Cross-aligned Fusion For Multimodal Understanding

Abshishek Rajora*, Shubham Gupta*, Suman Kundu
Indian Institute of Technology Jodhpur, India
{rajora.1,gupta.37,suman}@iitj.ac.in

Abstract

*Recent multimodal frameworks often grapple with semantic misalignment and noise, impeding effective integration of diverse modalities. In order to solve this problem, this study presents CaMN (**C**ross-**a**ligned **M**ultimodal **N**etwork), a framework designed to enhance multimodal understanding through a robust cross-alignment mechanism. Unlike conventional fusion methods, our framework aligns features extracted from images, text, and graphs via a tailored loss function, enabling seamless integration and exploitation of complementary information. Leveraging Abstract Meaning Representation (AMR), we extract intricate semantic structures from textual data, enriching the multimodal representation with contextual depth. Furthermore, to enhance robustness, we employ a masked autoencoder to simulate noise-independent feature space. Through comprehensive evaluation on the crisisMMD dataset, CaMN demonstrates superior performance in crisis event classification tasks, highlighting its potential in advancing multimodal understanding across diverse domains. Our code is available at https://github.com/brillard1/CaMN.*

1. Introduction

Crisis event classification from social media content is an important and challenging problem [11]. Specifically, when the information is available in various modality such as text and images. Although it is clear that multimodal classification that uses information from both visual and language cues can achieve higher efficiency [36], aligning this information in the same semantic space is difficult. Ensemble of unimodels have been used to address this issue [1,15,20,31] by learning the correlations between modalities.

However, during the fusion of modalities, in these methods [1, 15, 20, 31] a modality with less information dominates the other. The issue can be challenging in the cases where modalities are noisy, a prevalent problem when gathering data from social platforms. For example, images can be blurry, fragmented, obscured, or can contain various other types of noise. Hence, this noise not only degrades the visual quality of the features but also affects the aggregation with the other modality. Similarly, noisy text can have an impact in accurately capturing and interpreting complex semantics and logical relationships present within the text.

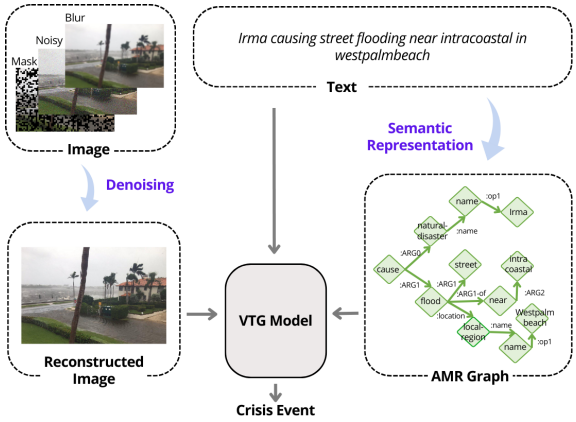

Figure 1. A Twitter example of multimodal disaster event classification using VTG (Vision-Text-Graph) model.

In order to deal with the aforementioned challenges, we propose a novel denoising multimodal framework named as CaMN (**C**ross-**a**ligned **M**ultimodal **N**etwork). This framework utilizes a semantically enhanced structure derived from text to improve the integration of different modalities for classifying crisis events. At the core of CaMN lies a masked denoising autoencoder that closely approximates the original image's latent space through improved reconstruction loss. In addition to understand the logical structure of sentences in the text, our model integrates the Abstract Meaning Representation (AMR) [5], a linguistic paradigm designed to abstractly and structurally convey the essence of natural language sentences. Proposed VTG (Vision-Text-Graph) model then extracts features from image, text, and graph modalities as shown in Figure 1, fusing them via a modality-wise guided cross-attention module. A new objective function is designed that facilitates the creation of a unified latent space where representations from text, image,

*Both authors contribute equally.

and graph coexist seamlessly. The extensive experiments on crisisMMD [3] dataset demonstrate the superior performance of the CaMN compared to existing state-of-the-art solutions. Also, generalizability of the model tested on the different domain such as fake news detection. In particular, we sum up contributions of this study as follows:

- We propose CaMN, a novel denoising based multimodal architecture that seamlessly integrates vision, text, and graph representation through a proposed guided cross attention mechanism. Noise-independent feature space has been generated by a Masked Autoencoder for visual data and Abstract Meaning Representation for complex linguistic relationships.
- We design a new objective function to align modalities within a unified semantic space, thereby improving the model's coherence for better classification.
- Our model demonstrates superior performance on the publicly available dataset, outperforming existing state-of-the-art methodologies in terms of various evaluation metrics.
- We provide a thorough analysis of our model's performance, offering insights into how different modalities interact and contribute to the decision-making process in crisis event and fake news domain.

2. Related Works

The related work is categorized into three parts. First, we discuss multimodal methodologies that integrate different modalities in the latter part of the model. We highlight the challenges faced by these methods to align semantic spaces. Then we review the literature on how various models manage noise in images and the challenges they face. Finally, an exploration of how complex relationships can be effectively extracted from text using the Abstract Meaning Representation (AMR) graph is discussed.

Multimodal learning is used to integrate information from different modalities into a single representation, enriching data representation and allowing more reliable predictions. Recent works such as image-text matching [33], sarcasm detection in memes [6], emergency response [2,20] and many more take advantage of diverse information available in both image and text forms. Integration of different modalities can be broadly categorized in three approaches: high-level, intermediate, and low-level feature fusion. In the first strategy, independent deep neural networks generate high-level features for each modality [17, 21]. Fusion occurs at the model's final layers using aggregation methods such as summation [26], tensor fusion [42], etc. Wang et al. [38] proposed an event detection model that combines low and high-level features to capitalize on their respective advantages, while a community-based unsupervised event detection model [18] is proposed that forms the interaction graph over the text modality and applies community detection algorithm to find out micro-level events. On the contrary, intermediate-level feature fusion strategies focus on fine-grained token features of image and text modality, such as multimodal BERT [25, 29]. Recent models [40, 44] uses multiple individual modality features layers to create a strong final representation but these strategies still face limitations in semantically aligning features when information in the modalities are not coherently aligned with each other.

