This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Expanding Large Pre-trained Unimodal Models with Multimodal Information Injection for Image-Text Multimodal Classification

Tao Liang[1,2] Guosheng Lin[3] Mingyang Wan[2] Tianrui Li[1] Guojun Ma[2] Fengmao Lv[1*]
[1] Southwest Jiaotong University
[2] Engineering Productivity & Quality Assurance of IES, Bytedance
[3] Nanyang Technological University
{fengmaolv,taoliangdpg}@126.com {wanmingyang,maguojun}@bytedance.com
gslin@ntu.edu.sg trli@swjtu.edu.cn

Abstract

Fine-tuning pre-trained models for downstream tasks is mainstream in deep learning. However, the pre-trained models are limited to be fine-tuned by data from a specific modality. For example, as a visual model, DenseNet cannot directly take the textual data as its input. Hence, although the large pre-trained models such as DenseNet or BERT have a great potential for the downstream recognition tasks, they have weaknesses in leveraging multimodal information, which is a new trend of deep learning. This work focuses on fine-tuning pre-trained unimodal models with multimodal inputs of image-text pairs and expanding them for image-text multimodal recognition. To this end, we propose the Multimodal Information Injection Plug-in (MI2P) which is attached to different layers of the unimodal models (e.g., DenseNet and BERT). The proposed MI2P unit provides the path to integrate the information of other modalities into the unimodal models. Specifically, MI2P performs cross-modal feature transformation by learning the fine-grained correlations between the visual and textual features. Through the proposed MI2P unit, we can inject the language information into the vision backbone by attending the word-wise textual features to different visual channels, as well as inject the visual information into the language backbone by attending the channel-wise visual features to different textual words. Armed with the MI2P attachments, the pre-trained unimodal models can be expanded to process multimodal data without the need to change the network structures.

1. Introduction

In social media such as Twitter, a tweet usually contains both the text and image contents which share the same concept. With the increased use of social media, a massive number of multimodal user-generated contents can be available for training deep models. It is clear that multimodal classification can gain a nontrivial advantage over the unimodal counterpart by using information from both the visual and language modalities [22]. Over the past years, image-text multimodal classification has been widely applied to different social media projects such as emergency response [1, 2], emotional recognition [31], fake news detection [25], etc.

The core idea in image-text multimodal classification is to integrate the image and texts together. In general, the current works for image-text multimodal recognition can be categorized into two strategies. The first strategy maintains two separate backbones (e.g., DenseNet or BERT) to process each modality and performs multimodal fusion on the classification scores or the high-level features produced by each backbone [1, 7, 15]. On the other hand, the second strategy goes in-depth into the intermediate layers of the backbones and performs multimodal fusion on the fine-grained mid-level features of each modality [13, 16, 17, 29]. However, the current works along this line mainly focus on the homogeneous setting in which the modalities are just different views of the same input (e.g., RGB and depth images) [13, 29, 34]. Due to the strong heterogeneity between the mid-level features of images and texts, the second strategy is less studied for the image-text multimodal fusion task. The recently proposed multimodal BERT can model the inter-modal interactions between the fine-grained mid-level features of the visual and language modalities based on the recent advances of Transformer [12, 16–19, 23]. As large pre-trained models, multimodal BERT can be fine-tuned for image-text multimodal recognition.

The previous works have shown that an efficient multimodal classification algorithm needs to consider both the intra-modal processing and inter-modal interaction [13, 29].

* Corresponding author: F. Lv (email: fengmaolv@126.com).

To be specific, the intra-modal processing requires to extract the discriminative semantic information from each modality, which is crucial for the classification task, while the inter-modal interaction requires to fully integrate the content of each modality. In general, the first strategy does well in intra-modal processing by maintaining separate unimodal backbones to process each modality, but has weaknesses in modeling sufficient inter-modal interaction [1, 7, 15]. On the other hand, the multimodal BERT models from the second strategy do well in inter-modal interaction by attending to the fine-grained token features of each modality, but underestimate the end-to-end intra-modal processing of each modality (e.g., directly take the region features extracted from faster-RCNN as visual inputs [17–19]; directly aggregate the original image patches or textual features [16]). Although the recent PixelBERT proposes to leverage an end-to-end CNN backbone to extract the image features [12], the intra-modal processing is still prone to be underestimated once the mid-level features of each modality have been input into the transformer layers [29]. The stacked transformer layers cut off the direct connection between the CNN backbone and the final prediction. Compared with the pre-trained multimodal BERT models, the large pre-trained unimodal models (e.g., DenseNet or BERT) carefully consider the end-to-end intra-modal processing and have a strong ability in extracting the discriminative semantic information from each modality.

Motivated by the above discussion, this work focuses on directly expanding the large pre-trained unimodal models for image-text multimodal recognition, with the consideration of both effective intra-modal processing and inter-modal interaction. Our core idea is to integrate the features from other modalities to augment the mid-level features of the unimodal models. To this end, we propose the Multimodal Information Injection Plug-in (MI2P) attached to the mid-level layers of the unimodal networks (e.g., DenseNet or BERT). In order to bridge the heterogeneity across different modality features, MI2P performs cross-modal feature transformation by learning the fine-grained cross-modal attentions between the visual and textual features. Through the MI2P unit, the language information can flow into the visual backbone by attending the word-wise textual features to different visual channels. Similarly, the visual information can also flow into language backbone by attending the channel-wise visual features to different textual words. By fine-tuning the unimodal backbone together with the attached MI2P units, the injected multimodal information can be adapted to augment the mid-level features in a proper manner, i.e., enrich the semantic patterns of the mid-level features but not suppress their intra-modal processing.

