Abstract

Tasks that rely on multi-modal information typically include a fusion module that combines information from different modalities. In this work, we develop a Refiner Fusion Network (ReFNet) that enables fusion modules to combine strong unimodal representation with strong multimodal representations. ReFNet combines the fusion network with a decoding/defusing module, which imposes a modality-centric responsibility condition. This approach addresses a big gap in existing multimodal fusion frameworks by ensuring that both unimodal and fused representations are strongly encoded in the latent fusion space. We demonstrate that the Refiner Fusion Network can improve upon performance of powerful baseline fusion modules such as multimodal transformers. The refiner network enables inducing graphical representations of the fused embeddings in the latent space, which we prove under certain conditions and is supported by strong empirical results in the numerical experiments. These graph structures are further strengthened by combining the ReFNet with a Multi-Similarity contrastive loss function. The modular nature of Refiner Fusion Networks lends itself to be combined with different fusion architectures easily, and in addition, the refiner step can be applied for pre-training on unlabeled datasets, thus leveraging unsupervised data towards improving performance. We demonstrate the power of Refiner Fusion Networks on three datasets, and further show that they can maintain performance with only a small fraction of labeled data.

1. Introduction

In several real-world applications, decision making involves integrating multiple modalities such as vision, text, auditory, and possibly even the content creator and how people engage with the input \[15, 3, 37\]. The application of multi-modal inference or decision making systems span several fields such as hate speech detection \[15\], misinformation detection \[22\], reasoning tasks \[7\], etc. Multimodal modeling includes two broad steps: extraction of features from different modalities such as images and text, and fusion of the different modalities. Several choices are available for fusion, including late fusion, mid-fusion or early fusion \[14, 33\]. Early fusion integrates features extracted from multiple modalities, and uses the integrated feature representation for learning downstream tasks. On the other hand, late fusion integrates classification scores of different features \[41\] to obtain the final classification score. Whether it is early or late fusion, different fusion strategies exist. These include Concat fusion \[50\], where different features are concatenated first and input into a MLP to get an integrated feature, as well as Set-based fusion \[42\] that is based on permutation invariant functions or Graph-based fusion modules \[1\].

The approach in this paper is partly motivated by a lim-
iteration of current fusion strategies that predominantly encode multimodal information, ignoring potentially the importance of retaining unimodal signals. Even powerful multimodal architectures such as those with co-attention transformer layers across modalities, which attend from the visual stream over language stream, leverage such approaches for the final fusion.

We propose a complementary approach wherein we refine these fusion representations to optimally balance independent information that each modality carries with multimodal representations to inform the downstream task. To this end, we leverage notions of responsibility and contrastive learning to build what we refer to as a Refiner Fusion Network (ReFNet). The main idea behind ReFNet is to balance the fusion module with a ‘responsibility’ driven refiner module as illustrated in Figure 1. The refiner module drives creation of sets of artificial neurons preferentially tuned to specific modal inputs, while the fusion module drives creation of artificial neurons with mixed representations. The responsibility condition is akin to a regularization that helps guard against over-fitting of the fusion module to the downstream task, and reduces vanishing gradients with respect to specific input fusion modes. By simultaneously balancing the number of unimodal and multimodal neurons during training, a better representation of the input modalities is achieved to perform multi-modal fusion informed by the downstream task.

One interesting finding of this work is that by imposing a responsibility condition in addition to the downstream task, the refiner module induces strong graphical structures between modalities, which we can prove under certain conditions. Such a latent inductive graph, when coupled with metric learning, encourages a latent graphical structure across modalities and samples, which we believe leads to a stronger fusion representation and performance on downstream tasks. More importantly, as opposed to transductive graph modeling, an inductive graph avoids the burden of needing to “carry” an adjacency matrix during inference time, which can get very large in real-world use cases.

Our contributions can be summarized as follow:

1. We propose ReFNet, a refiner module that can be added to any given fusion module that helps to induce neurons that are each responsible for a specific modality. We show that ReFNet can boost performance even over powerful transformers and can help induce latent graphical structures that we can show under certain conditions.

