Multi-Modality Guidance Network for Missing Modality Inference

Anonymous ACL submission

Abstract

Recent advancements in multimodal models have showed promise, yet their dependency on consistent modalities from training to inference limits their application. While existing methods mitigate the problem through reconstructing the missing modalities, they increase unnecessary computational cost, which could be just as critical, especially for large, deployed systems. To address these issues, we propose a novel multimodal guidance network that promotes knowledge sharing during training, taking advantage of the multimodal representations to train better single-modality models for inference. Real-life experiment in violence detection shows that our proposed framework trains single-modality models that significantly outperform its traditionally trained counterparts while maintaining the same inference cost. Code will be made public upon acceptance.

1 Introduction

Multimodal deep learning (Ngiam et al., 2011) has garnered significant interest for its capacity to integrate data from diverse modalities, mirroring human perception and often enhancing performance across various machine learning (ML) tasks (Baltrušaitis et al., 2018; Akkus et al., 2023). This approach has shown considerable success in numerous vision-language domains such as video understanding (Nagrani, 2020; Palaskar, 2022), image captioning (Yu et al., 2019; Zhao et al., 2019), and many others (Zong et al., 2023). However, a common assumption in these models is the consistent presence of all modalities from training through to inference, which can be a limiting factor (Ma et al., 2022). The dependency on having all modalities available during inference can restrict their practical application, as gathering multimodal data is often more challenging during inference (Woo et al., 2023). Therefore, developing new approaches that can take advantage of multiple modalities during training while being robust with missing modalities during inference is in urgent need.

Addressing the issue of missing modalities in multimodal deep learning has emerged as a new research area (Tsai et al., 2018; Ma et al., 2021). Many studies focus on generating features for the absent modality. For example, Ma et al. (Ma et al., 2021) introduced SMIL, a Bayesian Meta-Learning approach for estimating missing-modality features. However, this method is limited to specific model-modality combinations such as ResNet (He et al., 2016) for images and LSTM (Hochreiter and Schmidhuber, 1997) for texts. Ma et al. (Ma et al., 2022) adapted this concept to the more versatile Transformer model (Vaswani et al., 2017). Additionally, Woo et al. (Woo et al., 2023) proposed ActionMAE, which learns to reconstruct missing modality features by employing random drop.

While current research offers valuable insights into generating features for missing modalities, this suggests that multimodal features may be more informative than previously utilized in conventional training-inference methods. This raises the question of whether we can use these features to improve single-modality models for inference, making the issue of missing modalities less critical.

Following this idea, in this paper, we introduce a novel multi-modality guidance network that leverages multimodal representations for training more effective single- modality models for inference. Unlike previous approaches, our framework focuses on using multimodal fusion representations as a guide for training single-modality encoders, which are then used in inference. This method avoids the need to generate missing features and instead uses the multimodal representations to enhance the performance of single-modality models, potentially reducing latency and costs in situations where a modality is missing during inference.

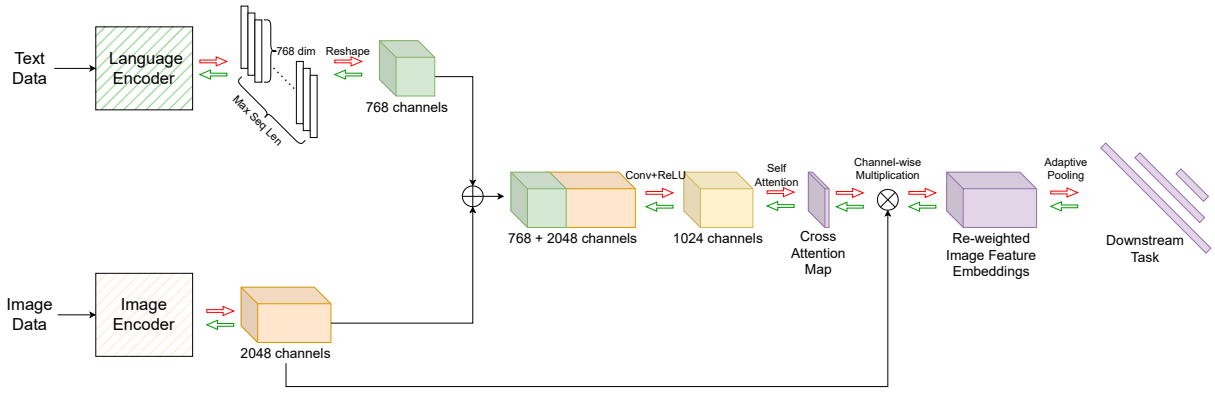

Figure 1: An overview of the proposed guidance network in a vision-language setup. The network begins with encoding text and image features separately, fusing the embeddings from both modalities, and then applies self-attention to obtain the attention map. However, instead of applying the attention map back to the fusion embeddings, we apply it to the image-only embeddings to promote knowledge sharing from cross-modality features to better singlemodal attention.