Compared with the existing image-text multimodal classification methods [1, 9, 14, 17, 18], our approach can better balance the inter-modal interaction and intra-modal processing. For the former purpose, the fine-grained cross-modal interactions are explicitly modeled within the MI2P attachment. In practice, the visual and textual modalities are usually correlated on different abstraction levels (e.g., in the sentence of "*An elephant is drinking from the stream with its long nose*", the word *nose* may relate to the mid-level visual features of images, while the word *elephant* may relate to the high-level features of images). The MI2P plug-ins can be flexibly attached to multiple layers of the unimodal networks, in order to model the cross-modal interactions of different abstraction levels. For the latter purpose, our approach completely preserves the original network structures of the large pre-trained unimodal models. As plug-ins, the MI2P units will not suppress the intra-modal processing of the unimodal models.

To sum up, the contributions of this work are three-fold:

- We propose to expand the large pre-trained unimodal models for image-text multimodal classification by arming them with the introduced Multimodal Information Injection Plug-in units. The proposed implementation of multimodal recognition can preserve the strong intra-modal processing ability of the large pre-trained unimodal models.
- Our approach can model the cross-modal interactions of different abstraction levels by attaching the MI2P units to multiple layers of the unimodal models, with the consideration of sufficient inter-modal interaction.
- Our approach can obtain state-of-the-art performance across different image-text multimodal classification benchmarks.

2. Related Works

2.1. Image-text multimodal classification

Image-text multimodal classification aims to improve the performance over the unimodal counterpart by integrating information from both the visual and language modalities [22]. Over the past years, multimodal classification has been widely applied into various social media projects such as emergency response [1, 2], emotional recognition [31], fake news detection [25], etc. According to where the modalities are integrated, we can categorize the current multimodal recognition approaches as two strategies. The first strategy is the predominant method which maintains two separate backbone network (e.g., DenseNet or BERT) to process each modality and performs multimodal fusion on the classification scores [8, 30] or the high-level features produced from each backbone network [1, 4, 7, 15, 33] via aggregation operations such as addition [15], outer product [7], cross-gating [1], tensor fusion [33], etc. The main drawback for this strategy lies in insufficient inter-modal interaction. Hence, the second strategy mainly focuses on performing multimodal fusion on the fine-grained mid-level

features of each modality [12, 14, 17–19, 29, 34]. In particular, the recently proposed multimodal BERT models stack transformer layers over the mid-level features of images and texts and can be fine-tuned for image-text multimodal recognition [12, 14, 17, 18]. The fine-grained inter-modal interactions between the textual words and the visual tokens can be modeled by the attention mechanism.

2.2. Pre-training

The pre-training paradigm is one of the main causes that lead to the great success of deep learning. Fine-tuning large pre-trained models for specific downstream tasks is currently a common notion in the field of both computer vision and natural language processing. For example, deep convolutional neural networks (e.g., ResNet [10] or DenseNet [11]) pre-trained on ImageNet have been widely used as the standard baselines to process visual signals like images or videos. In the recent year, various large pre-trained models which are not based on convolutional operations are also proposed [6, 24]. On the other hand, the recent advances in natural language processing are also greatly driven by the large pre-trained language models like BERT [5] or XLNet [32]. We call the above models as unimodal models since they are pre-trained with the corpus of a particular modality, as well as carefully designed for processing the features of that modality. Over the past years, various large pre-trained multimodal models (e.g., PixelBERT [12], VisualBERT [18] or ViLT [16]) are also proposed. These models are usually called as multimodal BERT models since they are originally inspired by language BERT. The multimodal BERT models are usually fine-tuned for different downstream tasks such as visual question answer or image-text retrieval [17, 19].

3. Methodology

3.1. Problem statement

In image-text multimodal classification, each sample is associated with an image $Z_i \in \mathbb{R}^{c\times h\times w}$ and a textual description $T_i \in \mathbb{R}^{l_i\times d}$. The notations l_i and d represent the textual length and feature dimension, respectively. Both the visual and textual modalities correspond to a class label $Y_i \in \{0, 1, ..., K\}$. Denote by $\mathcal{D} = \{(Z_i, T_i, Y_i)\}_{i=1}^{N}$ the training dataset. Our goal is to learn a classifier $h(Z_i, T_i)$ which can make good predictions on Y_i by integrating information from both the visual and language modalities.

3.2. Model overview

This work adopts the Convolutional Neural Networks (CNN) and the language BERT as our study objects since they have been widely recognized as the standard baselines in their respective fields. In our approach, the pre-trained CNN and BERT are respectively expanded for image-text multimodal recognition. We call the above expanded unimodal models as Multimodal Expanded CNN and Multimodal Expanded BERT, respectively.