2. When coupled with metric learning, which we call ReFNet$_{MS}$, we observe a further boost in performance, and surmise from the T-SNE plots that we can generate stronger clustering and representation of the different classes.

3. Lastly, we demonstrate that ReFNet has increased level of tolerance to lesser amount of labeled data, which also helps reduce annotation needs.

2. Related Work

**Responsibility problem** The notion of responsibility was first introduced by Zhang and colleagues. The goal of this paper was that when permutation invariant sets are mapped to a latent space, the neurons of the encoding function must be faithful to the discontinuities introduced from the input space to the set entries. This was expanded into Set and Graph Refiner networks in wherein an inner loop optimization was performed to divide inputs into set elements that satisfy the responsibility problem, in contrast to say splitting just the feature vector of the CNN of an image. This showed better performance in relational reasoning tasks, and this approach can be used with other permutation invariant architectures such as Deep Multimodal Sets.

**Multitask Learning** The idea behind multitask learning is to learn tasks in parallel but using a shared representation. A CentralNet architecture expanded this idea for multimodal fusion networks. Their approach was to create a central network that links modality specific networks. Each modality is allowed to make decisions independent of other modalities, while a central network aims to leverage the mixed modalities. Taskonomy builds the shared representation space by first learning several low level tasks that can be of generic value to several downstream tasks. ReFNet is designed to complement these approaches where the refiner operates on the shared representation space to decode back the unimodal signals.

**Graph based Fusion** In Graph based fusion modules, each input modality can be considered nodes of a graph with a known adjacency matrix or explicitly modeling interaction across modalities. The GINFusion model creates a representation of the graph in an embedding space using a dense graph connection. Adding ReFNet to the downstream loss pushes the fusion architecture to have strong unimodal and strong multimodal components, and to induce edges between modalities, when they exist (as we show later in the manuscript).

**Autoencoder** Autoencoders play a big role in unsupervised learning, transfer learning and dimensionality reduction. Set autoencoders can be used for dimensionality reduction of a set of features to a lower dimensional space. Multi-modal autodecoders have been developed for filling in missing data. A special case of the Refiner Network will be the cyclic loss function introduced in. ReFNet enables encoding of feature sets that feed into a fusion module.

**Metric Learning** Supervised deep metric learning has been the focus of several research efforts. Contrastive loss and triplet losses are being
widely used in several applications. In contrastive learning [23, 51], samples with similar labels are pulled together in the fusion embedding space while those with dissimilar labels are pulled apart. Contrastive loss has two ingredients for a given anchor: pool samples with similar labels (positive) and those with dissimilar labels (negative), and minimize distances in the embedding space for the former and maximize distances for the latter. Triplet loss uses anchors, where one positive and negative sample are chosen per anchor, which are typically the hardest examples for a given anchor. Other approaches include lifted structures [36], n-pair losses [46], quadruplets [10], angular loss [48], adapted triplet loss [54], and multi-similarity loss functions [49] that utilize pair-wise relations across samples in a batch. A recent work [34] demonstrated that gains with metric learning are more modest compared to what is commonly reported.

In this paper, we use the Multi-Similarity contrastive loss function in combination with the refiner network. We call this method ReFNet$_{MS}$, wherein the responsible weights are trained in combination with maximizing separation of dissimilar embeddings in the fusion space, that can simultaneously elicit the underlying graphical structure across modalities and samples.

3. Multimodal Refiner Fusion Network Design

Let $F_1$, $F_2$, $\cdots$, $F_M$ refer to featurized inputs of $M$ modalities to a fusion module. A fusion module then aggregates these inputs, and creates a fused embedding of the multi-modal features

\[ F_{\text{emb}} = \mathcal{A}([F_1, F_2, \cdots, F_M]) \]  
\( (1) \)

where $\mathcal{A}$ maps the input features to an embedding space. In the context of this paper, $\mathcal{A}$ can encompass many of the fusion methods in literature such as Concat/MLP or other fusion modules used with Concat-Bert, ViLBERT, MMBT, etc. The fused embedding is subsequently used to train a downstream task such as classification.

