REFINING MULTIMODAL REPRESENTATIONS USING A MODALITY-CENTRIC SELF-SUPERVISED MODULE

Anonymous authors
Paper under double-blind review

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 self-supervised module, called REFINER, that refines multimodal representations using a decoding/defusing module applied downstream of the fused embedding. Our approach strengthens representation of features computed upstream of the fusion module that are relevant to the downstream task. Our approach provides both stronger generalization and reduced over-fitting. REFINER is only applied during training keeping the inference time intact. The modular nature of REFINER lends itself to be combined with different fusion architectures easily. We demonstrate the power of REFINER on three datasets over powerful baseline fusion modules, and further show that they give a significant performance boost in few shot learning tasks.

1 INTRODUCTION

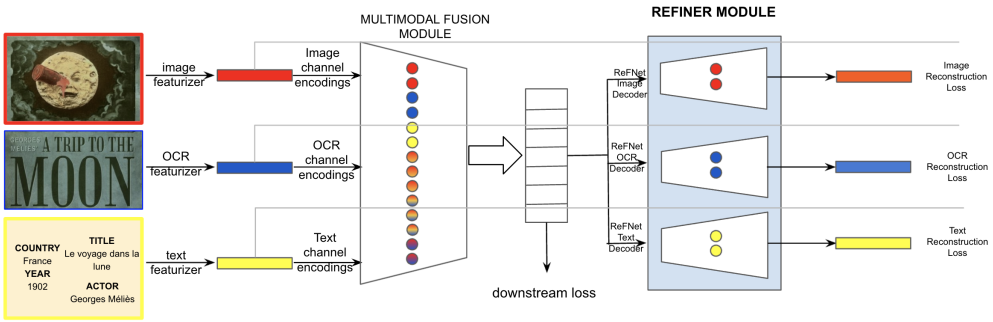

Figure 1: Illustration of the idea behind the REFINER module.

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 Ngiam et al. (2011); Atrey et al. (2010); Oviatt et al. (2003). The application of multi-modal inference or decision making systems spans several fields such as hate speech detection Gomez et al. (2020), misinformation detection Khattar et al. (2019), reasoning tasks Cadene et al. (2019), etc. Multimodal modeling includes two broad steps: extraction of features from different modalities, and fusion of the different modalities. Several choices are available for fusion, including late fusion, mid-fusion or early fusion Gadzicki et al. (2020); Minotto et al. (2014). Early fusion integrates features extracted from multiple modalities, and uses the integrated feature representation for learning downstream tasks. Late fusion integrates classification scores of different features Poria et al. (2017) to obtain the final classification score. Some of these fusion methods include concat fusion Wang et al. (2020) based on concatenated features, Set-based fusion Reiter et al. (2020) based on permutation invariant functions or Graph-based fusion modules Angelou et al. (2019).

In this work, we propose to refine fusion representations to optimally represent features computed upstream of the fusion module to inform the downstream task. This approach is partly motivated

by a limitation of current fusion strategies that predominantly encode multimodal information, ignoring potentially the importance of retaining unimodal or weakly multimodal (e.g. transformers with cross-modal key-value pairs) signals. The REFINER module, used with a multimodal fusion architecture is referred to as Refiner Fusion Network (ReFNet). The main idea behind ReFNet is to balance the fusion module with a REFINER module using a reconstruction loss. 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 target for REFINER can be unimodal representations when using non-transformer architectures such as GMU, concat etc. or weakly multimodal signals for early fusion methods such as multimodal transformers, and in general can strive for the representation of any intermediate feature upstream of the fusion module.

Our contributions can be summarized as follow:

1. We propose a REFINER module that can be added to any given fusion module to induce modality-specific neurons. We show that ReFNet, a fusion architecture combined with the REFINER module, can boost performance even over powerful transformers. We also show that ReFNet boosts performance on retrieval tasks where the query is unimodal but (key, value) pair is multimodal.
2. When coupled with metric learning, which we call $\text{ReFNet}_{\text{MS}}$, we observe a further boost in performance, and surmise from the T-SNE plots that this approach creates two strong clusters per class, a modality-responsible and a mixed cluster.
3. Lastly, we demonstrate that the REFINER has increased level of tolerance to lesser amount of labeled data.

2 Related Work

Multitask Learning The idea behind multitask learning is to learn tasks in parallel but using a shared representation Caruana (1997). A CentralNet architecture expanded this idea for multimodal fusion networks Vielzeuf et al. (2018); Pérez-Rúa et al. (2019) by creating 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 Zamir et al. (2018) builds the shared representation space by first learning several low level tasks. Recently, the UniT architecture was introduced Hu & Singh (2021) providing an end-to-end framework for multimodal multitasking.

Graph based Fusion In Graph based fusion modules, each input modality can be considered nodes of a graph with a known adjacency matrix Zhang et al. (2019); Angelou et al. (2019) or explicitly modeling interaction across modalities Mai et al. (2020). The GINFusion model Xu et al. (2018) 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 (refer Appendix, Section A.4 for more details). The GRN technique for relational reasoning problems used the notion of responsibility to refine graph embeddings Huang et al. (2020).

Autoencoder Autoencoders play a big role in unsupervised learning, transfer learning and dimensionality reduction Baldi (2012); Burda et al. (2015); Makhzani et al. (2015). Set autoencoders can be used for dimensionality reduction of feature sets Chen et al. (2018). Multi-modal autodecoders have been developed for filling in missing data Jaques et al. (2017). A special case of the REFINER will be the cyclic loss function introduced in Zhu et al. (2017).