Image based representation learning has witnessed significant advancements, particularly with the advent of autoencoder-based models. Masked image encoding methods have garnered considerable attention, offering promising avenues for robust feature extraction. Pioneering work by [9] introduced masking as a noise type in Denoising Autoencoders (DAE), demonstrating its efficacy in learning representations from corrupted images. Inspired by the success of masking methods in Natural Language Processing [13], recent approaches have adapted these principles to image processing tasks. For instance, the iGPT model [8] by OpenAI operates on sequences of pixels, predicting unknown pixels to reconstruct images effectively.Similarly, Vision Transformer (ViT) [14] explores masked patch prediction for self-supervised learning, leveraging transformers for image representation. Alongside PixelBERT [43] and VisualBERT [30], these works demonstrate the effectiveness of masking and autoencoder-based approaches in capturing visual features for diverse downstream tasks.

AMR as introduced by [5], represents relations between nodes using PropBank, frameset, and sentence vocabularies. It utilizes over a hundred semantic relations, including negation, conjunction, command, and wikification. It aims to represent different sentences with the same semantic meaning using the same AMR graph. Various NLP fields, such as summarization [28], event detection [39], question answering [32], fake news detection [19, 21] etc., have effectively used AMR. Recently, Zhang et al. [43] used AMR for the identification of out-of-context multimodal misinformation in the detection of multimodal discrepancies between visual and textual data. In order to understand the importance of semantic relations in disaster event classification, we embarked on an exploration of utilizing AMR for the same. By incorporating AMR, we enhance the capability of the detection model to identify and analyze the intricate semantic structures present in text documents.

3. Problem Definition

Given an image and associated textual content, the objective is to learn the features and categorize it into one of the predefined crisis event classes. Formally, for a given training dataset $D = \{(I_j, T_j, G(T_j), y_j)\}_{j=1}^{N}$, where image I_j,

text T_j, and graphical representation of text $G(T_j)$ represent a disaster event labeled with $y_j \in C$, the aim is to learn a objective function f such that $f : f(I_j, T_j, G(T_j)) \rightarrow y_j$. Here, C represents the set of disaster events. The goal is to improve predictive accuracy by utilizing information from multiple modalities.

4. Model Overview

Figure 2 shows the proposed CaMN. The input to the model is the data (I, T, G) and it generates a probability distribution of labels across various classes of crisis events using proposed modality-wise guided cross-attention and cross-alignment loss. Before detailing out the proposed moadlity-wise guided cross attention in Section 4.4 and cross-alignment loss in Section 4.5, we outline the noise independent features extraction used in the architecture below.

4.1. Feature Extraction from Image

This module is pivotal in distilling rich and robust representations from input images, effectively capturing their intrinsic visual characteristics in the form of latent features. We have used Masked Autoencoder (MAE) [22], a state-of-the-art model pioneered by Meta's Fundamental AI Research for the purpose. Built upon the transformer architecture, MAE excels in learning hierarchical features from images. At its core, the MAE encoder comprises a series of transformer blocks, forming the backbone ViT (Vision Transformer). The input image is masked into patches. Each transformer block sequentially refines the features extracted from these patches, thereby encapsulating both local and global information.

Encoding: The encoding process begins by embedding image patches into a high-dimensional space, represented as $Z_0 = \text{PatchEmbed}(I)$ where input images, sized 228, is divided into patches of size 14. From there, each subsequent transformer block Block_i iteratively refines the feature representation. Let us denote Z_i as the output of the i-th transformer block, following the relation:

$$Z_i = \text{Block}_i(Z_{i-1}), \quad \text{for } i = 1, 2, \ldots, \text{depth} \qquad (1)$$

with the final output of the encoder denoted as $\mathcal{Z}_I = Z_{i=\text{depth}}$. Increasing the encoder depth, which is set to 32, enhances its ability to capture both local and global patterns, allowing it to extract more comprehensive hierarchical features from the image patches, as discussed for asymetrical model design in [22].

Masking: During training, the MAE encoder employs masking, wherein a fraction of input patches is randomly masked out. This mechanism compels the model to reconstruct missing portions of the image from available context, fostering robust feature learning. With a masking ratio typically set to 75%, the encoder learns to accommodate partial observations, mirroring real-world scenarios where complete information may be lacking. Self-supervised learning underpins the training regime of MAE, harnessing salient features extracted by the encoder. Reconstructing the original image from partial encodings and mask tokens, the model discerns informative image regions, thereby enriching the learned representations. This process enables adaptability to noisy or incomplete images, contributing to the robustness of extracted features.

Fine-tuning: In order to harness the latent features gleaned by the MAE encoder, a subsequent fine-tuning phase is imperative. Initially, the encoder is pre-trained on a varied dataset. Then, fine-tuning orients the encoder for the specific task, which in our case is crisis events classification. This process capitalizes on the enriched visual features residing in the latent space, thus enhancing the performance of classification tasks.

The MAE is built upon asymmetrical design where the encoder outweighs the decoder in terms of computational complexity. Given that the decoder demands less than 10% computation per token compared to the encoder, an alternative approach could involve utilizing the decoder during the fine-tuning phase. However, we deduce that the use of the encoder alone suffices for our intended purpose.

4.2. Feature Extraction from Text

The objective of the text encoder is to generate the representation, denoted as $\mathcal{Z}_T$, of text content. We use ELECTRA [12] model for the same. The rationale behind selecting the ELECTRA is that it is more parameter efficient and faster to train than other transformer-based models. This is attributed to its substituted token detection objective and the generator-discriminator framework. Moreover, through positional encoding, it effectively captures sequence information and possesses robust capabilities in extracting semantic features.