The refiner module that we introduce in this paper decomposes the fused embedding into set elements, and imposes a refiner loss that the decomposed set elements can capture the featurized inputs to the fusion module (Fig. 2).

\[ R_i = \mathcal{D}_i(F_{\text{emb}})^{\forall i = 1, 2, \cdots, M} \]  
\( (2) \)

where $R_i$ are the set elements generated by the refiner module, $\mathcal{D}$. We then introduce a self-supervised loss function, $\mathcal{C}_{1,ss}$, for each fusion input

\[ \mathcal{C}_{1,ss} = 1 - \text{Cosine Similarity}(R_i, H_i(F_i))^{\forall i = 1, 2, \cdots, M} \]  
\( (3) \)

where $H_i$ is a mapping of the features to the refiner space. In general $H_i$ can be the identity transformation unless the problem necessitates a lower dimensional refiner space than the feature space (e.g. when not sufficient training samples are present). The total refiner cost function is $\sum_{i=1}^{M} \gamma_i \mathcal{C}_{1,ss}$ where $\gamma_i$ are the weights of the refiner cost function associated with the different features.

3.1. Refiner as a self-supervision module

One advantage of the refiner module is that it is self-supervised and can leverage unlabeled data (refer Fig. 2). Hence, unsupervised data, when available, can be pretrained initially by the refiner module before training on the downstream task. This helps in reducing the amount of training labels required, especially when considering that labeling on multi-modal tasks is typically more expensive than in unimodal tasks such as classification.

3.2. Refiner Fusion Module on Multi-modal transformers

Refiner module can be applied on top any fusion architecture that takes as input feature streams from the multiple modalities that are finally fused together. These streams can be in the form of a set or a graph. Multi-modal transformers such as MMBT [24], Visual Bert [26] and ViLBERT [29] have emerged as strong multi-modal models in the recent past. They combine the power of BERT model [12] for processing text, caption or OCR, with powerful ResNet models [17] to capture image features. In ViLBERT, a multi-modal co-attention model is used that has demonstrated powerful state-of-the-art performance on many public multi-modal benchmark tasks.

In the rest of the paper, we apply ReFNet on top of ViLBERT architecture. The output post the co-attention layers of the text and visual streams are fused, and a decoder is applied to the fused embedding to decode back the text and visual embeddings using an MLP with hidden layers.

4. Inducing Latent Graph Structures

While we do not explicitly model graph in this paper, we show in this section that our proposed method can exploit hidden graphical connections in the data, both across modes within a training sample and across samples in a batch as illustrated in Figure 3. In Theorem 4.1 we show that when an (unknown) adjacency matrix, $A$, exists that contains the connections across modalities and when the fusion network and refiner network are linear, then the inverse of the weights of the refiner network contains the weighted adjacency matrix.

**Theorem 4.1.** Let $A$ be an unknown adjacency matrix and $W$ be the weights of an affine transformation to generate the fusion embedding, $E$. The weights $\Gamma$ of a linear refiner satisfy the property, $\Gamma^T A \mathbf{W} = I_m$.

**Proof.** Let $F \times d$ represent features where $m$ is the number of modalities and $d$ is the size of each feature vector. We
Figure 2. Schematic of the proposed algorithm with a refiner and contrastive loss module. The image based features $F_1, F_2, F_3, F_4$ and text based features $F_5, F_6$ are fused together, and a refiner is applied on the fused embedding to generate refiner outputs $R_1, R_2, \ldots, R_6$ which are used to define a self-supervised loss function and a supervised Multi-similarity contrastive loss is also used across samples in a batch.

Assume that each modality has a feature vector of dimension $d$ without any loss of generality (otherwise they can be padded with zeros).