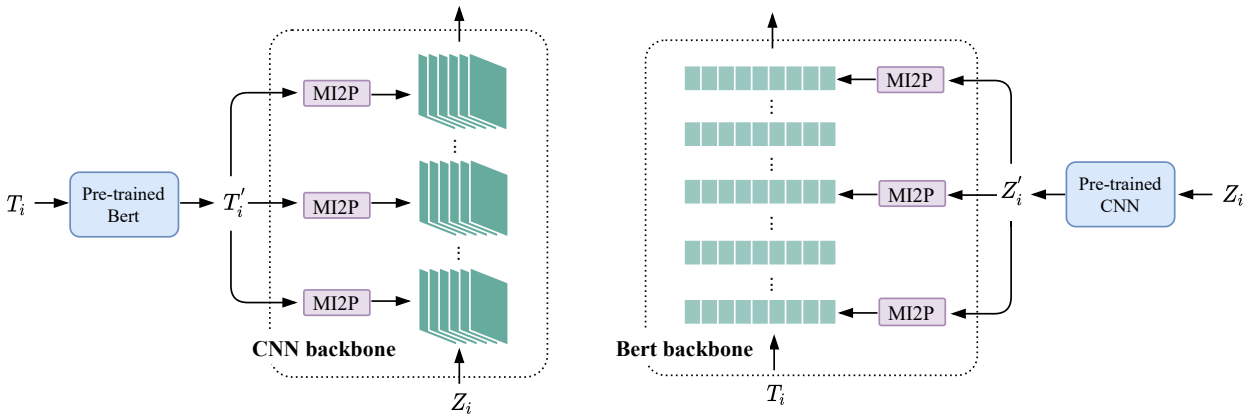
The overall architectures of the multimodal expanded models are shown in Fig. 1. To arm the CNN model with the language modality, we first pass the text features T_i through an external pre-trained BERT model and obtain the high-level representations $T_i^{'} \in \mathbb{R}^{l_i\times d}$. For each image-text pair, we integrate $T_i^{'}$ into the information flow of Z_i across the CNN backbone by the MI2P plug-ins attached to different layers of CNN (see Fig. 1(a)). Similarly, to arm the language BERT model with the image modality, we first pass the image features Z_i through an external pre-trained CNN model and obtain the high-level representations before the aggregation layer $Z_i^{'} \in \mathbb{R}^{c'\times h'\times w'}$. For each image-text pair, we integrate $Z_i^{'}$ into the information flow of T_i across the language BERT backbone by the MI2P plug-ins attached to different layers of BERT (see Fig. 1(b)). During the fine-tuning process, the unimodal models are trained jointly with the attached MI2P units. The injected multimodal information can be adapted to augment the mid-level features properly.

3.3. Multimodal information injection plug-in

In order to expand the large pre-trained unimodal models for image-text multimodal recognition, we propose the MI2P attachment which provides the path to integrate the features from other modalities to augment the mid-level features of the unimodal models. Due to the strong heterogeneity across different modalities, the features from other modalities cannot directly flow into the unimodal backbones. To bridge the modality gap, the MI2P modules perform cross-modal feature transformation based on the fine-grained cross-modal interactions between the visual and textual features.

Before going on, we need to figure out how do the mid-level features of the image Z_i and its counterpart T_i interact with each other. From the aspect of the image features, each channel of the feature maps is associated with a specific semantic pattern about the input image Z_i. In the multimodal setting, these semantic patterns are also expressed by the textual words of the language counterpart T_i. Hence, the channel-wise visual features can be closely correlated with the textual words via their shared semantic patterns.

According to the above discussion, the MI2P units need to model the cross-modal interactions between the channel-wise visual features and the word-wise textual features. Based on the modeled cross-modal interactions, the MI2P attachments will inject the language information into the unimodal CNN model by attending the word-wise textual features to different visual channels, as well as inject the visual information into the unimodal BERT model by attending the channel-wise visual features to different textual



(a) Multimodal Expanded CNN. (b) Multimodal Expanded Bert.

Figure 1. The overall architecture of the proposed approach. The information of other modalities can be integrated into the unimodal backbones via the MI2P units attached to different layers of the unimodal models. The unimodal backbones are fine-tuned jointly with the MI2P units. The parameters of the external pre-trained models displayed in blue are fixed during the training phase.

words. As plug-ins, MI2P can be flexibly attached to multiple layers of the unimodal models, in order to model the inter-modal interactions of different abstraction levels. Our approach requires minimum changes in the original network structures of the unimodal models.

Multimodal Expanded CNN. In this part, we introduce the detail how the CNN backbone is expanded for multimodal recognition. For a image-text pair consisted by Z_i and T_i, we first pass the text features T_i through an external pre-trained BERT model and obtain the high-level representations $T_i^{'} \in \mathbb{R}^{l_i \times d}$. The language features $T_i^{'}$ will then be integrated into the CNN backbone via the MI2P plug-ins attached to different layers of CNN.

Suppose a MI2P plug-in is attached at the k-th layer of the CNN backbone. Denote by $Z_i^k \in \mathbb{R}^{c^k \times h^k \times w^k}$ the image features in the k-th layer of CNN. The MI2P plug-in integrates the language features $T_i^{'}$ into the CNN backbone by attending the word-wise textual features of $T_i^{'}$ to different visual channels. To this end, we use Z_i^k to compute the query and $T_i^{'}$ to compute the key and value. Considering the spatial characteristic of the channel-wise features, we compute the query vectors $Q_i^k \in \mathbb{R}^{c^k \times d_q}$ by performing the convolution operation with d_q kernels on each channel of Z_i^k and then aggregating the feature maps via average pooling (see Fig. 2(a)). The key and value vectors are generated via linear transformation: $K_i^k = T_i^{'} W_K^k$, $V_i^k = T_i^{'} W_V^k$, where $W_K^k \in \mathbb{R}^{d \times d_k}$ and $W_V^k \in \mathbb{R}^{d \times d_v}$. One individual head of the cross-modal attention operation is formulated as follows:

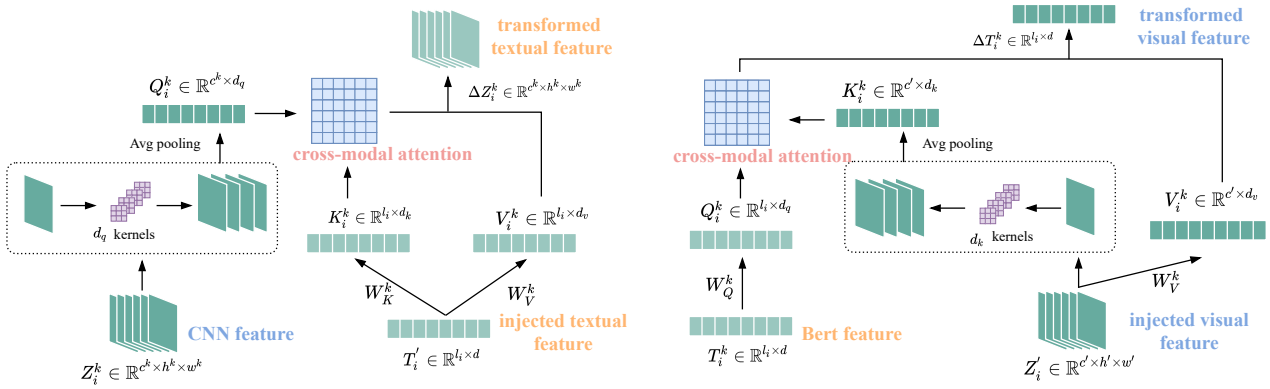
$$\begin{aligned} \Delta Z_i^k &= \mathrm{CA}_{l \to v}^k(T_i^{'}, Z_i^k) \\ &= \mathrm{softmax}(\frac{Q_i^k {K_i^k}^T}{\sqrt{d_k}}) V_i^k, \end{aligned} \tag{1}$$

where $\Delta Z_i^k \in \mathbb{R}^{c^k \times d_v}$. With h attention heads, the dimension of ΔZ_i^k will be $c^k \times h d_v$ (the values of h and d_v need to satisfy the condition $h d_v = h^k w^k$). We then reshape ΔZ_i^k as $Z_i^k \in \mathbb{R}^{c^k \times h^k \times w^k}$. ΔZ_i^k can be considered as the cross-modal transformation of $T_i^{'}$. The semantic patterns of $T_i^{'}$ are injected into different visual channels according to modeled inter-modal interactions and augment the visual content of Z_i^k in each channel: $Z_i^k = Z_i^k + \Delta Z_i^k$. We illustrate the above operations in Fig. 2(a). In order to implement inter-modal interactions between the visual and textual modalities on multiple abstraction levels (see the discussion in Section 1), the MI2P units are attached to different layers of the CNN backbone.

The CNN backbone is fine-tuned together with the attached MI2P units. The MI2P units will be trained to augment the mid-level features of the CNN backbone in a proper manner, i.e., enrich the semantic patterns of the visual channels but not suppress the intra-modal processing of the image features. Armed with the MI2P attachments, the unimodal CNN can obtain better recognition performance by integrating the language information from texts.

Multimodal Expanded BERT. In this part, we introduce the detail how the language BERT is expanded for multimodal recognition. For each image-text pair (Z_i, T_i), we first pass the image features Z_i through an external pre-trained CNN model and obtain the high-level representations $Z_i^{'} \in \mathbb{R}^{c^{'} \times h^{'} \times w^{'}}$. The image features $Z_i^{'}$ will then be integrated into the BERT backbone via the MI2P units attached to different layers of BERT.

Suppose a MI2P plug-in is attached at the k-th layer of the language BERT model. Denote by $T_i^k \in \mathbb{R}^{l_i \times d}$ the textual features in the k-th layer of BERT. The MI2P plug-in integrates the visual features $Z_i^{'}$ into the BERT backbone



(a) The MI2P units attached to the CNN model.

(b) The MI2P units attached to the Bert model.

Figure 2. The detailed operations in the MI2P attachments. (a) The MI2P attachments inject the information of $T_i^{'}$ (i.e., the language modality) into the CNN model by attending the word-wise textual features to the channel-wise visual features. (b) For the language BERT mode, the MI2P attachments inject the information of $Z_i^{'}$ (i.e., the visual modality) into the language BERT model by attending the channel-wise visual features to the word-wise textual features.

by attending the channel-wise visual features of $Z_i^{'}$ to different textual words. To this end, we use T_i^k to compute the query and $Z_i^{'}$ to compute the key and value. The operations are similar to the ones introduced in multimodal expanded CNN. In particular, we compute the key vectors $K_i^k \in \mathbb{R}^{c' \times d_k}$ by performing the convolution operation with d_k kernels on each channel of $Z_i^{'}$ and then aggregating the feature maps via average pooling (see Fig. 2(b)). The query and value vectors are generated via linear transformation: $Q_i^k = T_i^k W_K^k$, $V_i^k = \hat{Z}_i^{'} W_V^k$, where $\hat{Z}_i^{'} \in \mathbb{R}^{c' \times h'w'}$ is reshaped from $Z_i^{'}$, $W_Q^k \in \mathbb{R}^{d \times d_q}$, $W_V^k \in \mathbb{R}^{h'w' \times d_v}$. One individual head of the cross-modal attention operation is formulated as:

$$\begin{aligned} \Delta T_i^k &= \mathrm{CA}_{v \to l}^k(Z_i^{'}, T_i^k) \\ &= \mathrm{softmax}(\frac{Q_i^k {K_i^k}^T}{\sqrt{d_k}}) V_i^k \ , \end{aligned} \tag{2}$$