Given a textual input denoted as $T = \{w_1, w_2, \cdots, w_n\}$, consisting of a sequence of n tokens related to crisis information, were tokenized and fed into the ELECTRA model to obtain the final layer embedding h^i as follows:

$$\mathcal{Z}_T = h_T^0; \; [h_T^0, h_T^1, \ldots, h_T^n] = \text{ELECTRA}(T) \qquad (2)$$

where, h_T^i represents the ELECTRA embedding for the i-th token. $\mathcal{Z}_T$, representing the embedding of the "[CLS]" token, is seen as the initial textual representation for T. $\mathcal{Z}_T \in \mathbb{R}^{N \times L \times D}$, where N, L, and D denote the batch size,

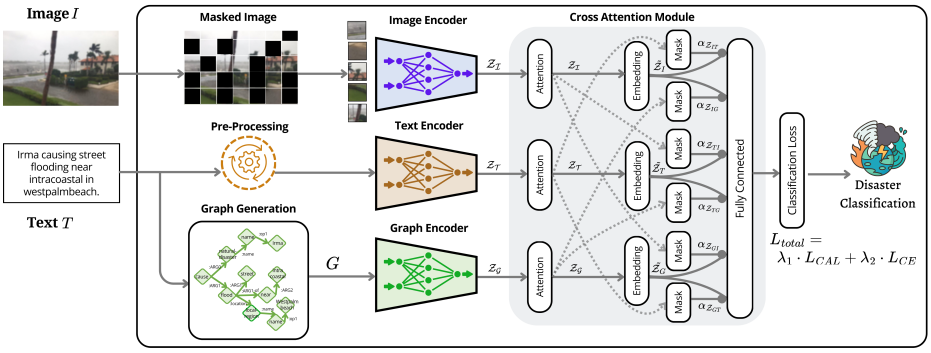

Figure 2. The overall architecture of CaMN.

maximum sequence length, and dimension of the feature vector, respectively.

4.3. Feature Extraction from Graph

The Graph Encoder module leverages Abstract Meaning Representation (AMR) [5] to encode the semantics of textual information T into a graph structure. This graph encoder consists of two fundamental elements spanning from AMR generation to path-aware graph learning.

4.3.1 AMR Generation

The generation process converts the text into a network of nodes and edges, capturing the relationships between different entities. AMR generation process involves parsing the sentences to extract linguistic information, including semantic roles, relations, and core events. For a text T, we represent the AMR graph as $\mathcal{G}^{amr} = (\mathcal{V}^{amr}, \mathcal{E}^{amr})$ where $\mathcal{V}^{amr}$ signifies a set of node entities, and $\mathcal{E}^{amr}$ denotes relation edges. As an illustrative example, consider the sentence: "*Irma causing street flooding near intracoastal in westpalmbeach.*" The corresponding AMR graph is presented below.

$$
\begin{aligned}
&(c \,/\, cause-01\\
&\quad :arg0\ (n \,/\, natural-disaster\\
&\qquad :name\ (n2 \,/\, name :op1\ \text{“}Irma\text{”}))\\
&\quad :arg1\ (f \,/\, flood-01\\
&\qquad :arg1\ (s \,/\, street)\\
&\qquad :arg1-of\ (n3 \,/\, near-02\\
&\qquad\quad :arg2\ (ii \,/\, intracoastal))\\
&\qquad :location\ (l \,/\, local-region\\
&\qquad\quad :name\ (n4 \,/\, name :op1 \,/\, Westpalmbeach))
\end{aligned}
$$

The AMR graph is a directed acyclic graph that represents a hierarchical structure with nodes denoting entities ($Irma$, $intracoastal$, $Westpalmbeach$, etc). Edges ($arg0, arg1, name$, etc) capture the relationships between these entities, forming a semantically structured representation of T.

4.3.2 Graph Learning with Path Optimization

This module assumes a crucial role in extracting informative features from the obtained AMR graph. In order to gain a deeper understanding of textual data, these features encapsulate critical semantic relationships. The main part of the module is a Graph Transformer [7], which uses different ways of paying attention to process the graph representation. This helps the model think and learn about the text more effectively.

The graph obtained earlier is sent to the node initialization and relation encoder to convert the AMR into a format represented in $\mathbb{R}^{N\times L\times D'}$, where D' is the dimension of the graph encoding.

To help the model understand specific paths in the graph from $\mathcal{G}^{AMR}$, the relation encoder is used to find the shortest path between two entities. This sequence representing the path is then turned into a relation vector using a bi-directional Gated Recurrent Unit (GRU) based RNN [10]. The mathematical representtaion for the encoding is:

$$\overrightarrow{p}_t = \text{GRU}_f(\overrightarrow{p}_{t-1}, sp_t)\ \overleftarrow{p}_t = \text{GRU}_g(\overleftarrow{p}_{t+1}, sp_t)$$

In this context, sp_t represents the shortest path of the relation between the two entities. The last hidden states of the forward GRU network and the backward GRU networks are concatenated to form the final relation encoding r_{ij}. To compute the attention score, r_{ij} is split into two distinct encodings: $r_{i\to j}$ and $r_{j\to i}$ using a linear layer and the parameter matrix W_r.

$$r_{ij} = [\overrightarrow{p}_n; \overleftarrow{p}_0],\ \ [r_{i\to j}; r_{j\to i}] = W_r r_{ij}$$

After that, the attention scores are computed on both the entity and relation representation present in $\mathcal{G}^{AMR}$ and then Graph Transformer (GT) encodes the AMR representations $\mathcal{G}^{AMR}$ as follows:

$$\mathcal{Z}_G = \text{GT}(\mathcal{G}^{AMR}) \in \mathbb{R}^{N\times L\times D'} \tag{3}$$

Where $\mathcal{Z}_G$ represents the final graph embedding obtained from the Graph Transformer, and D' is the dimension of the feature vector.