2 Methodology

Multimodal deep learning spans a wide range of fields in ML, with significant focus on natural language processing (NLP) (Garg et al., 2022; Yin et al., 2023), computer vision (Bayoudh et al., 2021; Bi et al., 2022), and etc. (Zong et al., 2023). In the realm of NLP, the integration of textual and visual modalities is a key area of study, leading to the development of vision-language models (VLMs) (Zhang et al., 2023). Our paper concentrates on showcasing the proposed guidance network within a vision-language framework, particularly emphasizing the fusion and enhancement of textual modalities in conjunction with visual data.

As outlined in Sec. 1, VLMs also typically require consistency in modalities between training and inference phases. Our approach, diverging from the generation of textual features from images or vice versa (Ma et al., 2021, 2022; Woo et al., 2023), focuses on developing a single-modal encoder that incorporates both image and text features during training. This encoder is designed for effective standalone application during inference.

The process begins with separate text and image encoders, then uniquely applies a cross-modality attention map from fusion embeddings to image embeddings. These re-weighted image embeddings are then used for training tasks, exemplified by a violence detection case study we show in Sec. 3. The overall structure of our guidance network during the training phase is depicted in Fig. 1.

Text Embeddings. To prepare text features, the language encoder begins by tokenizing input texts, followed by padding the tokens to a maximum sequence length. This results in embeddings for each token with a specific hidden dimension, which are user-defined hyper-parameters. In our experiments, unless otherwise stated, the maximum sequence length and hidden dimension are set to 121 and 768, respectively. The resulting text embeddings are then formatted into a block with the shape $\left[\texttt{hidden\_dim}, \sqrt{\texttt{seq\_length}}, \sqrt{\texttt{seq\_length}}\right]$ in the channel-first format.

Image Embeddings. To prepare image features, our process involves the use of feature-extracting backbones as image encoders, as depicted in Fig. 1. In practice, we utilize well-known architectures such as ResNet50 (He et al., 2016), Vision Transformer (ViT) (Dosovitskiy et al., 2020), and MobileOne (Vasu et al., 2023). These varied backbones allow for a comprehensive approach to creating image embeddings.

Text-Image Embeddings Fusion. After post-processing and reshaping the text embeddings, we can efficiently concatenate them with the image embeddings along the hidden dimension axis. A straightforward convolutional neural network (CNN) consisting of three convolutional layers, followed by ReLU (Fukushima, 1975) and Batch Normalization (Ioffe and Szegedy, 2015) layers is then used for multimodal fusion. This fusion method, despite its simpleness, can effectively integrate the distinctive features of both modalities.

Text-Guided Image Re-Weighting. To enhance image encoder training with text features, we apply self-attention (Vaswani et al., 2017) to the fusion embeddings, generating a cross-modality attention

map. This map is then applied to the original, pre-fusion image embeddings, instead of to the fusion embeddings as seen more commonly in existing works. The re-weighted image embedding, now influenced by textual information, is used for task-specific training. This method transfers textual insights to the image feature domain, enabling text-derived high-level descriptions or specific details, which are not extractable from images alone, to aid in training the image backbone.

3 Experiment

We evaluated our multimodal guidance network on a real-life violence detection task (Yao and Hu, 2023), aiming to identify images containing violent content. Technically speaking, violence detection is a binary classification task. However, it presents greater difficulty due to the overlap in the visual distribution between realistic in-game images and actual violent events, leading to high rates of false-positives where non-violent in-game graphics are incorrectly classified as real violence. Unfortunately, this has proved to be complex even for advanced deep learning models (Yao and Hu, 2023). And in practice, live monitoring of the violent streaming requires an extensive amount of human verifier to account for these high false-positives.

Although advancements in vision encoders alone can further solve this issue (Yao and Hu, 2023), very often we notice this could be mitigated much easier if textual information, such as the streamer's streaming history, could be integrated when making the predictions. However, these information are often hard to gather during inference, even for matured industry platforms, which makes this task a natural fit for our proposed guidance network.

3.1 Data Collection

Since there is no existing public dataset that contains both images and the corresponding textual information, we collect our own dataset from the logged historical streaming contents with ground truths (`violent` or `non-violent`) labeled by professionally trained human judges.

