations $\boldsymbol{c}$ and $\boldsymbol{r}$ as follows:

$$\boldsymbol{c} = f^T(\boldsymbol{s}), \boldsymbol{r} = f^I(\boldsymbol{v}). \quad (2)$$

Since the appearance of triplet learning, it has shown the effective ability to learn data representations based on different distance relationships [26, 27]. Thus in MRML, we design a metric-based triplet learning to capture the intra-modal relationship in both textual and visual modalities.

Given the visual modality as a showcase, we first sample various triplets from the dataset, for example, (I_a, I_p, I_n). For each anchor I_a, the negative sample I_n has the opposite label with I_a (label means rumor or non-rumor), while the positive sample I_p has the same label as the anchor. To accurately distinguish the rumors and non-rumors, we aim to learn a distance metric that maps the anchor and positive sample closer (these two have the same label) and keeps the anchor far away from the negative sample (these two have the opposite label), as shown in Fig. 2. We can see that if the closest negative sample is far from the farthest positive sample, then the other triplets will also satisfy the distance relationship. Thus we sample such triplets to train our model.

Specificity, we define the distance metric function with the learned parameters $\boldsymbol{W}^I$ as:

$$D^I(I_i, I_j) = (\boldsymbol{r}_i - \boldsymbol{r}_j)\boldsymbol{W}^I(\boldsymbol{r}_i - \boldsymbol{r}_j). \quad (3)$$

Then we calculate the metric-based triplet learning loss based on the sampled triplets in image modality as:

$$L^I_{\text{mtl}} = \sum_{\mathcal{H}^I} \max\big\{0, \alpha_{\text{mtl}} - \Delta^T(D^I(I_a, I_p) - D^I(I_a, I_n))\big\}, \quad (4)$$

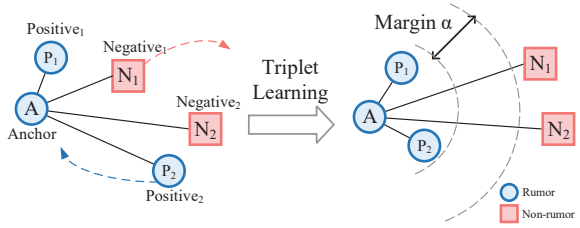

Fig. 2. The distance of samples in triplet learning.

where $\mathcal{H}^I = \{(I_a, I_p, I_n)\}$ is the set of sampled triplets, and α_{mtl} is a margin that keeps away the distance between the negative and positive samples. Further, we use Δ^T above to capture the complementary information in the textual modality to assist the learning of the intra-model relationship in the visual modality. Given the corresponding triplets (T_a, T_p, T_n) in textual modality, we have:

$$\Delta^T = \text{Sign}(\|\boldsymbol{c}_a - \boldsymbol{c}_p\|_2 - \|\boldsymbol{c}_a - \boldsymbol{c}_n\|_2), \tag{5}$$

where $\text{Sign}(x)$ indicates the sign of x.

By minimizing the metric-based triplet learning loss in the visual modality, the representations of rumors (non-rumors) are closely and separate from non-rumors (rumors), which reflects the intra-modal relationship in the visual modality. To further capture the textual intra-modal relationship, we follow the above process to sample triplets in textual modality and get the loss L^T_{mtl} as (4) with the complementary information Δ^I from the visual modality. We denote the other parameters of $L^{T/I}_{\text{mtl}}$ except for $W^{T/I}$ as θ_{mtl}. In MRML, we design iterative learning to effectively extract intra-modal relationships from both modalities. When minimizing L^I_{mtl}, W^T is fixed, while W^I and θ_{mtl} are updated. Similarly, when minimizing L^T_{mtl}, W^I is fixed, W^T and θ_{mtl} are updated.