4.4. Modality wise Guided Cross Attention-based Fusion

Once we obtain the individual feature maps for the image ($\mathcal{Z}_I$), text ($\mathcal{Z}_T$), and graph ($\mathcal{Z}_G$), we implement guided cross-attention across all modalities. This is done to minimize semantic inconsistencies during training that can adversely impact the overall performance of the network. To achieve this, we formed the guided cross-attention module for the image, text, and graph modalities as described below:

$$Attn(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{4}$$

In this context, Q, K, and V represent the query, key, and value respectively, while $\sqrt{d}$ serves as a normalization factor. The new representations $\mathcal{Z}_I$, $\mathcal{Z}_T$, and $\mathcal{Z}_G$ are then derived as follows:

$$\begin{aligned} &\overline{\mathcal{Z}_I} = Attn(\mathcal{Z}_I), \quad \overline{\mathcal{Z}_T} = Attn(\mathcal{Z}_T), \\ &\overline{\mathcal{Z}_G} = Attn(\mathcal{Z}_G) \end{aligned} \tag{5}$$

Now, project new representation of image $\overline{\mathcal{Z}_I}$, text $\overline{\mathcal{Z}_T}$, and graph $\overline{\mathcal{Z}_G}$ into a fixed dimensionality K in following manner:

$$\begin{aligned} &\tilde{\mathcal{Z}}_I = F\left(W_I^{\mathcal{T}}\overline{\mathcal{Z}_I} + b_I\right), \quad \tilde{\mathcal{Z}}_T = F\left(W_T^{\mathcal{T}}\overline{\mathcal{Z}_T} + b_T\right), \\ &\tilde{\mathcal{Z}}_G = F\left(W_G^{\mathcal{T}}\overline{\mathcal{Z}_G} + b_G\right) \end{aligned} \tag{6}$$

where F represents an activation function such as ReLU and $\tilde{\mathcal{Z}}_I$, $\tilde{\mathcal{Z}}_T$, and $\tilde{\mathcal{Z}}_G$ are of dimension K and K is fixed to 100.

In order to apply attention across modalities, attention masks on $\overline{\mathcal{Z}_I}$, $\overline{\mathcal{Z}_T}$, and $\overline{\mathcal{Z}_G}$ are calculated as follows:

$$\begin{aligned} &\alpha_{\mathcal{Z}_{IT}} = \sigma\left(W_I'^{\mathcal{T}}\overline{\mathcal{Z}_T} + b_I'\right), \alpha_{\mathcal{Z}_{IG}} = \sigma\left(W_I'^{\mathcal{T}}\overline{\mathcal{Z}_G} + b_I'\right), \\ &\alpha_{\mathcal{Z}_{TI}} = \sigma\left(W_T'^{\mathcal{T}}\overline{\mathcal{Z}_I} + b_T'\right), \alpha_{\mathcal{Z}_{TG}} = \sigma\left(W_T'^{\mathcal{T}}\overline{\mathcal{Z}_G} + b_T'\right), \\ &\alpha_{\mathcal{Z}_{GI}} = \sigma\left(W_G'^{\mathcal{T}}\overline{\mathcal{Z}_I} + b_G'\right), \alpha_{\mathcal{Z}_{GT}} = \sigma\left(W_G'^{\mathcal{T}}\overline{\mathcal{Z}_T} + b_G'\right) \end{aligned} \tag{7}$$

where σ denotes the Sigmoid function. The attention mask $\alpha_{\mathcal{Z}_{IT}}$ for the image relies entirely on the text embedding $\overline{\mathcal{Z}_T}$, while the attention mask $\alpha_{\mathcal{Z}_{IG}}$ for the image depends solely on the graph embedding $\overline{\mathcal{Z}_G}$. Similar relationships can be derived for the other modalities. Once we obtain the attention masks for the image, text, and graph, we enhance the projected image, text, and graph embeddings $\tilde{\mathcal{Z}}_I$, $\tilde{\mathcal{Z}}_T$, and $\tilde{\mathcal{Z}}_G$ by performing element-wise multiplication as follows:

$$\begin{aligned} &\alpha_{\mathcal{Z}_{IT}} \cdot \tilde{\mathcal{Z}}_I, \alpha_{\mathcal{Z}_{IG}} \cdot \tilde{\mathcal{Z}}_I, \quad \alpha_{\mathcal{Z}_{TI}} \cdot \tilde{\mathcal{Z}}_T, \alpha_{\mathcal{Z}_{TG}} \cdot \tilde{\mathcal{Z}}_T, \\ &\alpha_{\mathcal{Z}_{GI}} \cdot \tilde{\mathcal{Z}}_G, \alpha_{\mathcal{Z}_{GT}} \cdot \tilde{\mathcal{Z}}_G \end{aligned} \tag{8}$$

Last phase of this module involves processing the combined embedding, which represents the image, text, and graph pair, through a fully-connected network. The classification is then performed using the proposed cross-alignment loss method described in the following section.

$$\begin{aligned} \mathcal{Z} = Concat([&\alpha_{\mathcal{Z}_{IT}} \cdot \tilde{\mathcal{Z}}_I, \alpha_{\mathcal{Z}_{IG}} \cdot \tilde{\mathcal{Z}}_I, \alpha_{\mathcal{Z}_{TI}} \cdot \tilde{\mathcal{Z}}_T, \\ &\alpha_{\mathcal{Z}_{TG}} \cdot \tilde{\mathcal{Z}}_T, \alpha_{\mathcal{Z}_{GI}} \cdot \tilde{\mathcal{Z}}_G, \alpha_{\mathcal{Z}_{GT}} \cdot \tilde{\mathcal{Z}}_G]) \end{aligned} \tag{9}$$

4.5. Cross-Alignment Loss

The aggregate Loss (L_{total}) used in the model is summation of Cross-Alignment Loss (L_{CAL}) and Cross-Entropy loss (L_{CE}).