For each image, we formulate a corresponding text caption that contains descriptions consist of streamer's past streaming history as well as various meta-data such as the streaming title, user-defined streaming category, and streaming device. In total, we collect around 150,000 samples, and are divided into train and test set with a 85/15 split ratio.

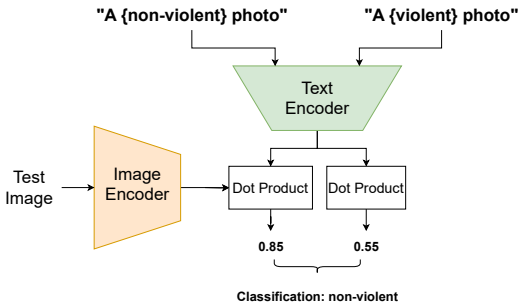

Figure 2: Illustration of CLIP zero-shot classification. CLIP encodes both image and all potential class captions, compares similarity scores between image and each text embedding and takes the higher one as the classification result. In the case shown here, it classifies the input image as non-violent.

3.2 Baseline Approaches

Contrastive Language Image Pretraining (CLIP). The first baseline in our experiment is the pre-trained CLIP, known for its robust performance in various zero-shot classification tasks (Radford et al., 2021). Considering our task involves identifying violent images, which might also be present in other internet-sourced datasets, we anticipate that large foundation models like CLIP, trained on extensive internet data, should exhibit commendable performance even in a zero-shot scenario.

Fig. 2 details our approach for the baseline evaluation using the pre-trained CLIP. We utilize CLIP's pre-trained image and text encoders and compare the cosine distance between the image and the text embeddings. Then the classification is determined by choosing the pair with the closer match. The performance tested with various image encoders pre-trained with CLIP is presented in Table 1, with latency measurements conducted on a single NVIDIA V100 GPU.

Considering both performance and latency, CLIP (ViT-B/16) performs the best. This configuration will be used as the representative for CLIP's performance in subsequent comparisons throughout this paper.

Model	Precision	Recall	Accuracy	Latency
RN50	94.79%	65.63%	90.32%	0.56ms
RN101	90.63%	67.19%	89.86%	0.69ms
ViT-B/32	**95.01%**	47.70%	86.03%	**0.26ms**
ViT-B/16	92.94%	**89.31%**	**95.54%**	0.88ms
ViT-L/14	93.94%	85.26%	94.84%	3.39ms

Table 1: Violence detection performance of the pre-trained CLIP with various vision encoders.

Model	Precision	Recall	Accuracy	Latency
MO-S0	97.27%	86.33%	95.90%	**0.11ms**
MO-S1	**97.61**%	81.30%	94.73%	0.11ms
MO-S2	96.16%	86.45%	95.67%	0.12ms
MO-S3	94.33%	89.68%	95.99%	0.11ms
MO-S4	94.05%	**93.53%**	**96.84%**	0.14ms

Table 2: Fine-tuned MobileOne performance on violence detection using our dataset.

Model	Precision	Recall	Accuracy	Latency
CLIP	92.94%	89.31%	95.54%	0.88ms
MO-S4	94.05%	93.53%	96.84%	0.14ms
MO-S4 Frozen	95.79%	**96.92**%	98.13%	**0.13**ms
MO-S4 Unfrozen	**97.84**%	95.50%	**98.32**%	**0.13**ms

Table 3: Performance comparisons between our proposed multimodal guidance network and two baselines.

MobileOne. MobileOne (MO) (Vasu et al., 2023) is a state-of-the-art, efficient neural network backbone that has demonstrated superior performance in image classification tasks such as ImageNet (Deng et al., 2009). After initializing with weights from the pre-trained model, we adapt the classification head to our dataset and fine-tuned the model specifically for our violence detection task.

Table 2 displays the performance of all MO variants. As the variant number increases from MO-S0 to MO-S4, the model's complexity, number of trainable parameters, and performance also increase. We refer readers to the original MO paper (Vasu et al., 2023) for more details on these variants. Of all the variants, MO-S4 showed the best results and will be used as the representative for MO performance in subsequent comparisons.

3.3 Results of Our Guidance Approach

To evaluate our guidance network's ability to enhance single-modal model training using multimodal data, we employ MO-S4 as our image encoder, consistent with the baseline approach. For text processing, we use DistilBERT (Sanh et al., 2019), a streamlined version of BERT (Devlin et al., 2018), to handle the text captions. The text captions, being natural sentences, allow for flexibility in freezing or unfreezing backpropagation during training. The comparative results, including those of the two baselines, are compiled in Table 3.