The unknown adjacency matrix, $A^{m \times m}$ contains ones whenever modality $i$ and $j$ have an edge between them. Let $W^{k \times m}$ be the unknown coefficients that create the fusion embeddings, $E^{k \times d}$ from the features as shown in Equation 4.

$$E = WAF$$

where $E$ is the embedding in a $k \times d$ space. The fusion module generates weights, $W^* = WA$. The refiner calculates weights $\Gamma^{m \times k}$ such that $\tilde{F} = \Gamma E = \Gamma WAF$. Since the refiner network finds weights such that $\tilde{F} = F$ (refer Eq. 3 where $D(x) = \Gamma x$ and $H_i$ is identity), we have

$$F = \Gamma WAF.$$

Since the above holds $\forall F$, $\Gamma WA = I_m$ where $I_m$ is an identity matrix of size $m \times m$.

If $WA$ is invertible, then $\Gamma = (WA)^{-1}$. Without using refiner, the weights, $W^*$ will be tuned by a downstream task of much lower dimension than the refiner module, and therefore the weights will be tuned to generate a good representation of the graph in a much lower dimensional space, thereby failing to induce the latent graphical structure.

**Corollary 4.1.** When $k = m$ and $W$ has a rank equal to $m$, $\Gamma^{-1}$ is the weighted adjacency matrix, $WA$.

The above follows directly from Equation 4 because $WA$ has a rank of $m$.

### 4.1. Contrastive Loss

When the self-supervised refiner module is combined with a contrastive loss, weights are driven to induce graphical structures both within modalities and across samples. Without refiner, metric learning maximizes separation of fused embeddings based on the downstream classification task. But the addition of refiner to metric learning enables separation also across modalities (because the refiner is responsible) as we demonstrate with a T-SNE figure later in the manuscript. We use the Multi-similarity loss in this paper, though other contrastive loss functions can be used. The Multi-similarity loss for $T$ training samples in a batch is calculated as

$$L_{MS} = \frac{1}{T} \sum_{i=1}^{T} \left[ \frac{1}{\alpha} \log[1 + \sum e^{-\alpha(S_{ik}-\lambda)}] + \frac{1}{\beta} \log[1 + \sum e^{\beta(S_{ik}-\lambda)}] \right]$$

where $S_{ij}$ is the similarity (dot product) between samples and $\alpha, \beta$ and $\lambda$ are hyperparameters. We provide the overall
are the weights of the fusion and refiner networks and \( \eta \), \( \zeta \) coefficients, the learning rate.

5. Experiments

For this study, we use the pre-trained ViLBERT model \([29]\) as a baseline for training and testing ReFNet and ReFNet\(_{MS}\). Similar to \([25]\), the ViLBERT model was only unimodally pretrained. Instead of using the element-wise product \([29]\) to fuse the unimodal representations of images and texts, we concatenate them and feed them into a new linear layer to formulate the overall representation. We are able to reproduce the baselines reported for each of examples in this section. We feed the fused embedding to two new linear layers. The Multi-similarity loss function is integrated with the fused embedding, and included during the final training runs. We used the MMF framework \([44]\) built on PyTorch \([38]\) for setting up the training and evaluation pipelines.

5.1. MM-IMDB

The multi-modal IMDB dataset \([2]\) contains 25,959 movies and their poster, genre, plot and other metadata fields such as year, language, writer, director, and aspect ratio. The goal is to classify the movie into 24 categories. Each movie can contain more than one class. The micro-f1 and macro-f1 scores were used to evaluate performance. The original dataset contains a baseline using Gated Multimodal methods, and is compared with ReFNet and ReFNet\(_{MS}\). Since each movie can simultaneously have multiple classes present, the precision and recall scores are calculated based on the f-score as follows \([30]\).

The macro f1 score is calculated from the precision, \( p_j \) and recall \( r_j \) of each class as

\[
f_{1}^{\text{macro}} = \frac{1}{N} \sum_{i=1}^{N} \frac{2 \times p_j \times r_j}{p_j + r_j}.
\]

The micro f1 score is calculated using all the class labels together as

\[
f_{1}^{\text{micro}} = \frac{2 \times p_{\text{micro}} \times r_{\text{micro}}}{p_{\text{micro}} + r_{\text{micro}}}
\]

where \( p_{\text{micro}} \) and \( r_{\text{micro}} \) are the precision and recall across all classes calculated based on the total number of true positives, false positives and false negatives.