Metric Learning Supervised deep metric learning has been the focus of several research efforts Li & Tang (2015); Kaya & Bilge (2019); Duan et al. (2018). Contrastive loss Hadsell et al. (2006); Hu et al. (2014) and triplet losses Schroff et al. (2015); Cheng et al. (2016) are being widely used in several applications. In contrastive learning Khosla et al. (2020); Weinberger & Saul (2009), samples with similar labels are pulled together in the fusion embedding space while those with dissimilar labels are pulled apart. 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 Oh Song et al. (2016), n-pair losses Sohn (2016), quadruplets Chen et al. (2017), angular loss Wang et al. (2017), adapted triplet loss Yu et al. (2018), and multi-similarity loss

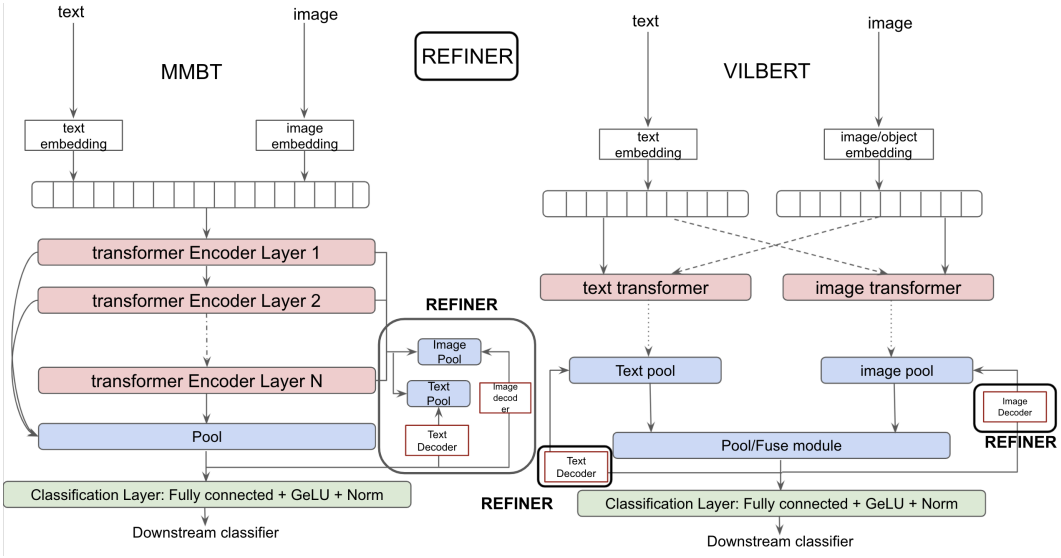

Figure 2: Schematic of the Refiner Network applied to (left) MMBT architecture and (right) VILBERT architecture. Decoders are added for each uni-modal feature to compute the Refiner loss. Note that the refiner targets can leverage any of the standard pooling operations used in Transformers ranging from just using the last encoder layer to pooling across several encoder layers.

functions Wang et al. (2019) that utilize pair-wise relations across samples in a batch. A modality-alignment loss maximizing distance on differing identities was introduced in Wen et al. (2021). In this paper, we use the Multi-Similarity contrastive loss function in combination with the REFINER module. We call this method ReFNet$_{\text{MS}}$, wherein the REFINER module is also trained to 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 as

$$\text{F}_{\text{emb}} = \mathcal{A}([F_1, F_2, \cdots F_M]) \tag{1}$$

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, Linear or transformer based 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 applies a set of decoders to the fused embedding, and imposes a reconstruction loss between the decoded outouts and REFINER targets that are chosen upstream of the fusion module based on the architevture (Fig. 6).

$$R_i = \mathcal{D}_i(\text{F}_{\text{emb}}) \forall i = 1, 2, \cdots M \tag{2}$$

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

$$\mathcal{C}_{\text{i,ss}} = \mathcal{L}(R_i, H_i(F_i)) \forall i = 1, 2, \cdots M \tag{3}$$

where H_i is a mapping of the features to the refiner space and $\mathcal{L}$ is a loss function. In general H_i can be the identity transformation in the context of simple architectures, but in the context of transformers, H_i usually involves modality specific pooling across transformer encoder layers. The total refiner cost function is $\sum_{i=1}^{M} \gamma_i \mathcal{C}_{\text{i,ss}}$ where γ_i are the weights of the refiner cost function associated with the different features.

3.1 REFINER AS A SELF-SUPERVISION MODULE

One feature of the REFINER module is that it is self-supervised and can leverage unlabeled data. REFINER helps the fusion module to be pre-trained initially using the refiner losses before training on the downstream task. Three self-supervised loss functions were explored in this work, cosine similarity, mean-squared error, and refiner-contrastive loss. The refiner-contrastive loss imposes that, within a batch, the decoded signals for a particular input are closer to it's target compared to the other targets to within a margin. To calculate refiner-contrastive loss, we first calculate $r_{ij} = \text{Similarity}(R_i, H_j(F))$ which is the similarity of i^{th} decoded signal with the j^{th} target within a batch. The loss within a batch for a given sample t is then calculated as

$$\mathcal{L}(x_t, y_t) = \sum_{k=1, k\neq t}^{N_B} \mathbf{H}(r_{kt} - r_{tt}) \tag{4}$$

where $\mathbf{H}(x) = x$ when x is positive and zero otherwise and N_B is the batch size.

3.2 REFINER MODULE ON MULTI-MODAL TRANSFORMERS

The REFINER module can be applied on top of 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 Kiela et al. (2019), VisualBert Li et al. (2019) and ViLBERT Lu et al. (2019) have emerged as strong multi-modal models in the recent past. They combine the power of BERT model Devlin et al. (2018) for processing text, caption or OCR, with powerful ResNet models He et al. (2016) 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 MMBT, tokens from each modality are concatenated and used as different sentences (segments) of a transformer encoder.

In the rest of the paper, we apply ReFNet on top of MMBT and ViLBERT architecture, as well as simpler architectures such as Concat, LinCat and GMU. 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. REFINER can elicit hidden structures in the input multi-modal data that we describe in detail in section A.4 and demonstrate some interesting properties of linear REFINER networks in section A.6. We believe that the REFINER can generally be useful to other early or mid fusion architectures not discussed in this paper.