The Cross-Alignment Loss L_{CAL} is computed as the average cosine embedding loss across all modality pairs (p) of attention masks ($\alpha_{\mathcal{Z}_i}$) and linear projections ($\tilde{\mathcal{Z}}_i$):

$$L_{CAL} = \frac{1}{p}\sum_{i=1}^{p}\left(1 - \frac{\alpha_{\mathcal{Z}_i} \cdot \tilde{\mathcal{Z}}_i}{\|\alpha_{\mathcal{Z}_i}\| \cdot \|\tilde{\mathcal{Z}}_i\|}\right) \tag{10}$$

where $p = \frac{M!}{(M-2)!}$, M denotes number of modalities used as an input for the model and M is 3 in our case.

Second, L_{CE} is calculated by using the softmax probabilities passed through fully connected layer:

$$f(\mathcal{Z}) = softmax(\text{FC}(\mathcal{Z})) \in \mathbb{R}^{N \times n}, y_{\text{pred}} = argmax(f(\mathcal{Z})) \tag{11}$$

where n is the total number of crisis event classes.

$$L_{CE} = -\sum_{i=1}^{N} y_{\{\text{true},i\}} \log(f(\mathcal{Z})_i) \tag{12}$$

By combining L_{CAL} with L_{CE}, the overall loss function is formulated as:

$$L_{total} = \lambda_1 \cdot L_{CAL} + \lambda_2 \cdot L_{CE} \tag{13}$$

The L_{CAL} serves as a regularization term, encouraging alignment and integration of features across modalities. This loss term measures the cosine distance between the attention mask and linear projection vectors, representing the degree of alignment between modalities. Ideally, for a better prediction, the attention masks ($\alpha_{\mathcal{Z}_i}$ terms) and linear projections ($\tilde{\mathcal{Z}}_i$ terms) should align closely, ensuring that relevant features from each modality contribute effectively to the prediction task. Deviations from this ideal alignment are penalized by L_{CAL}, encouraging the model to learn cohesive representations.

The overall loss function is formulated as a weighted combination of L_{CAL} and L_{CE}, where hyperparameters λ_1 and λ_2 control the relative importance of each loss term. By balancing both contributions, the model optimizes both prediction accuracy and multimodal alignment, thereby enhancing its performance across a range of tasks.

Table 1. Comparative analysis of Setting A and B presented in the terms of Accuracy (Acc)%, Macro F1-score (M-F1)%, Weighted F1-score (W-F1)% and Multi-task Model Strength (MTMS)%.

Method	Task 1			Task 2			Task 3			MTMS
	Acc	M-F1	W-F1	Acc	M-F1	W-F1	Acc	M-F1	W-F1	
Setting A										
DenseNet [23]	81.6	79.1	81.2	83.4	60.5	87.0	62.9	52.3	66.1	76.9
MAE [22]	84.9	82.1	84.6	89.3	64.2	89.3	67.4	55.2	67.1	81.9
ELECTRA [12]	85.8	82.9	85.3	87.1	59.4	86.4	60.2	47.3	57.4	78.7
MMBT [25]	86.4	85.3	86.2	88.7	64.9	89.6	70.1	59.2	68.7	82.7
GMU [4]	87.2	84.6	85.7	88.7	64.3	89.1	70.6	57.1	68.2	82.9
VilT [27]	87.6	85.1	88.0	86.7	61.2	87.2	67.6	58.4	65.0	81.2
CentralNet [37]	87.8	85.3	86.1	89.3	64.7	89.8	71.1	57.4	68.7	83.5
CBGP [26]	88.1	86.7	87.3	84.7	65.1	88.7	67.9	50.7	64.6	80.3
VisualBERT [30]	88.1	86.7	88.6	87.5	64.7	86.1	66.3	56.7	62.1	81.3
ViLBERT [34]	88.4	86.5	88.7	88.2	65.1	86.6	65.9	56.3	61.8	81.6
TinyCLIP [41]	84.2	81.1	83.7	86.7	59.5	86.4	64.8	41.2	56.0	79.6
PixelBERT [24]	88.7	86.4	87.1	89.1	66.5	88.9	65.2	57.3	63.7	81.8
Cross-attention [1]	88.4	87.6	88.7	90.0	67.8	90.2	72.9	60.1	69.7	84.5
UniS-MMC [44]	90.9	89.6	90.2	88.7	**68.1**	88.6	70.7	58.1	69.5	83.7
CaMN(Ours)	**92.8**	**91.3**	**92.7**	**92.4**	67.5	**92.2**	**73.2**	**60.7**	**71.1**	**86.7**
Setting B										
DenseNet [23]	84.4	82.8	84.6	74.8	60.7	79.9	-	-	-	77.5
MAE [22]	86.7	84.2	86.4	77.1	62.5	82.3	-	-	-	79.8
ELECTRA [12]	83.3	80.5	82.2	81.8	57.6	81.1	-	-	-	82.2
Cross Attention [1]	85.6	82.3	84.8	89.3	63.4	89.8	-	-	-	88.2
UniS-MMC [44]	86.3	84.7	86.3	84.1	66.5	84.1	-	-	-	84.7
CaMN(ours)	**87.6**	**85.2**	**87.3**	**90.4**	**66.7**	**90.1**	-	-	-	**89.6**

5. Experiments and Results

In order to evaluate the efficacy of CaMN, extensive experiments are conducted using the CrisisMMD dataset [3]. This section presents an overview of the results and ablation study. The detailed experimental setup and dataset settings are available in supplementary document.

5.1. Results

We assess our proposed methodology against multiple state-of-the-art approaches, both unimodal and multimodal. Specifically, we compare our method to unimodal networks like DenseNet and MAE for images, and the language based model such as ELECTRA for textual analysis, across a range of tasks. Additionally, we consider a second category comprising existing image-text multimodal classification methods referenced in several studies [1, 4, 16, 20, 24–27, 30, 37]. Some of these methods [1, 30, 37, 44] focus on multimodal fusion utilizing global features derived from each modality's backbone, while some apply compact bilinear pooling for fusion [16, 26]. Recently, CrisisKAN [20] uses external knowledge to do the crisis event classification. Because of external knowledge limitation, we have not included CrisisKAN in our comparative study. Quantitative results for the mentioned baselines are provided in Table 1 for two different settings, Setting A and Setting B. Our evaluation within the present dataset configuration, measures the effectiveness of each method.