We used the AdamW optimizer, with a Cosine warmup and Cosine decay learning rate scheduler. The value of \( \epsilon \) for AdamW optimizer was set to \( 1e^{-8} \) with corresponding \( \beta_1 \) and \( \beta_2 \) as 0.9 and 0.999. The batch size was set to 32, learning rate was set to \( 5e^{-5} \) and the fused embedding dimension was set to 512. An MLP with a hidden layer was used for the decoding refiner module. For the metric loss function, values of \( \alpha = 50 \) and \( \beta = 2 \) were chosen. We used values of \( \eta_1 = 0.1 \), \( \eta_2 = 0.1 \), \( \zeta = 0.1 \).

5.2. Hateful Memes

Hateful Memes dataset \([25]\) contains over 10,000 multimodal examples (image and test) with the goal of detecting if an input is hateful or not. The dataset is constructed...
such that unimodal models struggle and only multi-modal models can succeed (see Fig. 4). Difficult examples (“benign confounders”) are added to the dataset to make it hard to rely on unimodal signals. The dataset comprises of five different types of memes: multimodal hate, where benign confounders were found for both modalities, unimodal hate where one or both modalities were already hateful on their own, benign image and benign text confounders and finally random not-hateful examples. There were 1,000 samples in the validation dataset and 2,000 examples in the test dataset. We used values of \( \eta_i = 0.1 \), \( \zeta = 0.05 \).

5.3. SNLI Visual Entailment

The SNLI Visual Entailment (SNLI-VE) dataset \[52\] consists of image-sentence pairs whereby a real world image premise and a natural language hypothesis are given. The goal is to determine if the natural language hypothesis can be concluded given the information provided by the image premise. Three labels, entailment (hypothesis is true), neutral or contradiction (hypothesis is false), are assigned to image-sentence pairs. The dataset has 550k image-sentence pairs generated based on the Stanford Natural Language Inference (SNLI) \[5\] corpus and Flickr30k \[40\] dataset. The training setup is the same as Hateful Memes dataset except we use a batch size of 480, an initial learning rate of 5.0e-05, and a total of 10000 updates. We used values of

| model               | macro f1 test   | micro f1 test   |
|---------------------|-----------------|-----------------|
| ViLBERT             | 58.48 ± 0.25    | 66.77 ± 0.14    |
| ViLBERT-VQA2        | 57.70           | 66.42           |
| ViLBERT-COCO        | 57.72           | 65.63           |
| ViLBERT-cc small    | 58.20           | 66.70           |
| ReFNet              | 58.75 ± 0.07    | 67.02 ± 0.15    |
| ReFNetMS            | 58.96 ± 0.09    | 67.31 ± 0.19    |

Table 1. Comparison of the macro and micro f1 score on the test set across ViLBERT in combination with different pretraining datasets \[45\].

6. Results

6.1. MM-IMDB

Table 1 compares the performance of ReFNet and ReFNetMS on the test set of MM-IMDB dataset to a ViLBERT baseline model in combination with different pre-trained models. In general, pretraining ViLBERT reduced the performance on MM-IMDB dataset, but ReFNet was able to improve upon the performance. A relative gain of 0.8% in the micro f1 score and 0.82% in the macro f1 score were observed. Compared to other pretraining modules, ReFNet had gains between 0.48% and 2.12 %. The improvement on micro and macro f1 score using ReFNetMS were statistically significant based on a t-test (p = 0.02 and 0.03 respectively).

6.2. Hateful Memes

Table 2 compares the performance of ReFNet and ReFNetMS with respect to the ViLBERT baseline. ReFNet improves the accuracy by 3.30% with a relative gain in the AUC Of 1.84%. The use of ReFNetMS improves the Accuracy further by 0.80%. The overall relative gain in AUC was 2.17%. Based on a t-test, both ReFNet and ReFNetMS had a statistically significant improvement on the AUC (p-value = 1e-4 and 0.006 respectively) and the accuracy (p-value < 1e-6 for both) on the test set.