3.3 MULTI-SIMILARITY LOSS

The addition of supervised metric learning with REFINER enables separation of embeddings per class across modalities (because the REFINER is responsible) as we demonstrate with a T-SNE figure later in the manuscript. Without REFINER, metric learning maximizes separation of fused embeddings based on the downstream classification task. 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 Wang et al. (2019)

$$\mathcal{L}_{MS} = \frac{1}{T}\sum_{i=1}^{T}\left[\frac{1}{\alpha}log[1 + \sum e^{-\alpha(S_{ik}-\lambda)}]\right] + \left[\frac{1}{\beta}log[1 + \sum e^{\beta(S_{ik}-\lambda)}]\right], \tag{5}$$

where S_{ij} is the similarity (dot product) between samples and α, β and λ are hyperparameters. We provide the overall algorithm in Alg. 1 where γ_i are the self-supervised loss coefficients, ζ is the loss coefficient for contrastive loss, w_k are the weights of the fusion and refiner networks and η is the learning rate. Refer to section A.1 for additional details on the integration.

4 EXPERIMENTS

For this study, we use the ViLBERT and MMBT models Lu et al. (2019) as baseline for training and testing ReFNet and ReFNet$_{\text{MS}}$ in addition to GMU, Linear Sum and Concat baselines Arevalo et al. (2017). Similar to Kiela et al. (2020), the ViLBert model was only unimodally pretrained.

Algorithm 1 Algorithm for training Multi-modal Fusion Networks

Pretraining (optional)

Compute features, $F_1, F_2, \cdots, F_M$.
while $epoch \leq$ max epochs **do**
  $\mathcal{L}_{pretrain} = \sum \gamma_i \mathcal{C}_{i,ss}$
  $w_{k+1} = w_k - \eta \frac{\partial \mathcal{L}_{pretrain}}{\partial w}$
end while

Training

while $epoch \leq$ max epochs and stop criterion is not met **do**
  $\mathrm{F}_{\text{emb}} = \mathcal{A}([F_1, F_2, \cdots F_M])$
  $R_i = \mathcal{D}_i(\mathrm{F}_{\text{emb}}) \forall i = 1, 2, \cdots M$
  $\mathcal{L}_{train} = \mathcal{L}_{downstream} + \sum \gamma_i \mathcal{C}_{i,ss} + \zeta * \mathcal{L}_{MS}$
  $w_{k+1} \leftarrow w_k - \eta \frac{\partial \mathcal{L}_{train}}{\partial w}$
end while

For ViLBERT, we use the tokens from each of text and image channels as targets for REFINER, whereas for MMBT, we use average pooled outputs in the last transformer layer for each modality as REFINER targets as illustrated in Figure 2. We used the MMF framework Singh et al. (2020a) built on PyTorch Paszke et al. (2019) for setting up the training and evaluation pipelines. We plan to open source the developments in this paper with the MMF framework.

4.1 DATASETS

We use three datasets - the multimodal IMDB, Hateful Memes and SNLI visual entailment for this paper, which are described below.

MM-IMDB: The multi-modal IMDB dataset Arevalo et al. (2017) 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 belong to 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 units(GMU) Arevalo et al. (2017). Three models from the original paper, as well as ViLBERT baseline Singh et al. (2020b) and MMBT baselines are used here. Note that ReFNet combines each of the base fusion modules with the REFINER module. For the non-transformer baselines, the output of the unimodal text and image encoders were used as targets, and the REFINER decoder takes as input the activation layer post-gating. Since each movie can simultaneously have multiple classes present, the precision and recall scores are calculated based on the f-score as follows Madjarov et al. (2012).

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

$$f_1^{\text{macro}} = \frac{1}{N} \sum_{j=1}^{N} \frac{2 \times p_j \times r_j}{p_j + r_j}$$

where N is the number of classes. The micro f1 score is calculated using all the class labels together as

$$f_1^{\text{micro}} = \frac{2 \times p^{\text{micro}} \times r^{\text{micro}}}{p^{micro} + r^{micro}}$$

where p^{micro} and r^{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 the limiting factor for AdamW optimizer was set to $1e^{-8}$ and corresponding coefficients for first and second moments were set to 0.9 and 0.999 respectively. 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 in Eq. 5, values of $\alpha = 50$ and $\beta = 2$ were chosen. Hyperparameters η_1, η_2, and ζ were in the range of 0.0 to 0.3.

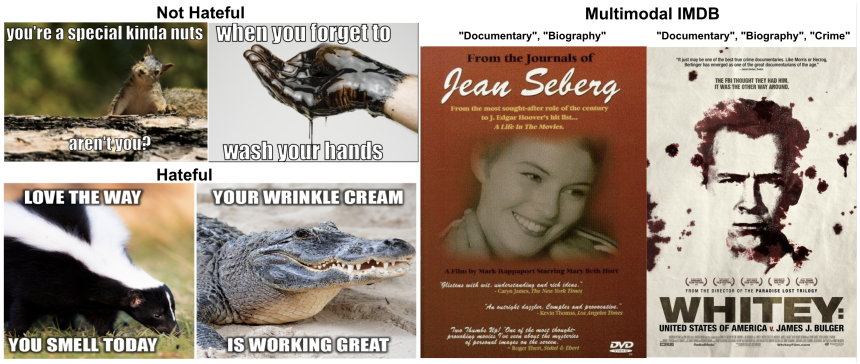

Figure 3: Examples from the (left) Hateful Memes dataset showing two not hateful (top) and two hateful (bottom) memes from the Hateful Memes dataset, and (right) two examples from the multimodal IMDB dataset with the genre that is to be predicted. There is also a corresponding text description for each input which is not shown, and each input may have multiple labels associated.