2.3. Contrastive Pairwise Learning

Contrastive learning [28] has been proven effective in capturing data features. Thus in this work, we design the contrastive pairwise learning to capture the inter-modal relationship across multimodal. Specificity, for the sampled news pairs (T_a, I_a) and (T_n, I_n) with opposite labels, that is, if (T_a, I_a) is a rumor (non-rumor), (T_n, I_n) is a non-rumor (rumor), we enforce the similarity score of original pair (T_a, I_a) higher than the unpaired sample (T_a, I_n). In this work, we define the contrastive pairwise learning loss as:

$$L_{\text{cpl}} = \sum_{\mathcal{D}} \max\{0, \alpha_{\text{cpl}} - \text{Sim}(T_a, I_a) + \text{Sim}(T_a, I_n)\}, \tag{6}$$

where $\mathcal{D} = \{(T_a, I_a), (T_n, I_n)\}$ is the set of valid news pairs, and α_{cpl} is the margin. The Sim(,) function calculates the similarity score of a given pair:

$$\text{Sim}(T_i, I_j) = \sigma(\boldsymbol{W}_1 \boldsymbol{m}(T_i, I_j)), \tag{7}$$

where σ is the sigmoid function, $\boldsymbol{W}_1$ is the weight matrix, and $\boldsymbol{m}$ is the multimodal representation of (T_i, I_j) defined as:

$$\boldsymbol{m}(T_i, I_j) = \sum_{x=1}^{k} (W^T_x \boldsymbol{c}_i) \odot (W^I_x \boldsymbol{r}_j). \tag{8}$$

The $\boldsymbol{c}_i$ and $\boldsymbol{r}_j$ are latent representations of T_i and I_j as in (2). The W^T_x and W^I_x are the learned parameters, and $\odot$ is the element-wise product. The k is used to fuse the interactions between two modalities from different levels.

2.4. Rumor Detection

Finally, a rumor detection module g consisting of two fully connected layers with the relu activation function followed by the softmax function is applied to detect the news authenticity. The input is the multimodal representation $\boldsymbol{m}(T_i, I_i)$ of news, and the output is the probability of the input news being a rumor, denoted as $\hat{y} = g(\boldsymbol{m}(T_i, I_i))$. We use y to represent the ground-truth label and then employ cross-entropy to calculate the rumor detection loss:

$$L_{\text{rd}} = \sum -[y log(\hat{y}) + (1 - y) log(1 - \hat{y})]. \tag{9}$$

Overall, the objective of our rumor detection method is:

$$L = \lambda_{\text{mtl}} L^{T/I}_{\text{mtl}} + \lambda_{\text{cpl}} L_{\text{cpl}} + L_{\text{rd}}. \tag{10}$$

The λ_{mtl} and λ_{cpl} are weights of metric-based triple learning loss and contrastive pairwise learning loss, respectively.

3. EXPERIMENTS

3.1. Datasets and Settings

We conduct experiments on two widely used multimodal rumor detection datasets. The Weibo [21] dataset is collected from China's social media platform Weibo. The Twitter [29] dataset comes from the MediaEval Verifying Multimedia Use benchmark. Following the others [18, 15], we remove the news with videos and news without texts or images and divide the rest data into training and testing sets with a ratio of 8:2. In the experiments, there are 4211 rumors and 3642 non-rumors in Weibo with 12941 attached images. In Twitter, there are 5797 rumors and 5253 non-rumors with 451 images.

The token length l of text content is padded or truanted to 50, and the word dimension is 768. The dimension of unimodal visual feature is 4096. The textual/visual projection network is a fully connected layer of size 1024 with a relu activation function. We set the iterative frequency to 2 and the dimension of multimodal representation to 1024 with the fused layer $k = 2$. The learning rate is 0.00001 for Weibo and 0.0005 for Twitter. The dropout rate is 0.4. We use Adam optimizer to train our model with a batch size of 128 and 0.0001 weight decay. The optimal hyperparameters of our model are determined by grid searching. Thus we have $\alpha_{\text{mtl}} = \alpha_{\text{cpl}} = 0.2$, $\lambda_{\text{mtl}} = 0.3$, and $\lambda_{\text{cpl}} = 0.5$. The code of MRML is available at https://github.com/plw-study/MRML.

Table 1. Results of different rumor detection methods on two datasets. (* means the results are from the baseline paper.)