3.4 Discussions

The guidance network we proposed excels in all metrics, including precision, recall, accuracy, and latency, significantly outperforming the strong baselines. Given that the image encoder we employ is the same as the one in MO-S4 baseline, this performance gain empirically validates our hypothesis that the guidance network can leverage multimodal representations to train a more efficient single-modal model, while maintaining the same network complexity and low latency.

Interestingly, no definitive advantage is observed between guidance-trained image encoders when the language encoder's training status (frozen or unfrozen) varied. We suspect this is due to the natural-language text captions aligning with DistilBERT's training distribution, rendering additional finetuning of the language encoder less impactful.

4 Conclusions and Future Work

In this paper, we address the missing modality inference problem within multimodal deep learning. While current research often favors generative methods to compensate for missing modalities by using existing features, our paper proposes a different strategy focusing on harnessing the strengths of multimodal data during the training phase to develop a more efficient single-modal model. Empirical results confirm our hypothesis that our guidance network can significantly improve single-modal models compared to counterparts with the same architecture but fine-tuned traditionally, making our method balance the effectiveness of multimodal approaches with cost efficiency.

Inspired by our findings, we believe that it is worthwhile to continue researching in this direction, and could start with the following aspects. First, better-designed attention mechanism within the guidance network could be studied to promote better knowledge sharing and transferring. Next, different modality pairs, such as image-audio, or text-audio, could be further explored for additional downstream tasks such as video understanding, sentiment analysis, and many others. Last by not least, we also believe that such knowledge sharing idea could be further extended to more than two modalities to develop even more powerful yet efficient solutions for missing modality inference.

Limitations

While our proposed guidance network shows effectiveness in mitigating missing modality problem as well as proficiency in handling specific tasks such as violence detection, its adaptability to a broader range of applications remains to be explored. Additionally, the current model architecture might not be optimally efficient for real-time applications due to computational demands, particularly when processing large-scale datasets. Future iterations of this research will aim to address these limitations, focusing on enhancing the versatility and computational efficiency of the model.

References

Cem Akkus, Luyang Chu, Vladana Djakovic, Steffen Jauch-Walser, Philipp Koch, Giacomo Loss, Christopher Marquardt, Marco Moldovan, Nadja Sauter, Maximilian Schneider, et al. 2023. Multimodal deep learning. *arXiv preprint arXiv:2301.04856*.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

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*, pages 1–32.

Ying Bi, Bing Xue, Pablo Mesejo, Stefano Cagnoni, and Mengjie Zhang. 2022. A survey on evolutionary computation for computer vision and image analysis: Past, present, and future trends. *IEEE Transactions on Evolutionary Computation*, 27(1):5–25.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Kunihiko Fukushima. 1975. Cognitron: A self-organizing multilayered neural network. *Biological cybernetics*, 20(3-4):121–136.

Muskan Garg, Seema Wazarkar, Muskaan Singh, and Ondřej Bojar. 2022. Multimodality for nlp-centered applications: Resources, advances and frontiers. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6837–6847.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. 2022. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18177–18186.

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. 2021. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2302–2310.

Arsha Nagrani. 2020. *Video understanding using multimodal deep learning*. Ph.D. thesis, University of Oxford.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696.

Shruti Palaskar. 2022. *Multimodal Learning from Videos: Exploring Models and Task Complexities*. Ph.D. thesis, Carnegie Mellon University.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2018. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176*.

Pavan Kumar Anasosalu Vasu, James Gabriel, Jeff Zhu, Oncel Tuzel, and Anurag Ranjan. 2023. Mobileone: An improved one millisecond mobile backbone. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7907–7917.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Sangmin Woo, Sumin Lee, Yeonju Park, Muhammad Adi Nugroho, and Changick Kim. 2023. Towards good practices for missing modality robust action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2776–2784.

Huiling Yao and Xing Hu. 2023. A survey of video violence detection. *Cyber-Physical Systems*, 9(1):1–24.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Jun Yu, Jing Li, Zhou Yu, and Qingming Huang. 2019. Multimodal transformer with multi-view visual representation for image captioning. *IEEE transactions on circuits and systems for video technology*, 30(12):4467–4480.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2023. Vision-language models for vision tasks: A survey. *arXiv preprint arXiv:2304.00685*.

Dexin Zhao, Zhi Chang, and Shutao Guo. 2019. A multimodal fusion approach for image captioning. *Neurocomputing*, 329:476–485.

Yongshuo Zong, Oisin Mac Aodha, and Timothy Hospedales. 2023. Self-supervised multimodal learning: A survey. *arXiv preprint arXiv:2304.01008*.