Hateful Memes: Hateful Memes dataset Kiela et al. (2020) contains over 10,000 multi-modal examples (image and text) 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. 3). 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 injected 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. Range for the hyperparameters η_i and ζ were between 0.0 to 0.3 in Algorithm 1.

We use a cross entropy loss for the two-label classification. We train on 8 NVIDIA Volt100 GPUs with a total batch size of 32 for a total of 22000 updates. We evaluate every 1000 updates and save the model with the best AUROC metric on the validation set. We use the AdamW optimizer with an initial learning rate of 1.0e-05. We use a linear decay learning rate schedule with warm up to train the model.

SNLI Visual Entailment: The SNLI Visual Entailment (SNLI-VE) dataset Xie et al. (2019) 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) Bowman et al. (2015) corpus and Flickr30k Plummer et al. (2016) 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.

4.2 Do Multimodal Fusion Architectures innately encode upstream(unimodal or weakly multimodal) information?

We performed an analysis of the impact of preserving upstream information on the performance on both the Hateful Memes and the mmIMDB datasets. We used the ReFNet weight (γ in Algorithm 1) as an index of preserving information upstream of the fusion module, with contrastive loss weight η set to zero. Fig. 4 shows that as we apply ReFNet with a loss weighted by γ, the accuracy initially increases but once the unimodality starts to dominate and prevent effective multimodal mixing, the accuracy starts dropping back.

model	Top-1	Top-2	Top-5	Top-10
baseline	27.44	42.83	60.51	67.02
ReFNet	42.67	46.99	58.15	70.05

Table 1: Image-based retrieval of the multimodal embeddings from the fusion module using unimodal image embedding from the upstream image encoder.

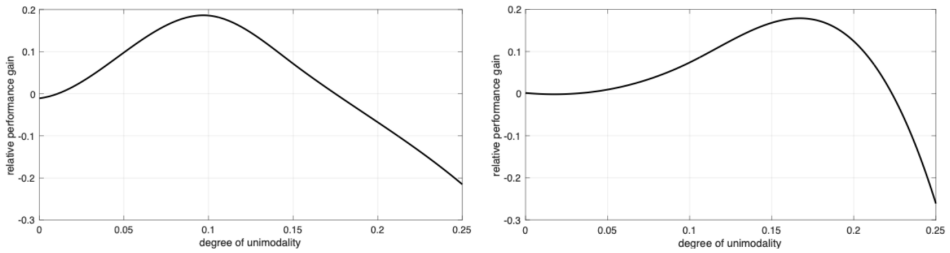

Figure 4: Relative change in accuracy as we increase the weight of the REFINER reconstruction loss, showcasing that no reconstruction loss (baseline) does not fully leverage information of the fusion module.

Next, we performed a study wherein we stored the embeddings using the baseline and ReFNet trained mmIMDB datasets in a bank, and tried to retrieve the target using just the image encoder using a k-nearest neighbor search. The query for the kNN search is an unimodal image encoder and the (key, value) pairs are the (fused embedding, target) pairs. Results are summarized in Table 1 demonstrating that REFINER achieves superior retrieval based on unimodal image signal compared to native Multimodal Fusion modules on the test set. REFINER consistently outperformed baseline when using unimodal text encoders as well (discussed in Appendix A.8). These experiments demonstrate a better encoding of the unimodal signals in the fused representation.

To further understand this problem, we performed studies with just the relevant unimodal features on the mmIMDB dataset. Full details of the study are provided in section A.5, with the results indicating that when only the relevant text and image signals are fed into the fusion network, REFINER is able to boost performance on both text-only and image-only inputs.

4.3 Impact of adding Multi-Similarity loss to embeddings

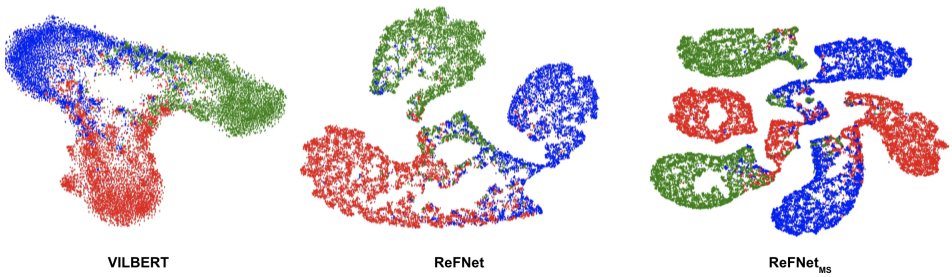

Figure 5: Visualization of fusion features in reduced dimensions using T-SNE. Left: fusion features of ViLBERT baseline showing the 3 clusters with entanglements. Center: fusion features of ReFNet showing the 3 clusters are better separated with less entanglements. Right: fusion features of RefNet$_{MS}$ 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.

We generated T-SNE plots on the fusion features for SNLI visual entailment dataset with ViLBERT, ReFNet and ReFNet$_{MS}$ algorithms. Fig. 5 shows that ViLBERT model generates 3 vaguely separated fusion feature clusters while the clusters generated by ReFNet has a better separation between clusters and clearly delineates them. The metric learning algorithm is able to further separate the clusters across modalities, therefore we observe six clusters, three different classes for each modality (vision and language).

5 Results

MM-IMDB: Table 2 compares the performance of non-transformer baselines for the multi-modal IMDB dataset against ReFNet. Table 3 compares the performance of ReFNet and ReFNet$_{MS}$ on a test set of MM-IMDB dataset to both ViLBERT and MMBT baseline models. ReFNet was generally

	Gated Multimodal Units		Linear Sum		Concat	
model	macro f1	micro f1	macro f1	micro f1	macro	micro f1
baseline	54.1	63.0	53.0	60.7	52.1	60.6
ReFNet	56.0	64.0	56.7	64.1	56.3	61.7

Table 2: Comparison of the performance of ReFNet with non-transformer baselines for the multimodal IMDB dataset as described in Arevalo et al. (2017).