where $\Delta T_i^k \in \mathbb{R}^{l_i \times d_v}$. With h attention heads, the dimension of ΔT_i^k will be $l_i \times h d_v$ (the values of h and d_v need to satisfy the condition $h d_v = d$). The semantic patterns of $Z_i^{'}$ are injected into different textual words according to modeled inter-modal interactions and augment the language content of T_i^k in each word: $T_i^k = T_i^k + \Delta T_i^k$. The above operations are illustrated in Fig. 2(b). Similarly, the MI2P units are also attached to different layers of the language BERT model, with the consideration of modeling the inter-modal interactions between the visual and textual modalities on multiple abstraction levels. After fine-tuned together with the attached MI2P units, the language BERT model can be expanded to integrate the visual information of images for prediction.

3.4. Late fusion strategy

Both the multimodal expanded CNN and BERT models can perform image-text multimodal recognition independently. In order to further improve the performance, we can also conduct late fusion on top of the expanded unimodal models. Different late fusion strategies, including the commonly used score fusion (i.e., averaging the classification scores), feature concatenation (i.e., concatenating the global features) and the recent cross-attention (i.e., filtering the concatenated features via the cross-attention operation in [1]), are implemented in our experiments. The multimodal expanded CNN and BERT models are trained jointly if the late fusion strategy is adopted.

4. Experiments

4.1. Experimental setup

We conduct experiments on the standard image-text multimodal classification benchmarks, including CrisisMMD [3], Food101 [28] and MM-IMDB [20].

CrisisMMD. This benchmark focuses on detecting crisis events for emergency response based on social media posts [3]. In the dataset, each sample is associated with an image-tweet pair collected by searching hashtags in Twitter. This benchmark contains three sub-tasks. Specifically, task1 mainly focuses on recognizing whether a social media post is informative or uninformative for humanitarian aid purposes. In task2, the objective is to recognize the humanitarian categories (i.e., infrastructure damage, vehicle damage, rescue efforts, affected individuals and others) based on each image-tweet pair. In task3, the objective is to assess the severity (i.e., severe, mild, and none) of the

Table 1. The sample number in different data splits of each setting.

Setting	Training split	Validation split	Testing split
CrisisMMD (Task1)	9601	1573	1534
CrisisMMD (Task2)	2874	477	451
CrisisMMD (Task3)	2461	529	530
Food101	58131	6452	21519
MM-IMDB	15552	2608	7799

Table 2. The hyper-parameters used in each setting. The notation Crisis-T1 denotes task1 of the CrisisMMD benchmark, and so on.

	Crisis-T1	Crisis-T2	Crisis-T3	Food101	IMDB
Batch size	128	128	128	256	128
Epoch number	50	40	40	100	80
Learning rate	3.5e-5	2.5e-5	5e-5	7.5e-5	5e-5

damages reported in social media posts.

Food101. In this benchmark, each sample is associated with a recipe description scraped from web pages and a corresponding image obtained from Google Image Search [28]. The web pages have been processed into raw texts via html2text. The task is to classify each recipe-image pair from 101 food labels.

MM-IMDB. In this benchmark, each sample is associated with a movie plot outline and a corresponding movie poster [20]. The goal is to predict the movie genre based on the plot-poster pairs. Different from the above settings, this benchmark is featured as a multi-label learning task, since each movie can have multiple genres.

Table 1 shows the sample number in different data splits of each setting. The data used in our experiments do not contain personally identifiable information or offensive content.

4.2. Implementation details

We adopt DenseNet pre-trained on ImageNet [11] as the CNN backbone and the standard BERT pre-trained on BooksCorpus and English Wikipedia [5] as the language backbone. DenseNet contains five dense blocks and we attach the MI2P units to the endings of the first four blocks. The attention head of the attention operation is set to 8. BERT contains 12 transformer layers and we attach the MI2P units to all the layers. The attention head h of the attention operation is set to 12. The other important hyper-parameters are displayed in Table 2. We adopt Adam as the optimizer. The learning rate is fixed during training. The hyper-parameters are determined on the validation set. The models are trained on 24 T40 GPUs.

4.3. Performance comparison

Our proposed approach is compared to the original unimodal networks (i.e., DenseNet and language BERT), as well as the existing state-of-the-art image-text multimodal classification methods, including [1,7,12,14–16,18,20,26]. Of these, the works [1, 7, 15, 20, 26] mainly focus on performing multimodal fusion on the global features produced from each unimodal backbone; the works [12,14,16,18] are the recently proposed pre-trained multimodal BERT models which can be fine-tuned for image-text multimodal classification. The fine-grained inter-modal interactions can be modeled by the attention mechanism of the multimodal BERT models. Moreover, we also compare our approach to score fusion (i.e., averaging the classification scores of each unimodal model) and feature concatenation (i.e., concatenating the global features produced by each unimodal backbone) which are commonly used as the standard baselines for the multimodal recognition tasks.

CrisisMMD. We display the comparison on the CrisisMMD benchmark in Table 3. Since the previous works have made changes in the standard dataset, we reproduce the performance of the compared baselines for a fair comparison. In agreement with the previous work [1], we evaluate the performance by the metrics of classification accuracy, Macro F1-score and weighted F1-score.