6.3. SNLI Visual Entailment

ReFNet showed a small relative improvement on the test set (improvement in accuracy 0.1% ± 0.07 on the test set which was not significant). However, ReFNetMS improves the Accuracy relatively by 0.71% which was significant (p-value = 0.001). Since the dataset contains more than a hundred thousand examples, even a 1% improvement results in thousands of images being correctly classified.

7. Ablation Studies

For the ablation study for MM-IMDB, we successively chose a fraction of the labeled dataset (5, 10 and 20%) for
Table 2. Comparison of the Accuracy and AUC on both the validation and test sets for the Hateful Memes dataset. "base" refers to the baseline ViLBERT model.

| model     | Acc. val.    | AUC val.    | Acc. test.   | AUC test.   |
|-----------|--------------|-------------|--------------|-------------|
| base      | 60.71 ± 0.29 | 70.62 ± 0.42| 59.70 ± 0.20 | 70.53 ± 0.07|
| ReFNet    | 62.45 ± 1.09 | 70.87 ± 0.41| 63.00 ± 0.31 | 71.83 ± 0.13|
| ReFNetMS  | 63.29 ± 1.31 | 70.99 ± 0.37| 63.80 ± 0.36 | 72.06 ± 0.49|

Table 3. Ablation study on the MM-IMDB dataset. Fraction labeled is the fraction of labeled samples used during training for the logit binary cross entropy function.

| model     | frac labeled | macro f1 test | micro f1 test |
|-----------|--------------|---------------|---------------|
| baseline  | 5%           | 34.06         | 56.62         |
| ReFNet    | 5%           | 45.89         | 60.65         |
| ReFNetMS  | 5%           | 47.83         | 63.13         |
| baseline  | 10%          | 46.73         | 62.65         |
| ReFNet    | 10%          | 48.70         | 63.41         |
| ReFNetMS  | 10%          | 51.73         | 64.31         |
| baseline  | 20%          | 56.08         | 65.75         |
| ReFNet    | 20%          | 56.43         | 66.24         |
| ReFNetMS  | 20%          | 57.12         | 66.63         |

Figure 5. Visualization of fusion features in reduced dimensions using T-SNE with perplexity set to 100. Top: fusion features of ViLBERT baseline showing the 3 clusters with entanglements. Middle: fusion features of ReFNet showing the 3 clusters are better separated with less entanglements. Bottom: fusion features of ReFNetMS showing the Refiner and Contrastive loss inducing six clusters across modalities and classes (three clusters for each of the vision and text modalities). The colors red, blue, and green represent three classes contradiction, entailment, and neutral.

8. Discussion

We started with the hypothesis that imposing a responsibility condition on Multimodal fusion architectures can improve performance on downstream tasks. We imposed the responsibility condition using a refiner module to decode each feature feeding into the fusion. This self-supervised loss function for each input feature used the cosine similarity between the decoded and original fusion features. We demonstrated this by integrating refiner on a strong transformer based multimodal baseline - the ViLBERT model. We showed that imposing responsibility can result in relative gains of over 5.53% on the accuracy and 1.84% on the Hateful memes dataset.

Reducing the amount of labeled data available showed that the self-supervised ReFNet can effectively leverage unlabeled data and maintain performance much better than the baseline ViLBERT model. For instance, with 5% of downstream classification while the self-supervised refiner module leverages unlabeled data. The baselines were re-run using the ViLBERT model. Metrics were reported on the test dataset based on the model corresponding to the best validation performance on 20000 iterations. ReFNet was able to achieve a 4.03% higher micro f1 and 11.83% higher macro f1 score compared to the ViLBERT baseline. ReFNetMS boosted the performance over baseline to 6.51% micro f1 score and 13.77% macro f1 score. These were 1.66 and 5.00 when 10% of the labeled data was used and 0.88 and 1.04 when 20% of the labeled data was used. A full summary of the ablation study is provided in Table 3 and illustrated in Fig. 6.