	Hateful Memes				Multimodal IMDB	
model	Acc. val.	AUC val.	Acc. test.	AUC test.	macro f1 test.	micro f1 test.
ViLBERT	60.71 $\pm$0.29	70.62 $\pm$0.42	59.70 $\pm$0.20	70.53 $\pm$0.07	58.48 $\pm$ 0.25	66.77 $\pm$ 0.14
ReFNet	62.45 $\pm$ 1.09	70.87 $\pm$ 0.41	63.00 $\pm$ 0.31	71.83 $\pm$ 0.13	58.75 $\pm$ 0.07	67.02 $\pm$ 0.15
ReFNet$_{MS}$	63.29 $\pm$ 1.31	70.99 $\pm$ 0.37	**63.80 $\pm$ 0.36**	**72.06 $\pm$ 0.49**	**58.96 $\pm$ 0.09**	**67.31 $\pm$ 0.19**
MMBT	57.60 $\pm$ 0.76	63.14 $\pm$ 0.14	61.30 $\pm$ 0.35	67.76 $\pm$ 0.06	57.38 $\pm$ 0.04	66.97 $\pm$ 0.05
ReFNet	56.80 $\pm$ 1.87	61.68 $\pm$ 1.54	**62.30 $\pm$ 1.76**	69.32 $\pm$ 1.00	**58.43 $\pm$ 0.31**	**67.91 $\pm$ 0.08**
ReFNet$_{MS}$	58.00 $\pm$ 0.72	65.27$\pm$ 0.25	59.60 $\pm$ 0.44	**69.59 $\pm$ 0.15**	58.03 $\pm$ 0.06	67.57 $\pm$ 0.06

Table 3: Comparison of the performance on Hateful Memes and MMIMDB datasets.

able to improve upon the performance for all five baselines. Relative gains between 1.59-5.60% were observed for micro-f1 score and between 3.51-8.06% for macro-f1 scores of non-transformer baselines. For the transformer models, relative gains of 0.81-1.40% and 0.82-1.83% in micro and macro-F1 scores were observed, and these were statistically significant based on a t-test (p = 0.02 and 0.03 respectively).

Hateful Memes: Table 3 compares the performance of ReFNet and ReFNet$_{MS}$ with respect to the ViLBERT and MMBT baselines. ReFNet had a relative gain in the AUC Of 1.84% over ViLBERT, and 2.30% over MMBT baselines. The overall relative gain in AUC for ViLBERT and MMBT were 2.17% and 2.70% respectively. Based on a t-test, both ReFNet and ReFNet$_{MS}$ had a statistically significant improvement on the AUC (p-value = 1e-4 and 0.006 respectively) for both models.

SNLI Visual Entailment: ReFNet showed a small relative improvement on the test set (improvement in accuracy $0.1\% \pm 0.07$ on the test set which was not significant). However, ReFNet$_{MS}$ 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. This improvement with ReFNet$_{MS}$ is in the range of other SOTA improvements over baseline with similar capacity for this dataset (e.g. refer Table 1 in Xie et al. (2018) and table 1 in Shevchenko et al. (2021)).

6 Ablation Studies

Amount of Labeled Data 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 has a relative gain between 3.46-4.41% and 3.58-6.72% using ReFNet and ReFNet$_{MS}$ respectively. On the Multimodal IMDB dataset, ReFNet was able to achieve performance boosts in the range of 4 to 12% under reduced resource setting. Results of the reduced label study are discussed in detail in section A.3.

Targets for the REFINER module: The choice of targets for REFINER needs to be chosen carefully depending on the application. Choosing a target that is very upstream, such as the raw image pixels or text tokens, can result in the REFINER undoing the cross-modal attention weighting which is the hallmark of Multi-modal transformers. However, choosing a target which is just upstream of the fusion module might not be very useful as that information might already be represented in the fused embedding. To elucidate how the choice of REFINER target might affect the result, we perform a study of choosing five different modality pooler operations on the performance.

Table 5 compares performance of the five different modality poolers on the mmIMDB dataset. The results show that including penultimate layer on the modality pooling operation improves performance of REFINER compared to just including the last layer, which might indicate that different

model	5% V AUC	5% T Acc	5% T AUC	10% V AUC	10% T Acc	10% T AUC	20% V AUC	20% T Acc	20% T AUC
ViLBERT	59.4	52.8	57.12	60.32	54.45	60.28	50.20	50.2	61.31
ReFNet	58.65	56.94	59.64	61.84	54.65	**60.87**	61.95	53.70	**63.90**
$\text{ReFNet}_{\text{MS}}$	62.02	57.40	**60.96**	61.82	54.25	60.78	62.78	58.30	62.86
MMBT	59.08	54.50	59.48	62.37	55.93	65.89	64.30	56.74	66.41.
ReFNet	61.23	50.60	61.54	62.47	57.43	66.06	63.59	57.33	**68.06**
$\text{ReFNet}_{\text{MS}}$	59.72	58.00	**61.69**	62.26	57.00	**66.97**	63.95	56.10	65.97

Table 4: Ablation study on the Hateful Meme dataset (V: validation set, T: test set, x%: fraction of labels used, AUC: area under the ROC curve and ACC is the accuracy of the predicted model).

target	macro f1 val.	micro f1 val.	macro f1 test	micro f1 test
average last layer	58.85	67.21	57.64	66.76
average last two layers	59.18	67.96	58.30	67.20
average sum last layer	58.64	68.07	58.33	67.20
average sum last two layers	**59.05**	**68.09**	**58.66**	**67.78**
average concat	57.55	67.22	57.75	67.48

Table 5: Comparison of different REFINER targets using five different modality pooling operations on the test and validation performances.

encoder layers may contain relevant information for the corresponding modalities, which are not implicitly represented in the fused embeddings.