Method	Weibo							Twitter						
	Acc	Rumor			Non-rumor			**Acc**	Rumor			Non-rumor		
		Pre	Rec	**F1**	Pre	Rec	**F1**		Pre	Rec	**F1**	Pre	Rec	**F1**
Bert	0.845	0.858	0.833	0.845	0.833	0.857	0.844	0.642	0.666	0.766	0.711	0.602	0.474	0.526
VGG-19	0.647	0.640	0.700	0.668	0.657	0.591	0.621	0.767	**0.829**	0.753	0.787	0.704	0.785	0.740
att-RNN	0.772	0.854	0.656	0.742	0.720	0.889	0.795	0.664	0.749	0.615	0.676	0.589	0.728	0.651
EANN	0.782	0.827	0.697	0.756	0.752	0.863	0.804	0.648	0.810	0.498	0.617	0.584	0.759	0.660
MVAE	0.824	0.854	0.769	0.809	0.802	0.875	0.837	0.745	0.801	0.719	0.758	0.689	0.777	0.730
SpotFake	0.869	0.877	0.859	0.868	0.861	0.879	0.870	0.771	0.784	0.744	0.764	0.769	0.807	0.787
SpotFake+	0.870	0.887	0.849	0.868	0.855	0.892	0.873	0.790	0.793	0.827	0.810	0.786	0.747	0.766
SAFE	0.763	0.833	0.659	0.736	0.717	0.868	0.785	0.766	0.777	0.795	0.786	0.752	0.731	0.742
CAFE*	0.840	0.855	0.830	0.842	0.825	0.851	0.837	**0.806**	0.807	0.799	0.803	**0.805**	**0.813**	**0.809**
MRML-C	0.834	0.871	0.766	0.814	0.808	0.895	0.849	0.747	0.783	0.784	0.782	0.702	0.696	0.697
MRML-T	0.852	0.872	0.810	0.839	0.838	0.890	0.862	0.780	0.795	0.835	0.814	0.758	0.704	0.729
MRML	**0.897**	**0.898**	**0.887**	**0.892**	**0.896**	**0.905**	**0.901**	0.803	0.821	**0.844**	**0.832**	0.777	0.747	0.762

3.2. Baselines

We compare MRML against other rumor detection methods from two categories: unimodal and multimodal.

Unimodal methods rely on unimodal information to detect rumors. We use Bert/VGG-19 to get textual/visual unimodal representations. Then the learned unimodal representations are fed into a fully connected layer with a softmax function to perform the rumor detection. **Multimodal methods** use both text and image contents to distinguish rumors. Att-RNN [21] uses attention to fuse the image and text features. MVAE [18] uses two VAE models to reconstruct the unimodal features. EANN [17] uses an event discriminator network to extract event-invariant features. SpotFake [13] and SpotFake+ [14] concatenate the unimodal features extracted by pre-trained language and vision models. SAFE [10] designs a similarity-aware method to learn multimodal features. CAFE [11] designs a cross-modal ambiguity learning module for estimating the ambiguity between different modalities.

3.3. Results and Analysis

Table 1 displays the rumor detection results on the two datasets, including the overall Accuracy and Precision, Recall, and F1 scores for rumor and non-rumor, respectively.

On Weibo, our proposed method MRML outperforms all the other baselines on all the metrics, which means MRML successfully extracts the intra- and inter-modal relationships and learns better multimodal representations for detecting rumors. Also, the multimodal methods SpotFake and SpotFake+ achieve better results than the unimodal methods, which means extracting multimodal features from both textual and visual modalities is helpful in detecting rumors. On Twitter, MRML gets a comparable Accuracy with CAFE and achieves the highest F1 score in detecting rumors among all the methods. The multimodal methods receive higher accuracy scores than the unimodal methods, which means learning multimodal features can benefit rumor detection.

3.4. Ablation Study

To validate the effectiveness of different components of our method MRML, we conduct the ablation study and report the results in the bottom part of Table 1. The sub-model MRML-C has the same network structure as MRML but training without the contrastive pairwise learning loss. MRML-T is the sub-model without metric-based triplet learning loss.

Comparing the performance of MRML-C and MRML, we can observe that the model benefits a lot from extracting inter-modal relationships through contrastive pairwise learning. Without metric-based triplet learning, MRML-T receives a performance decline compared to MRML, which shows the effectiveness of capturing intra-modal relationships between rumors and non-rumors in each modality.

4. CONCLUSIONS

In this paper, we propose a multimodal rumor detection method based on deep metric learning (MRML) to effectively distinguish rumors. Specifically, we design metric-based triplet learning and contrastive pairwise learning to discover and capture the intra-modal relationships in different modalities and the inter-modal relationships across multiple modalities. Extensive experiments conducted on two widely used datasets demonstrate the superior performance of our proposed method compared to other rumor detection methods.

5. ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (No. 62002371, 62025208), Foundation of PDL (No. WDZC20205250104, WDZC20215250113), Foundation of National University of Defense Technology (No. ZK21-17).

6. REFERENCES

[1] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *SIGKDD*, vol. 19, no. 1, pp. 22–36, Sept. 2017.