According to Table 1, all multimodal methods outperform the unimodal models, highlighting the advantage of multimodal learning. Notably, our CaMN model shows an improvement of approximately 4-7% in Task 1, 3-8% in Task 2, and 1-8% in Task 3 under Setting A compared to multimodal baselines. Our model's Multi-Task Model Strength (MTMS) is also quantified across the three tasks, achieving a high score of 86.7%. It is also evident that MAE outperforms the Densenet unimodal, demonstrating that MAE generates a better feature representation.

Further insights from Table 1 under Setting B, where image and text pairs are inconsistently labeled for the same event, indicate that CaMN outperforms both unimodal and multimodal baselines. These findings confirm that our model effectively integrates textual, visual, and graphical features, making it a robust solution for diverse learning environments.

5.2. Ablation Study

For subsequent experiments using the CaMN model, we have conducted analyses on both Task 1 and 2 for the Setting A. The following subsections will provide detailed descriptions of each study.

5.2.1 Effect of different noises on model

To comprehensively assess the resilience of our model against various types of noise, we used three distinct noise patterns such as Uniform, Gaussian, and Masked into the dataset. Specifically, we modified the original images by

incorporating Uniform noise (set at level 80), Gaussian noise (with kernel size 15 and sigma equals to 5), and 50% Masked noise. Our experiments involved testing the CaMN model alongside various established unimodal and multimodal techniques to gauge its performance. For this comparative analysis, we selected the Cross-attention [1] and UniS-MMC [44] models, which are recognized for its superior performance in multimodal networks, particularly in terms of accuracy and F1-score, as presented in Table 1. According to the results presented in Table 2, it is clear that the CaMN model significantly outperforms other models on Tasks 1 and 2, demonstrating a notable margin of improvement. Further observation revealed that the MAE model yielded results that were on par with those of other multimodal networks. This finding supports the hypothesis that the masking-based encoder effectively shapes the latent representation, thereby enhancing the model's robustness. This enhanced representation likely contributes to the model's improved ability to handle diverse and challenging noise conditions.

Table 2. Effect of model's robustness on different noises.

Model	Uniform		Gaussian		Masked	
	Acc	W-F1	Acc	W-F1	Acc	W-F1
Task 1						
DenseNet [23]	77.6	78.2	77.9	78.4	74.5	73.9
MAE [22]	84.9	84.9	82.9	82.8	79.6	79.5
Cross-attention [1]	84.9	84.6	84.7	84.5	83.6	83.5
UniS-MMC [44]	86.4	86.3	87.1	86.8	87.4	87.3
CaMN (ours)	**90.9**	**90.8**	**91.4**	**91.5**	**90.3**	**90.1**
Task 2						
DenseNet [23]	79.3	80.4	77.5	77.8	76.3	77.2
MAE [22]	87.1	87.9	84.5	85.1	79.3	78.8
Cross-attention [1]	86.1	86.8	85.2	84.8	83.2	83.4
UniS-MMC [44]	87.2	86.9	86.9	86.5	85.4	85.3
CaMN (ours)	**90.3**	**91.3**	**89.4**	**89.9**	**90.2**	**89.8**

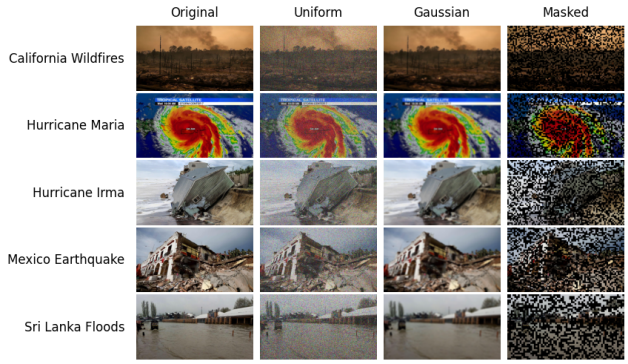

Figure 3. Curated samples and their noise correspondences for each disaster case. Here atleast two noise samples get misclassified by previous state-of-the-art.

To further elucidate the robustness of the model under various noise conditions, Figure 3 presents curated samples from disaster events along with their corresponding noise-modified versions. Each disaster case highlights instances where at least two noise conditions led to misclassifications by previous state-of-the-art models. For example, in the Sri Lanka Floods case, an image of a flood scene, which typically shows clear water levels and debris, is subjected to Gaussian noise, significantly blurring the details and leading to misclassification by other models. In contrast, CaMN maintains high accuracy by effectively leveraging its noise-independent feature space to recognize key visual elements. Similarly, the California Wildfires case, featuring a wildfire, demonstrates CaMN's ability to discern the fire's edge and smoke density despite heavy masking noise, where other models fail. These case studies highlight the model's effective denoising ability to interpret critical features under adverse conditions, substantiating its enhanced performance across varied disaster scenarios and noise types.

5.2.2 Effect of λ_1 and λ_2 on model

In order to calculate the total loss, L_{total}, in our model CaMN, we calculate it as a weighted sum of two component losses: L_{CAL} (cross-alignemnt loss) and L_{CE} (cross-entropy loss). The weights for these losses are denoted by λ_1 and λ_2, respectively.

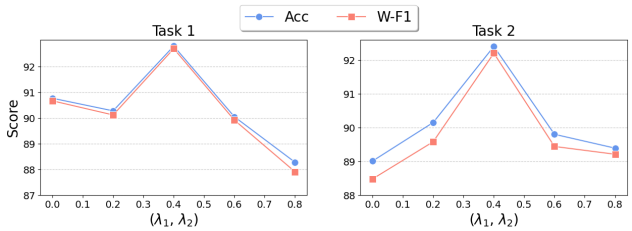

Figure 4. Hyper-parameter tuning of λ_1 and λ_2.