7 DISCUSSION

We started with the hypothesis that refining fused representations of Multimodal fusion architectures using a modality-centric decoder can improve performance on downstream tasks. We first demonstrated the gap in existing Multimodal transformer architectures by showing that even with unimodal inputs, refining the representation yields better performance. For these examples, the weights for the irrelevant modalities were identified as zero naturally by the REFINER, and achieved relative gain in performance of up to 2%. We further showed that ReFNet performs better on retrieval tasks where query is uni-modal but key is multi-modal. We also demonstrated that ReFNet provides bigger boosts in reduced resource setting and might help in reducing annotation requirements as in general. The choice of REFINER cost function also did not significantly change results (further details in section A.2).

For the transformer baselines, the targets for REFINER can be either image pixels and raw text, or the modality pooled inputs for each of the modalities. In general, the former did not give a boost in performance, likely because the REFINER acts against the information fusion happening within transformer encoders. The ablation studies on the REFINER targets indicated that using penultimate encoder layers had a positive effect on the performance. REFINER only had a modest increase in training time, while the use of Multi-Similarity loss had a training time increase of around 30-35% (refer section A.7 for more details). The relative contribution of the REFINER loss to the improvement in micro F1 score for the mmIMDB dataset was 46-56%, and the rest was due to contrastive loss. More than 80% of the relative contribution for Hateful Memes dataset was due to the self-supervised REFINER loss. Based on the overall improvements, we believe that our approach will also be helpful to refine embeddings with other baselines such as UNITER Lippe et al. (2020); Chen et al. (2019) (similar to MMBT) using the LXMERT fusion architecture or with other pretraining methods such as Li et al. (2020); Yu et al. (2020). Note that REFINER is also applied during fine-tuning and therefore can be used in tandem with other multimodal pretrained models.

Limitations and Future Work: We did not fully explore Metric Loss functions. Other modality pooling functions for REFINER targets were not fully explored and need to be understood.

REFERENCES

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.

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

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

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

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

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

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*, pp. 1989–1998, 2019.

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

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.

Weihua 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*, pp. 403–412, 2017.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. 2019.

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*, pp. 1335–1344, 2016.

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.

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*, pp. 2780–2789, 2018.

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

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

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, pp. 1735–1742. IEEE, 2006.

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*, pp. 770–778, 2016.

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*, pp. 1875–1882, 2014.

Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. *arXiv preprint arXiv:2102.10772*, 2021.

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.

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)*, pp. 202–208. IEEE, 2017.

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

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. Mvae: Multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference*, pp. 2915–2921, 2019.

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.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*, 2019.

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.

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.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pp. 121–137. Springer, 2020.

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

Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, and Helen Yannakoudakis. A multimodal framework for the detection of hateful memes. *arXiv preprint arXiv:2012.12871*, 2020.

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.

Gjorgji 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.

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, pp. 164–172, 2020.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

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

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *ICML*, 2011.

Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4004–4012, 2016.

Sharon Oviatt et al. Multimodal interfaces. *The human-computer interaction handbook: Fundamentals, evolving technologies and emerging applications*, 14:286–304, 2003.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.

Juan-Manuel Pérez-Rúa, Valentin Vielzeuf, Stéphane Pateux, Moez Baccouche, and Frédéric Jurie. Mfas: Multimodal fusion architecture search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6966–6975, 2019.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models, 2016.

Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125, 2017.

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

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*, pp. 815–823, 2015.

Violetta Shevchenko, Damien Teney, Anthony Dick, and Anton van den Hengel. Reasoning over vision and language: Exploring the benefits of supplemental knowledge. *arXiv preprint arXiv:2101.06013*, 2021.

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, 2020a.

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

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*, pp. 1857–1865, 2016.

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*, pp. 0–0, 2018.

Jian Wang, Feng Zhou, Shilei Wen, Xiao Liu, and Yuanqing Lin. Deep metric learning with angular loss. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2593–2601, 2017.

Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5022–5030, 2019.

Yikai Wang, Wenbing Huang, Fuchun Sun, Tingyang Xu, Yu Rong, and Junzhou Huang. Deep multimodal fusion by channel exchanging. *arXiv preprint arXiv:2011.05005*, 2020.

Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of machine learning research*, 10(2), 2009.

Peisong Wen, Qianqian Xu, Yangbangyan Jiang, Zhiyong Yang, Yuan He, and Qingming Huang. Seeking the shape of sound: An adaptive framework for learning voice-face association. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16347–16356, 2021.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment task for visually-grounded language learning. *arXiv preprint arXiv:1811.10582*, 2018.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding, 2019.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

Baosheng Yu, Tongliang Liu, Mingming Gong, Changxing Ding, and Dacheng Tao. Correcting the triplet selection bias for triplet loss. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 71–87, 2018.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. *arXiv preprint arXiv:2006.16934*, 1:12, 2020.

Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3712–3722, 2018.

Haigang Zhang, Shuyi Li, Yihua Shi, and Jinfeng Yang. Graph fusion for finger multimodal biometrics. *IEEE Access*, 7:28607–28615, 2019.

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*, pp. 2223–2232, 2017.

A Appendix

A.1 REFINER with Multi-similarity Loss

REFINER uses a self-supervised reconstruction loss, but we posited that when used with a Multi-similarity loss to contrast the embeddings associated with different labels, REFINER would create more powerful embeddings. Metric Learning methods have shown strong improvements for multi-class classification problems. When integrated with the self-supervised REFINER module, the metric Refiner network is able to elicit representations of the fused embeddings in a responsible setting, (i.e.) derive distance metrics in the embedding space characterized by strong unimodal and mixed representations of the input features. Figure 6 shows how REFINER Module can be integrated with the Multi-modal fusion module and combined with a Multi-Similarity loss. The latter is supervised, and therefore cannot be used in the pretraining stage. The results showed that generally the Multi-similarity loss can be helpful, especially in few-shot settings but it is not always the case. Plots comparing how $\text{ReFNet}_{\text{MS}}$ provides a boost over ReFNet for the ViLBERT baseline, especially in low shot settings, are illustrated in Fig. 7.