On the Hateful Memes dataset, the results based on successive reduction in the fraction of labeled data are summarized in Table 4. Using just 5% of the labeled data, the area under the ROC curve on the test set improves by 2.52% and 3.84% using ReFNet and ReFNetMS respectively. The corresponding improvements in accuracy are 4.14% and 4.60%. When using 10% of the labeled dataset, the accuracy improved by 1.50% and the AUC improved by 0.50%, and when using 20% of the labeled dataset, these were 3.50% and 2.59% respectively.
Figure 6. Comparison of performance of the baseline ViLBERT model, ReFNet and ReFNet\textsubscript{MS} across datasets with 5%, 10% and 20% of labeled training data available. Figure on the left shows micro and macro f1 scores on the validation dataset and that on the right shows the scores on the test dataset.

| model         | 5% V | 5% V | 5% T | 5% T | 10% V | 10% V | 10% T | 10% T | 20% V | 20% V | 20% T | 20% T |
|---------------|------|------|------|------|-------|-------|-------|-------|-------|-------|-------|-------|
| baseline      | 55.56 | 59.4 | 52.8 | 57.12 | 60.32 | 54.45 | 60.28 | 61.08 | 50.20 | 50.2 | 61.31 |
| ReFNet        | 56.20 | 58.65 | 56.94 | 59.64 | 61.84 | 54.65 | 60.87 | 54.96 | 61.95 | 53.70 | 63.90 |
| ReFNet\textsubscript{MS} | 57.34 | 62.02 | 57.40 | 60.96 | 61.82 | 54.25 | 60.78 | 58.13 | 62.78 | 58.30 | 62.86 |

Table 4. Ablation study on the Hateful Meme dataset. V refers to validation set, T refers to the test set, and 5%, 10% and 20% refer to the corresponding fraction of labeled training data used. AUC is the area under the ROC curve and ACC is the accuracy of the predicted model. All the reported values correspond to the model with the best performance on the validation dataset.
References

[1] Michalis Angelou, Vassilis Solachidis, Nicholas Vretos, and Petros Daras. Graph-based multimodal fusion with metric learning for multimodal classification. Pattern Recognition, 95:296–307, 2019.

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

[3] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanahalli. Multimodal fusion for multimedia analysis: a survey. Multimedia systems, 16(6):345–379, 2010.

[4] Pierre Baldi. Autoencoders, unsupervised learning, and deep architectures. In Proceedings of ICML workshop on unsupervised and transfer learning, pages 37–49. JMLR Workshop and Conference Proceedings, 2012.

[5] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference, 2015.

[6] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. arXiv preprint arXiv:1509.05019, 2015.

[7] Remi Cadene, Hedi Ben-Younes, Matthieu Cord, and Nicolas Thome. Murel: Multimodal relational reasoning for visual question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1989–1998, 2019.

[8] Rich Caruana. Multitask learning. Machine learning, 28(1):41–75, 1997.

[9] Hung-I Harry Chen, Yu-Chiao Chiu, Tinghe Zhang, Songyao Zhang, Yufei Huang, and Yidong Chen. Gsae: an autoencoder with embedded gene-set nodes for genomics functional characterization. BMC systems biology, 12(8):45–57, 2018.

[10] Wei-Che Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 403–412, 2017.

[11] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1335–1344, 2016.

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

[13] Yueqi Duan, Wenzhao Zheng, Xudong Lin, Jiwen Lu, and Jie Zhou. Deep adversarial metric learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2780–2789, 2018.

[14] Konrad Gadzicki, Razieh Khamehshahari, and Christoph Zetzsche. Early vs late fusion in multimodal convolutional neural networks. In 2020 IEEE 23rd International Conference on Information Fusion (FUSION), pages 1–6. IEEE, 2020.

