Bridging Pixels and Words: Mask-Aware Local Semantic Fusion for Multimodal Media Verification

Supplementary Material

6. Loss Calculation

For the binary classification, the loss is obtained by calculating the cross-entropy loss between the classification head ($H_{\text{b}} : \mathbb{R}^D \rightarrow \mathbb{R}^2$) outputs and the ground truth labels y_{b}.

$$\mathcal{L}_{\text{bcls}} = \mathcal{L}_{\text{ce}}(L_{\text{b}}(c_{\text{b}}), y_{\text{b}}), \tag{12}$$

where $\mathcal{L}_{\text{ce}}$ denotes the cross-entropy loss.

For the manipulation type classification task, we input features c_{v} and c_{l} to classification heads $H_{\text{v}} : \mathbb{R}^D \rightarrow \mathbb{R}^2$ and $H_{\text{l}} : \mathbb{R}^D \rightarrow \mathbb{R}^2$, respectively, and calculate the loss as follows:

$$\mathcal{L}_{\text{mcls}} = \mathcal{L}_{\text{ce}}(\text{Cat}(H_{\text{v}}(c_{\text{v}}), H_{\text{l}}(c_{\text{l}})), y_{\text{m}}), \tag{13}$$

where y_{m} represents the truth labels for manipulation type classification and "Cat" indicates the concatenation operation.

For text manipulation grounding, we need to determine the index of the manipulated token. We can get the prediction at each position by inputting $s_{\text{cap}} \in \mathbb{R}^{L \times D}$ to the linear classification header $L_{tg} : \mathbb{R}^{L \times D} \rightarrow \mathbb{R}^{L \times 2}$. The loss for the text grounding is:

$$\mathcal{L}_{\text{tg}} = \mathcal{L}_{ce}(H_{tg}(s_{\text{cap}}), y_{tg}), \tag{14}$$

where the $y_{tg} = \{y_i\}_{i=1}^{L}$ denotes whether the i-th token is manipulation or not.

For image manipulation grounding, we can get the bounding box by inputting $c_{\text{bbox}} \in \mathbb{R}^D$ to the linear classification header $H_{ig} : \mathbb{R}^D \rightarrow \mathbb{R}^4$. Then the loss is:

$$\mathcal{L}_{\text{ig}} = \mathcal{L}_{\text{L1}}(H_{ig}(c_{\text{bbox}}) - y_{ig}) + \mathcal{L}_{\text{GIoU}}(H_{ig}(c_{\text{bbox}}) - y_{ig}), \tag{15}$$

where the y_{ig} represents the manipulated image grounding label. The $\mathcal{L}_{\text{L1}}$ and $\mathcal{L}_{\text{GIoU}}$ denote the L1 loss and GIoU loss [25].

We combine the above losses to obtain the final loss function, where the α, β, and γ are hyperparameters that control the importance:

$$\mathcal{L} = \mathcal{L}_{bcls} + \alpha\mathcal{L}_{mcls} + \beta\mathcal{L}_{ig} + \gamma\mathcal{L}_{tg}. \tag{16}$$

In our experiment, we used the following hyperparameters: $\alpha = 1.5, \beta = 0.1, \gamma = 1$.

6.1. Implementation Details of Datasets

Since DGM4 is the only dataset currently available for media manipulation detection and grounding, in order to more comprehensively validate the effectiveness of our approach, we also conducted experiments on the multimodal fake news detection datasets Weibo21 and Weibo17. DGM4 is an artificial dataset, and Weibo17 and Weibo21 are real datasets; the excellent results achieved on both of them also serve to further demonstrate the applicability of our model.

DGM4. For media manipulation detection and grounding, we used the DGM4 dataset[26]. This is the only dataset currently available for this task. DGM4 consists of 230k image-text pairs with more than 77k original pairs and 152k processed pairs. The DGM4 dataset is constructed based on the VisualNews dataset [35], which was collected from several news agencies. In the DGM4 dataset, image manipulation includes face swapping (FS) and facial attribute manipulation (FA), while text manipulation includes text swapping (TS) and text attribute manipulation (TA).

Weibo17,Weibo21. For multimodal fake news detection, we choose the Weibo17 dataset and the Weibo21 dataset. Both Weibo17 and Weibo21 are Chinese datasets containing image and text pairs collected from the social media platform Weibo. The Weibo dataset collected by [6] contains 3749 fake news and 3783 real news for training, 1000 fake news and 996 real news for testing. In experiments, we follow the same steps in the work [6, 35] to remove the duplicated and low-quality images to ensure the quality of the entire dataset. We keep the same data split scheme as [2, 34]. The Weibo21 was created in 2021 by [20], where more recent social posts were collected with 4640 real news and 4487 fake news in total. We keep the same train-test split at a ratio of 9:1 of [34, 39].

6.2. Evaluation Metrics

DGM4. The evaluation metrics on DGM4 consist of a total of 12 metrics for the four tasks. For binary classification, we evaluate the Accuracy (ACC), Area Under the receiver operating characteristic curve (AUC), and equal error rate (EER). For manipulation classification, we evaluate mean mean F1 per class, (CF1) overall F1 (OF1), and mean average precision (MAP). For image manipulation grounding, we evaluate mean intersection over union (IoUmean), and IoU at thresholds of 0.5 (IoU50) and 0.75 (IoU75). For text manipulation grounding, we evaluate precision, recall, and F1 score.

Weibo17,Weibo21. The evaluation metrics of fake news detection include Accuracy on all samples. The metrics evaluated on real and fake news are Precision, recall and F1-score.