[2] Z. Jin, J. Cao, Y. G. Jiang, and Y. Zhang, "News credibility evaluation on microblog with a hierarchical propagation model," in *ICDM*, 2015.

[3] Z. Jin, J. Cao, Y. Zhang, and J. Luo, "News verification by exploiting conflicting social viewpoints in microblogs," in *AAAI*, 2016, AAAI'16, p. 2972–2978.

[4] K. Popat, S. Mukherjee, J Strötgen, and G. Weikum, "Credibility assessment of textual claims on the web," in *CIKM*, 2016.

[5] T. Chen, X. Li, H. Yin, and J. Zhang, "Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection," in *PAKDD*, 2018, pp. 40–52.

[6] Y. Liu and Y. Wu, "Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks," in *AAAI*, 2018.

[7] N. Ruchansky, S. Seo, and Y. Liu, "Csi: A hybrid deep model for fake news detection," in *CIKM*, New York, NY, USA, 2017, CIKM '17, p. 797–806.

[8] Z. Jin, J. Cao, Y. Zhang, J. Zhou, and T. Qi, "Novel visual and statistical image features for microblogs news verification," *TMM*, vol. 19, no. 3, pp. 598–608, 2017.

[9] W. Ke, Y. Song, and K. Q. Zhu, "False rumors detection on sina weibo by propagation structures," in *IEEE International Conference on Data Engineering*, 2015.

[10] X. Zhou, J. Wu, and R. Zafarani, "Safe: Similarity-aware multi-modal fake news detection," in *PAKDD*, 2020, pp. 354–367.

[11] Y. Chen, D. Li, P. Zhang, J. Sui, Q. Lv, L. Tun, and L. Shang, "Cross-modal ambiguity learning for multimodal fake news detection," in *WWW*, 2022, p. 2897–2905.

[12] Z. Wei, H. Pan, L. Qiao, X. Niu, P. Dong, and D. Li, "Cross-modal knowledge distillation in multi-modal fake news detection," *ICASSP*, pp. 4733–4737, 2022.

[13] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, and S. Satoh, "Spotfake: A multi-modal framework for fake news detection," in *BigMM*, 2019.

[14] S. Singhal, A. Kabra, M. Sharma, R. R. Shah, T. Chakraborty, and P. Kumaraguru, "Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract)," 05 2020.

[15] S. Qian, J. Wang, J. Hu, Q. Fang, and C. Xu, "Hierarchical multi-modal contextual attention network for fake news detection," in *SIGIR*, 2021, p. 153–162.

[16] Y. Wu, P. Zhan, Y. Zhang, L. Wang, and Z. Xu, "Multimodal fusion with co-attention networks for fake news detection," in *ACL*, Aug. 2021, pp. 2560–2569.

[17] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, and J. Gao, "Eann: Event adversarial neural networks for multi-modal fake news detection," in *KDD*, 2018.

[18] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, "Mvae: Multimodal variational autoencoder for fake news detection," in *WWW*, 2019, p. 2915–2921.

[19] H. Zhang, Q. Fang, S. Qian, and C. Xu, "Multi-modal knowledge-aware event memory network for social media rumor detection," in *ACM MM*, 2019, p. 1942–1951.

[20] Y. Wang, S. Qian, J. Hu, Q. Fang, and C. Xu, "Fake news detection via knowledge-driven multimodal graph convolutional networks," in *ICMR*, 2020.

[21] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in *MM*, 2017, p. 795–816.

[22] J. Devlin, M.i Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL*, June 2019, pp. 4171–4186.

[23] K. Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2015.

[24] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained models for natural language processing: A survey," *CoRR*, vol. abs/2003.08271, 2020.

[25] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *CoRR*, vol. abs/1901.06032, 2019.

[26] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, 2015, pp. 815–823.

[27] T. Wang, X. Xu, Y. Yang, A. Hanjalic, H. T. Shen, and J. Song, "Matching images and text with multi-modal tensor fusion and re-ranking," in *MM*, 2019, p. 12–20.

[28] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *CVPR*, 2006, vol. 2, pp. 1735–1742.

[29] C. Boididou, S. Papadopoulos, D. T. Dang Nguyen, G. Boato, M. Riegler, A. Petlund, and I. Kompatsiaris, "Verifying multimedia use at mediaeval 2016," in *MediaEval 2016 Workshop*, 2016.