[15] Raul Gomez, Jaume Gibert, Lluís Gomez, and Dimosthenis Karatzas. Exploring hate speech detection in multimodal publications. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 1470–1478, 2020.

[16] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR ’06), volume 2, pages 1735–1742. IEEE, 2006.

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

[18] Junlin Hu, Jiwen Lu, and Yap-Peng Tan. Discriminative deep metric learning for face verification in the wild. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1875–1882, 2014.

[19] Qian Huang, Horace He, Abhay Singh, Yan Zhang, Ser Nam Lim, and Austin R Benson. Better set representations for relational reasoning. Advances in Neural Information Processing Systems, 33, 2020.

[20] Natasha Jaques, Sara Taylor, Akane Sano, and Rosalind Picard. Multimodal autoencoder: A deep learning approach to filling in missing sensor data and enabling better mood prediction. In 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), pages 202–208. IEEE, 2017.

[21] Mahmut Kaya and Hasan Şakir Bilge. Deep metric learning: A survey. Symmetry, 11(9):1066, 2019.

[22] Dhruv Kathar, Jaipal Singh Goud, Manish Gupta, and Vaibhav Ghatikar. Mvae: Multimodal variational autoencoder for fake news detection. In The World Wide Web Conference, pages 2915–2921, 2019.

[23] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. arXiv preprint arXiv:2004.11362, 2020.

[24] Douwe Kiela, Suvarat Bhooshan, Hamed Firooz, and Davide Testuggine. Supervised multimodal bitransformers for classifying images and text. arXiv preprint arXiv:1909.02950, 2019.

[25] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes, 2020.

[26] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and powerful visuo-linguistic model. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1470–1478, 2019.

[27] Zechao Li and Jinhui Tang. Weakly supervised deep metric learning for community-contributed image retrieval. IEEE Transactions on Multimedia, 17(11):1989–1999, 2015.

[28] Wei Liu and Shih-Fu Chang. Robust multi-class transductive learning with graphs. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pages 381–388. IEEE, 2009.
[29] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. arXiv preprint arXiv:1908.02265, 2019.

[30] Giorgi Madjarov, Dragi Kocev, Dejan Gjorgjevikj, and Sašo Džeroski. An extensive experimental comparison of methods for multi-label learning. Pattern recognition, 45(9):3084–3104, 2012.

[31] Sijie Mai, Haifeng Hu, and Songlong Xing. Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 164–172, 2020.

[32] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, and Brendan Frey. Adversarial autoencoders. arXiv preprint arXiv:1511.05644, 2015.

[33] Vicente P Minotto, Claudio R Jung, and Bowon Lee. Simultaneous-speaker voice activity detection and localization using mid-fusion of svm and hmmgs. IEEE Transactions on Multimedia, 16(4):1032–1044, 2014.

[34] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In European Conference on Computer Vision, pages 681–699. Springer, 2020.

[35] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Deep multi-modal sets. arXiv preprint arXiv:1912.01703, 2019.

[36] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion, 37:98–125, 2017.

[37] Austin Reiter, Menglin Jia, Pu Yang, and Ser-Nam Lim. Deep multi-modal sets. arXiv preprint arXiv:2003.01607, 2020.

[38] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 815–823, 2015.

[39] Amanpreet Singh, Vedanuj Goswami, Vivek Natarajan, Yu Jiang, Xinlei Chen, Meet Shah, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Mmf: A multimodal framework for vision and language research, 2020.

[40] Amanpreet Singh, Vedanuj Goswami, and Devi Parikh. Are we pretraining it right? digging deeper into visio-linguistic pretraining. arXiv preprint arXiv:2004.08744, 2020.

[41] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In Proceedings of the 30th International Conference on Neural Information Processing Systems, pages 1857–1865, 2016.

[42] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 815–823, 2015.

[43] Amanpreet Singh, Vedanuj Goswami, Vivek Natarajan, Yu Jiang, Xinlei Chen, Meet Shah, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Mmf: A multimodal framework for vision and language research, 2020.
[59] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision, pages 2223–2232, 2017.