A.2 Effect of the choice of Cost Function

The choice of cost function did not have a drastic impact on the outcome. Cosine similarity and Mean-Squared Error generally converged faster and the difference between them was within the standard deviation and not statistically significant. Contrastive loss converged slightly slower and lagged behind the the other two losses for the same number of epochs by around 0.5%.

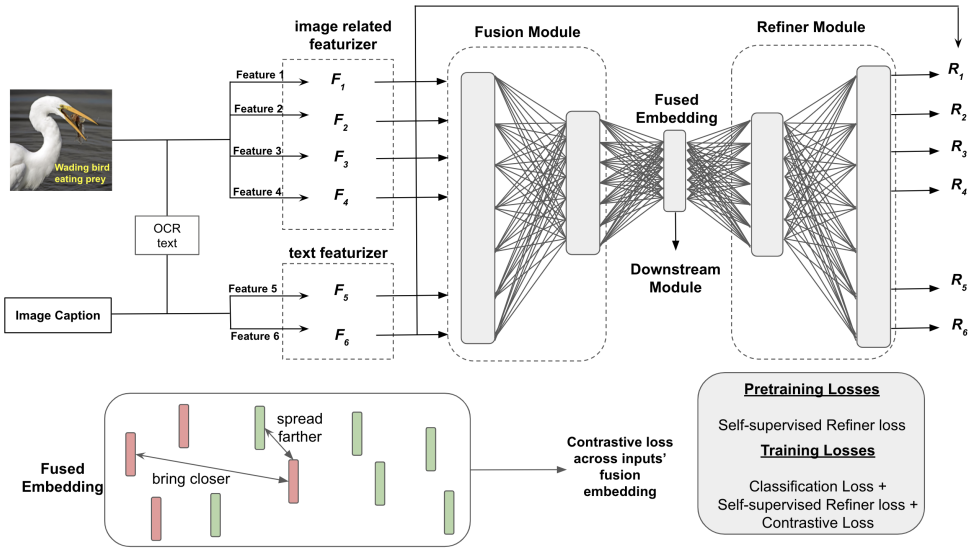

Figure 6: Schematic of the proposed algorithm with a refiner and contrastive loss module. The image based features F_1, F_2, F_3, F_4 and text based features F_5, F_6 are fused together, and a refiner is applied on the fused embedding to generate refiner outputs $R_1, R_2, \cdots, R_6$ which are used to define a self-supervised loss function and a supervised Multi-similarity contrastive loss is also used across samples in a batch.

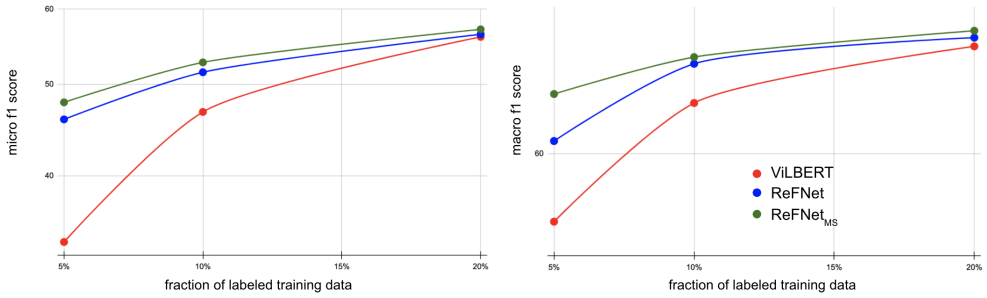

Figure 7: Comparison of performance of the baseline ViLBERT model, ReFNet and ReFNet$_{\text{MS}}$ across datasets with 5%, 10% and 20% of labeled training data available. Figure on the left shows micro and macro f1 scores on the validation dataset and that on the right shows the scores on the test dataset.

A.3 ABLATION STUDY ON THE MULTIMODAL IMDB DATASET

We successively chose a fraction of the labeled dataset in MM-IMDB (5, 10 and 20% of the training data) for downstream classification. The test set was kept intact. The baselines were rerun using the ViLBERT and MMBT models. Metrics were reported on the test dataset based on the model corresponding to the best validation performance on 20000 iterations. With just 5% of labeled dataset available, ReFNet was able to achieve a higher micro f1 score (**+4.03**) and a higher macro f1 score (**+11.83**) compared to the ViLBERT baseline, and a higher micro f1 score (**+4.71**) and a higher macro f1 score (**+6.78**) on the MMBT baseline. ReFNet$_{\text{MS}}$ boosted the performance over baseline to +6.51 micro f1 score and +13.77 macro f1 score for the former. A full summary of the ablation study is provided in Table 6.

model	5% labeled macro-f1	5% labeled micro-f1	10% labeled macro-f1	10% labeled micro-f1	20% labeled macro-f1	20% labeled micro-f1
ViLBERT	34.06	56.62	46.73	62.65	56.08	65.75
ReFNet	45.89	60.65	48.70	63.41	56.43	66.24
ReFNet$_{\text{MS}}$	**47.83**	**63.13**	**51.73**	**64.31**	**57.12**	**66.63**
MMBT	40.49	57.00	48.25	61.77	53.37	66.23
ReFNet	**47.27**	**61.71**	47.34	**62.45**	55.86	66.86
ReFNet$_{\text{MS}}$	42.05	57.47	**48.78**	61.64	**56.37**	**66.94**

Table 6: 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.

A.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 and elicit hidden graphical connections in the data, both across modes within a training sample and across samples in a batch. In Theorem A.1, we show that when an (unknown) adjacency matrix, $\mathbf{A}$, 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 when $k = m$, where m is the number of modalities and k is the rank of the fused embnedding representation (assuming $d > k$).