In order to identify the most effective combination of these weights, we conducted a series of experiments on the CaMN model, varying λ_1 and λ_2 from 0.0 to 0.8. The results, depicted in Figure 4, show that the graph follows a bell-shaped curve and optimal performance is achieved when $\lambda_1 = 0.4$ and $\lambda_2 = 0.6$. This specific combination of λ_1 and λ_2 leads to the best balance between the cross-alignment and the cross-entropy losses, enhancing CaMN's overall performance.

5.2.3 Interpreting cross-alignment

We calculate attention weights for the curated samples, in the original setup to understand the decision-making process. The attention map scores depicted in Figure 5 provide insight into the interplay of attention scores across different

Table 3. Comparison on different CaMN variants.

CaMN			Task 1			Task 2			Task 3			MTMS
IE	TE	GE	Acc	M-F1	W-F1	Acc	M-F1	W-F1	Acc	M-F1	W-F1	
✓	✗	✗	84.9	82.1	84.6	89.3	64.2	89.3	67.4	55.2	67.1	81.9
✗	✓	✗	85.8	82.9	85.3	87.1	59.4	86.4	60.2	47.3	57.4	78.7
✗	✗	✓	86.2	84.5	86.8	89.5	64.7	89.1	68.6	57.8	67.9	82.6
✓	✓	✗	88.9	87.9	89.1	90.5	65.9	90.4	72.6	59.4	69.3	84.8
✗	✓	✓	89.6	88.5	89.2	91.1	66.4	90.9	72.7	59.9	69.1	85.3
✓	✗	✓	91.2	90.6	90.9	91.8	66.9	91.7	73.0	60.4	70.6	86.1
✓	✓	✓	**92.8**	**91.3**	**92.7**	**92.4**	**67.5**	**92.2**	**73.2**	**60.7**	**71.1**	**86.7**

modality pairs within CaMN, highlighting the effectiveness of our cross-modal feature alignment. In the figure, Text-Image pair values range from 0-200, Text-AMR values are depicted from 200-400 and remaining represents the Image-AMR pair.

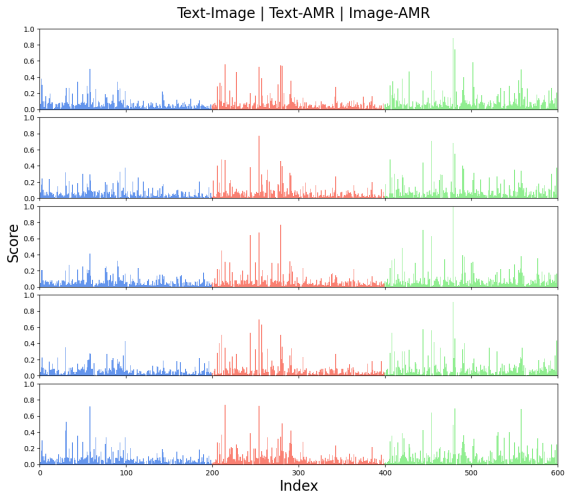

Figure 5. Projected attention weights for the curated samples.

The Image-AMR pair achieves the highest attention scores, indicating a robust integration of visual features from MAE with semantic structures extracted from the AMR via the graph transformer.

5.2.4 Comparison on CaMN Variants

We conducted experiments with our model CaMN by excluding or including difference encoders such as Image Encoder (IE), Text Encoder (TE), and Graph Encoder (GE). Our findings from Table 3 indicate that image and graph encoder pair to be more effective than other pairs. The choice of text and image only model produces lower accurate results due to presence of noise in individual modality. Once we integrate all the encoders into our final model, there is a enhancement of 1-2% over the image and graph pair model.

5.2.5 Generalisation study on CaMN

In order to generalize our methodology, we test our model's reliability on different domains such as Fake News Detection. For the experiment, we use Fakeddit dataset [35] for the binary classification task. The dataset settings is a binary classification module where 342 true and 464 fake instances are randomly sampled from the multimodal data. The results from Table 4 shows that CaMN outperforms other models in terms of accuracy and F1-score with an improvement over 3-4%.

Table 4. Evaluation scores on Fakeddit dataset.

Model	Acc	M-F1	W-F1
VisualBERT [30]	76.9	74.7	76.2
ViLBERT [34]	77.1	74.3	76.1
TinyCLIP [37]	75.8	73.9	74.8
Cross-attention [1]	77.5	76.3	77.2
UniS-MMC [44]	78.3	77.9	78.1
CaMN (ours)	**82.8**	**79.6**	**81.7**

6. Conclusion

In our research, we introduce CaMN, a novel Cross-Aligned Multimodal Network. Leveraging a masking mechanism-based encoder, this network effectively filters noise from images and improves reconstruction quality. To capture the semantic essence of text, we employ abstract meaning representation. Additionally, to enhance communication across diverse modalities and filter out irrelevant or misleading data, we introduce a guided cross-attention module. This module significantly reduces the semantic gap between modalities, allowing for selective fusion of valuable information across all modalities. This research proposes a classification loss which is the weighted summation of cross-alignment loss and cross entropy loss. Cross-alignment loss works as a regularization term which aligns features across different modalities to enhance consistency. The methodology can be further improved by integrating lengthy textual inputs into the designed model. Moreover, an explainability module can be integrated for visualization.