From Table 3, we can draw the following observations. First, the unimodal models perform worse than the multimodal classification approaches. Second, our proposed MI2P units can clearly improve the performance of the unimodal models (see the performance of ME BERT and ME DenseNet). Moreover, the same late fusion strategies performed on the multimodal expanded models can obtain significantly better performance than fusing the original unimodal models when we compare Score Fusion, Feature Concat and Cross-attention with ME Score Fusion, ME Feature Concat and ME Cross-attention, respectively. The performance improvement can be attributed to the fine-grained inter-modal interactions modeled in the intermediate layers of DenseNet and BERT. Finally, we can see that the large pre-trained multimodal BERT models are sub-optimal for multimodal classification (see the performance of MMBT, VisualBERT, PixelBERT and VilT), which is consistent with our previous discussion. In general, our approach consistently outperforms the compared baselines with a large performance gain.

Food101 & MM-IMDB. We display the performance comparison on MM-IMDB and Food101 in Table 4. Similarly, we also reproduce the performance of the compared baselines. In agreement with the previous works [14, 15], we evaluate the performance by the metrics of classification accuracy in the Food101 benchmark and by the metrics of Macro F1-score and Micro F1-score in the MM-IMDB benchmark. Similar observations can be drawn as in the above setting. Our approach can consistently outperform the compared baselines.

Table 3. Comparisons on CrisisMMD in terms of classification accuracy (%), Macro F1-score (%) and weighted F1-score (%). The notation "ME DenseNet" denotes the multimodal expanded DenseNet, and so on. The notation "ME Score Fusion" denotes performing score fusion on the multimodal expanded DenseNet and BERT, and so on.

Method	Task 1			Task 2			Task 3		
	Acc	M-F1	W-F1	Acc	M-F1	W-F1	Acc	M-F1	W-F
Unimodal DenseNet	81.6	79.1	81.2	83.4	60.5	87.0	62.9	52.3	66.1
Unimodal BERT	84.9	81.2	83.3	86.1	66.8	87.8	68.2	45.0	61.1
Score Fusion	88.2	83.5	85.3	86.9	54.0	88.9	71.2	53.5	66.3
Feature Concat	87.6	85.2	86.5	89.1	65.9	90.3	68.4	43.1	55.7
Cross-attention [1]	88.4	87.6	88.7	90.0	67.8	90.2	72.9	60.1	69.7
CentralNet [26]	87.8	85.3	86.1	89.3	64.7	89.8	71.1	57.4	68.7
GMU [20]	87.2	84.6	85.7	88.7	64.3	89.1	70.6	57.1	68.2
CBP [7]	87.9	85.6	86.4	90.2	66.1	89.8	65.8	60.4	69.3
CBGP [15]	88.1	86.7	87.3	84.7	65.1	88.7	67.9	50.7	64.6
MMBT [14]	86.4	85.3	86.2	88.7	64.9	89.6	70.1	59.2	68.7
VisualBERT [18]	88.1	86.7	88.6	87.5	64.7	86.1	66.3	56.7	62.1
PixelBERT [12]	88.7	86.4	87.1	89.1	66.5	88.9	65.2	57.3	63.7
VilT [16]	87.6	85.1	88.0	86.7	61.2	87.2	67.6	58.4	65.0
ME DenseNet	89.3	89.1	88.6	90.7	75.8	91.6	71.3	61.5	72.1
ME BERT	90.3	89.8	89.3	91.4	83.2	91.7	72.1	61.4	72.6
ME Score Fusion	91.6	90.8	90.6	93.3	84.9	93.0	75.8	**63.9**	75.0
ME Feature Concat	90.8	**91.6**	90.3	92.9	85.1	93.1	74.3	62.1	74.3
ME Cross-attention	**92.0**	91.2	**91.3**	**93.5**	**85.6**	**93.6**	**76.5**	63.8	**75.7**

Table 4. Comparisons on the MM-IMDB benchmark in terms of Macro-F1 score (%) and Micro-F1 score (%) and comparisons on the Food101 benchmark in terms of classification accuracy (%).

Method	MM-IMDB		Food101
	Macro F1	Micro F1	Acc
Unimodal DenseNet	37.3	46.7	60.8
Unimodal BERT	57.9	60.7	87.9
Score Fusion	59.3	61.6	89.3
Feature Concat	59.8	61.9	89.9
Cross-attention [1]	60.4	63.8	91.3
CentralNet [26]	54.8	63.2	91.5
GMU [20]	53.9	62.7	90.6
CBP [7]	53.2	63.1	89.4
CBGP [15]	52.9	61.8	89.7
MMBT [14]	62.3	67.1	91.7
VisualBERT [18]	62.8	68.1	92.3
PixelBERT [12]	63.1	69.3	92.6
VilT [16]	63.0	68.6	92.9
ME DenseNet	61.4	66.3	90.6
ME BERT	62.6	67.5	91.9
ME Score Fusion	63.2	69.8	93.6
ME Feature Concat	63.1	70.2	**94.7**
ME Cross-attention	**64.2**	**70.8**	94.6

Table 5. Ablation study on the humanitarian categorization task of the CrisisMMD benchmark. The notation MI2P{} denotes the MI2P units attached to the corresponding dense blocks or transformer layers.

Model design	Acc(%)	M-F1(%)	W-F1(%)
DenseNet	83.4	60.5	87.0
DenseNet + MI2P{1}	88.1	67.8	88.7
DenseNet + MI2P{2}	87.3	65.3	87.1
DenseNet + MI2P{3}	87.9	66.1	88.2
DenseNet + MI2P{4}	88.4	68.2	88.6
DenseNet + MI2P{1-4}	**90.7**	**75.8**	**91.6**
BERT	86.1	66.8	87.8
Bert + MI2P{1-3}	88.1	71.4	87.9
Bert + MI2P{4-6}	87.5	71.3	87.2
Bert + MI2P{7-9}	87.6	73.4	87.3
Bert + MI2P{10-12}	88.9	74.8	89.1
Bert + MI2P{1-12}	**91.4**	**83.2**	**91.7**

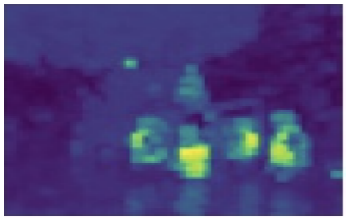
4.4. Analysis

Ablation study. Table 5 displays the ablation study on the humanitarian categorization task of the CrisisMMD benchmark. The first row displays the performance of unimodal DenseNet. In the next five rows, we attach the MI2P units to different layers of the DenseNet backbone. We can see that it is effective to inject the multimodal information into multiple layers of the unimodal backbone, which is consistent with our motivation of modeling inter-modal interactions of different abstraction levels. We also conduct the similar ablation study on the language BERT model and can draw the similar observations (see the next six rows).

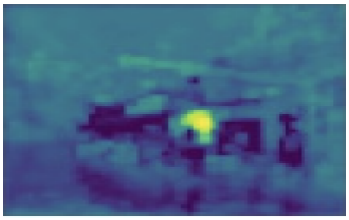
In order to verify the scalability of our proposed approach, we also conduct experiments by attaching the MI2P units to other large pre-trained neural models, i.e., ResNet-50 [10], XLNet [32], and CLIP [21]. From Table 6, we can clearly see that the proposed MI2P units again improve the performance of the unimodal models by a large margin.

Visualization. Finally, we display the visualization exam-

Thanks to the Texas National Guard for their help to rescue flooded Texans.

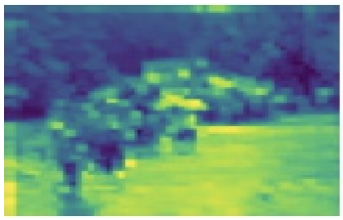


Thanks to the Texas National Guard for their help to rescue flooded Texans.

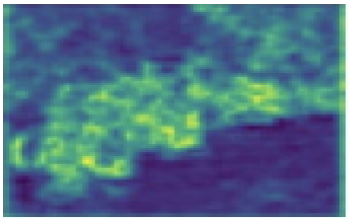


Thanks to the Texas National Guard for their help to rescue flooded Texans.

Sri Lanka Floods Update: Safe to Travel.



Sri Lanka Floods Update: Safe to Travel.



Sri Lanka Floods Update: Safe to Travel.

Figure 3. Visualization examples of the attended visual channels and the textual words in the CrisisMMD benchmark. We conduct the visualization by observing the inter-modal attention weights of the MI2P units attached to the last transformer layer of BERT.

ples of the modeled inter-modal interactions between the visual channels and the textual words in the CrisisMMD benchmark. From Fig. 3, we can see that the MI2P units can model reasonable inter-modal interactions between images and texts. The semantic patterns of the word features can be enriched by the attended channel-wise visual features.

5. Limitation and Future Work

Various large pre-trained models (e.g., Vision Transformer (ViT) [6], MLP-Mixer [24] and ConvMixer) are proposed in the recent year. We do not apply our approach to these models considering that they have not been widely recognized by the computer vision community compared with the CNN models. Moreover, some key factors why these models can be effective (e.g., what represents a visual channel in ViT) remain to be revealed. Our future works will try to expand these new models for multimodal recognition with reliable interpretability. Moreover, we will also extend our approach to the acoustic field by integrating multimodal information into the large pre-trained acoustic models (e.g., SLU [27] and acoustic Transformer [35]).

6. Conclusion

This work proposes to expand the large pre-trained unimodal models for image-text multimodal classification. To this end, we propose the MI2P plug-in which can be flexibly attached to different layers of the unimodal models. The MI2P plug-in attachments can integrate the features of other modalities into the unimodal models by modeling the fine-grained cross-modal interactions between the channel-wise visual features and the word-wise textual features. Compared with the existing methods for image-text multimodal classification, our approach can better balance the inter-modal interaction and intra-modal processing. We conduct extensive experiments on different benchmarks of image-text multimodal classification.

Table 6. The performance of other multimodal expanded pre-trained models on the humanitarian categorization task of the CrisisMMD benchmark.

Model design	Acc(%)	M-F1(%)	W-F1(%)
ResNet-50	83.9	61.4	87.6
ME ResNet-50	**91.4**	**76.3**	**91.9**
XLNet	87.8	67.4	88.3
ME XLNet	**92.1**	**84.6**	**92.8**
CLIP-Text-Encoder	86.3	61.1	86.8
ME CLIP-Text-Encoder	**90.4**	**76.2**	**90.2**
CLIP-Img-Encoder	84.3	60.3	84.1
ME CLIP-Img-Encoder	**89.6**	**74.7**	**90.1**

Acknowledgements. This work was conducted when T. Liang worked as a research assistant at Southwest Jiaotong University, supervised by F. Lv. F. Lv and T. Liang contributed equally to this work. This work was supported by the National Natural Science Foundation of China (No.62106204 and 62176221), the Fundamental Research Funds for the Central Universities of China (No.2682022CX068), and the Sichuan Science and Technology Program (No.2021YFS0178). G. Lin's participation was supported by the MOE AcRF Tier-1 research grant: RG95/20.