References

[1] Mahdi Abavisani, Liwei Wu, Shengli Hu, Joel Tetreault, and Alejandro Jaimes. Multimodal categorization of crisis events in social media. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14679–14689, 2020. 1, 6, 7, 8

[2] Mansi Agarwal, Maitree Leekha, Ramit Sawhney, and Rajiv Ratn Shah. Crisis-dias: Towards multimodal damage analysis - deployment, challenges and assessment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):346–353, Apr. 2020. 2

[3] Firoj Alam, Ferda Ofli, and Muhammad Imran. Crisismmd: Multimodal twitter datasets from natural disasters. In *Proceedings of the international AAAI conference on web and social media*, volume 12, 2018. 2, 6

[4] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017. 6

[5] Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kev Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria, Aug. 2013. 1, 2, 4

[6] Dibyanayan Bandyopadhyay, Gitanjali Kumari, Asif Ekbal, Santanu Pal, Arindam Chatterjee, and Vinutha BN. A knowledge infusion based multitasking system for sarcasm detection in meme. In *European Conference on Information Retrieval*, pages 101–117. Springer, 2023. 2

[7] Deng Cai and Wai Lam. Graph transformer for graph-to-sequence learning. In *AAAI*, pages 7464–7471. AAAI Press, 2020. 4

[8] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1691–1703. PMLR, 13–18 Jul 2020. 2

[9] Xinlei Chen, Zhuang Liu, Saining Xie, and Kaiming He. Deconstructing denoising diffusion models for self-supervised learning, 2024. 2

[10] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *EMNLP*, pages 1724–1734, Doha, Qatar, Oct. 2014. ACL. 4

[11] Shatadru Roy Chowdhury, Srinka Basu, and Ujjwal Maulik. A survey on event and subevent detection from microblog data towards crisis management. *International Journal of Data Science and Analytics*, 14(4):319–349, 2022. 1

[12] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: pre-training text encoders as discriminators rather than generators. In *ICLR*. OpenReview.net, 2020. 3, 6

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. 2

[14] 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, 2021. 2

[15] 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 Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 457–468, Austin, Texas, Nov. 2016. Association for Computational Linguistics. 1

[16] 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. *arXiv preprint arXiv:1606.01847*, 2016. 6

[17] Ignazio Gallo, Gianmarco Ria, Nicola Landro, and Riccardo La Grassa. Image and text fusion for upmc food-101 using bert and cnns. In *2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pages 1–6. IEEE, 2020. 2

[18] Shubham Gupta and Suman Kundu. Interaction graph, topical communities, and efficient local event detection from social streams. *Expert Systems with Applications*, 232:120890, 2023. 2

[19] Shubham Gupta, Abhishek Rajora, and Suman Kundu. EA2n: Evidence-based AMR attention network for fake news detection, 2024. 2

[20] Shubham Gupta, Nandini Saini, Suman Kundu, and Debasis Das. Crisiskan: Knowledge-infused and explainable multimodal attention network for crisis

event classification. In *Advances in Information Retrieval*, pages 18–33, Cham, 2024. Springer Nature Switzerland. 1, 2, 6

[21] Shubham Gupta, Narendra Yadav, Suman Kundu, and Sainathreddy Sankepally. Fakedamr: Fake news detection using abstract meaning representation network. In *International Conference on Complex Networks and Their Applications*, pages 308–319. Springer, 2023. 2

[22] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021. 3, 6, 7

[23] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017. 6, 7

[24] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020. 6

[25] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*, 2019. 2, 6

[26] Douwe Kiela, Edouard Grave, Armand Joulin, and Tomas Mikolov. Efficient large-scale multi-modal classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 2, 6

[27] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 6

[28] Panagiotis Kouris, Georgios Alexandridis, and Andreas Stafylopatis. Text summarization based on semantic graphs: An abstract meaning representation graph-to-text deep learning approach, 2022. 2

[29] 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 *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11336–11344, 2020. 2

[30] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2, 6, 8

[31] Tao Liang, Guosheng Lin, Mingyang Wan, Tianrui Li, Guojun Ma, and Fengmao Lv. Expanding large pre-trained unimodal models with multimodal information injection for image-text multimodal classification. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15471–15480, 2022. 1

[32] Jungwoo Lim, Dongsuk Oh, Yoonna Jang, Kisu Yang, and Heuiseok Lim. I know what you asked: Graph path learning using AMR for commonsense reasoning. In *ICCL*, pages 2459–2471, Barcelona, Spa(Online), Dec. 2020. International Committee on Computational Linguistics. 2

[33] Siqu Long, Soyeon Caren Han, Xiaojun Wan, and Josiah Poon. Gradual: Graph-based dual-modal representation for image-text matching. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2463–2472, 2022. 2

[34] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10434–10443, 2020. 6, 8

[35] Kai Nakamura, Sharon Levy, and William Yang Wang. Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France, May 2020. European Language Resources Association. 8

[36] 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 Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. 1

[37] Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. Centralnet: a multilayer approach for multimodal fusion. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 6, 8

[38] 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 *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021. 2

[39] Ziqi Wang, Xiaozhi Wang, Xu Han, Yankai Lin, Lei Hou, Zhiyuan Liu, Peng Li, Juanzi Li, and Jie Zhou.

CLEVE: Contrastive Pre-training for Event Extraction. In *ACL-IJCNLP (Volume 1: Long Papers)*, pages 6283–6297, Online, Aug. 2021. ACL. 2

[40] Hexiang Wu, Peifeng Li, and Zhongqing Wang. Multimodal event classification in social media. In *International Conference on Neural Information Processing*, pages 338–350. Springer, 2023. 2

[41] Kan Wu, Houwen Peng, Zhenghong Zhou, Bin Xiao, Mengchen Liu, Lu Yuan, Hong Xuan, Michael Valenzuela, Xi (Stephen) Chen, Xinggang Wang, Hongyang Chao, and Han Hu. Tinyclip: Clip distillation via affinity mimicking and weight inheritance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21970–21980, October 2023. 6

[42] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*, 2017. 2

[43] Yizhou Zhang, Loc Trinh, Defu Cao, Zijun Cui, and Yan Liu. Detecting out-of-context multimodal misinformation with interpretable neural-symbolic model, 2023. 2

[44] Heqing Zou, Meng Shen, Chen Chen, Yuchen Hu, Deepu Rajan, and Eng Siong Chng. UniS-MMC: Multimodal classification via unimodality-supervised multimodal contrastive learning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 659–672, Toronto, Canada, July 2023. Association for Computational Linguistics. 2, 6, 7, 8