References

[1] Mahdi Abavisani, Liwei Wu, Shengli Hu, Joel R. Tetreault, and Alejandro Jaimes. Multimodal categorization of crisis events in social media. In *CVPR*, pages 14667–14677, 2020. 1, 2, 5, 6, 7

[2] Mansi Agarwal, Maitree Leekha, Ramit Sawhney, and Rajiv Ratn Shah. Crisis-dias: Towards multimodal damage analysis - deployment, challenges and assessment. In *AAAI*, pages 346–353, 2020. 1, 2

[3] Firoj Alam, Ferda Ofli, and Muhammad Imran. Crisismmd: Multimodal twitter datasets from natural disasters. In *Proceedings of the International Conference on Web and Social Media (ICWSM)*, pages 465–473, 2018. 5

[4] Feiyu Chen, Zhengxiao Sun, Deqiang Ouyang, Xueliang Liu, and Jie Shao. Learning what and when to drop: Adaptive multimodal and contextual dynamics for emotion recognition in conversation. In *ACM MM*, pages 1064–1073, 2021. 2

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, 2019. 3, 6

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3, 8

[7] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*, pages 457–468, 2016. 1, 2, 6, 7

[8] Ignazio Gallo, Gianmarco Ria, Nicola Landro, and Riccardo La Grassa. Image and text fusion for UPMC food-101 using BERT and cnns. In *IVCNZ*, pages 1–6, 2020. 2

[9] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification. In *ICLR*, 2021. 2

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3, 7

[11] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *CVPR*, pages 2261–2269, 2017. 3, 6

[12] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *CoRR*, abs/2004.00849, 2020. 1, 2, 3, 6, 7

[13] Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L. Iuzzolino, and Kazuhito Koishida. MMTM: multimodal transfer module for CNN fusion. In *CVPR*, pages 13286–13296, 2020. 1

[14] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. Supervised multimodal bitransformers for classifying images and text. In *Visually Grounded Interaction and Language Workshop at NeurIPS*, 2019. 2, 3, 6, 7

[15] Douwe Kiela, Edouard Grave, Armand Joulin, and Tomás Mikolov. Efficient large-scale multi-modal classification. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *AAAI*, pages 5198–5204, 2018. 1, 2, 6, 7

[16] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, volume 139, pages 5583–5594, 2021. 1, 2, 3, 6, 7

[17] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, pages 11336–11344, 2020. 1, 2, 3

[18] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557, 2019. 1, 2, 3, 6, 7

[19] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, pages 13–23, 2019. 1, 2, 3

[20] John Edison Arevalo Ovalle, Thamar Solorio, Manuel Montes-y-Gómez, and Fabio A. González. Gated multimodal units for information fusion. In *ICLR Workshop*, 2017. 5, 6, 7

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139, pages 8748–8763, 2021. 7

[22] Dhanesh Ramachandram and Graham W. Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Process. Mag.*, 34(6):96–108, 2017. 1, 2

[23] Hao Tan and Mohit Bansal. LXMERT: learning cross-modality encoder representations from transformers. In *EMNLP*, pages 5099–5110, 2019. 1

[24] Ilya O. Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. Mlp-mixer: An all-mlp architecture for vision. *CoRR*, abs/2105.01601, 2021. 3, 8

[25] Nguyen Manh Duc Tuan and Pham Quang Nhat Minh. Multimodal fusion with BERT and attention mechanism for fake news detection. *CoRR*, abs/2104.11476, 2021. 1, 2

[26] Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. Centralnet: A multilayer approach for multimodal fusion. In *ECCV Workshop*, volume 11134, pages 575–589, 2018. 6, 7

[27] Pengwei Wang, Liangchen Wei, Yong Cao, Jinghui Xie, and Zaiqing Nie. Large-scale unsupervised pre-training for end-to-end spoken language understanding. In *ICASSP*, pages 7999–8003, 2020. 8

[28] Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frédéric Precioso. Recipe recognition with large multimodal food dataset. In *ICME Workshop*, pages 1–6, 2015. 5, 6

[29] Yikai Wang, Wenbing Huang, Fuchun Sun, Tingyang Xu, Yu Rong, and Junzhou Huang. Deep multimodal fusion by channel exchanging. In *NeurIPS*, 2020. 1, 2, 3

[30] Yifan Wang, Xing Xu, Wei Yu, Ruicong Xu, Zuo Cao, and Heng Tao Shen. Combine early and late fusion together: A hybrid fusion framework for image-text matching. In *ICME*, pages 1–6, 2021. 2

[31] Xiaocui Yang, Shi Feng, Daling Wang, and Yifei Zhang. Image-text multimodal emotion classification via multi-view attentional network. *IEEE Transactions on Multimedia*, 2020. 1, 2

[32] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*, pages 5754–5764, 2019. 3, 7

[33] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *EMNLP*, pages 1103–1114, 2017. 2

[34] Guodong Zhang, Jing-Hao Xue, Pengwei Xie, Sifan Yang, and Guijin Wang. Non-local aggregation for RGB-D semantic segmentation. *IEEE Signal Process. Lett.*, 28:658–662, 2021. 1, 3

[35] Ruixiong Zhang, Haiwei Wu, Wubo Li, Dongwei Jiang, Wei Zou, and Xiangang Li. Transformer based unsupervised pretraining for acoustic representation learning. In *ICASSP*, pages 6933–6937, 2021. 8