Theorem A.1. *Let $\mathbf{A}$ be an unknown adjacency matrix and $\mathbf{W}$ be the weights of an affine transformation to generate the fusion embedding, $\mathbf{E}$. The weights Γ of a linear refiner satisfy the property, $\Gamma\mathbf{W}\mathbf{A} = \mathbf{I}_m$.*

Proof. Let $\mathbf{F}^{m\times d}$ represent features where m is the number of modalities and d is the size of each feature vector. We 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, $\mathbf{A}^{m\times m}$ contains ones whenever modality i and j have an edge between them. Let $\mathbf{W}^{k\times m}$ be the unknown coefficients that create the fusion embeddings, $\mathbf{E}^{k\times d}$ from the features as shown in Equation 6.

$$\mathbf{E} = \mathbf{W}\mathbf{A}\mathbf{F} \tag{6}$$

where $\mathbf{E}$ is the embedding in a $k \times d$ space. The fusion module generates weights, $\mathbf{W}^* = \mathbf{W}\mathbf{A}$. The refiner calculates weights $\Gamma^{m\times k}$ such that $\tilde{\mathbf{F}} = \Gamma\mathbf{E} = \Gamma\mathbf{W}\mathbf{A}\mathbf{F}$. Since the refiner network finds weights such that $\tilde{\mathbf{F}} = \mathbf{F}$ (refer Eq. 2 where $\mathcal{D}(\mathbf{x}) = \Gamma\mathbf{x}$ and H_i is identity), we have

$$\mathbf{F} = \Gamma\mathbf{W}\mathbf{A}\mathbf{F}.$$

Since the above holds $\forall\mathbf{F}$, $\Gamma\mathbf{W}\mathbf{A} = \mathbf{I}_m$ where $\mathbf{I}_m$ is an identify matrix of size $m \times m$.

If $\mathbf{W}\mathbf{A}$ is invertible, then $\Gamma = (\mathbf{W}\mathbf{A})^{-1}$. Without using refiner, the weights, $\mathbf{W}^*$ 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. □

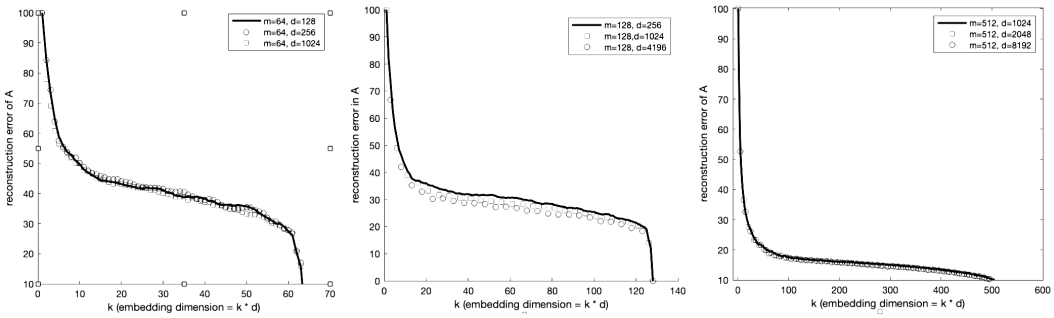

Figure 8: Plot of the reconstruction loss of the adjacency matrix against the dimension of the fused embedding space for different values of d for (top) m=64, (middle) m=128 and (bottom) m=512.

We show in section A.6 the ability of Γ to recover the adjacency matrix when the rank of $\mathbf{E}$ is less than m.

A.5 Experiments with unimodal inputs to drive fusion networks

We paired each image with an irrelevant text input and vice-versa. If the fusion embeddings innately encode unimodality, they should learn to ignore the irrelevant signals and get the strongest unimodal performance. However, in practice, since the multimodal fusion modules generate high dimensional fused representations and the learning geared towards lower dimensional tasks tend to get stuck in local minima, the performance was poorer without an explicit regularizer such as what the REFINER offers. Using the text only features, REFINER was able to boost the micro F1 score from 65.9 to 66.9 (+1%) and macro F1 score from 56.6 to 57.7 (+1.1%) respectively. Using the image only features, REFINER was able to boost the micro-F1 score over MMBT baseline from 47.6 to 48.6 (no change in the micro F1 score). Also, as expected, the REFINER hyper-parameter sweep for the text only inputs revealed an optimal value of 0.1 for text and 0.0 for image, while for the image only inputs revealed an optimal value of 0.2 for image and 0.0 for text. These indicate that the refiner can identify and tune the fused representations to use the relevant modality signal.

A.6 Analysis of Linear Fusion Modules

In order to understand further the implications of the various assumptions in the proof in section A.4, we analyze Linear fusion networks generated with a hidden unknown adjacency matrix.

First, we create a symmetric adjacency matrix comprised of zeros and ones of size $m \times m$ generated randomly. Diagonals comprise of ones. The matrix represents hidden structures/links across modalities. We then take a random feature vector of size d for each modality, and a random set of weights that represents some optimality with respect to a downstream loss (the actual weights are not central to the analysis; we averaged results across 5 randomly chosen sets of weights). The weights reduces the $d \times m$ feature vector into $d \times k$, which represents the fused embedding, with the caveat that k can be significantly less than m. The decoder weights Γ are calculated by taking the product of F with the pseudo-inverse of the fused embeddings. The adjacency matrix is recalculated by multiplying the weights with Γ, and taking its pseudo-inverse. The error norm between the original and reconstructed A were calculated, and averaged across features and weights for each d and m pair.

For the analysis, we first chose three different dimensions for the modality space, m=64, 128 and 512. We chose three different d values starting from $2 \times m$. Figure 8 shows the reduction in reconstruction error as a function of the embedding dimension. When k is equal to m, the matrix is perfectly invertible and we can reconstruct the matrix perfectly. We show in the figures that the errors rapidly decay for the initial values of k and the error is maintained at a much lower value than

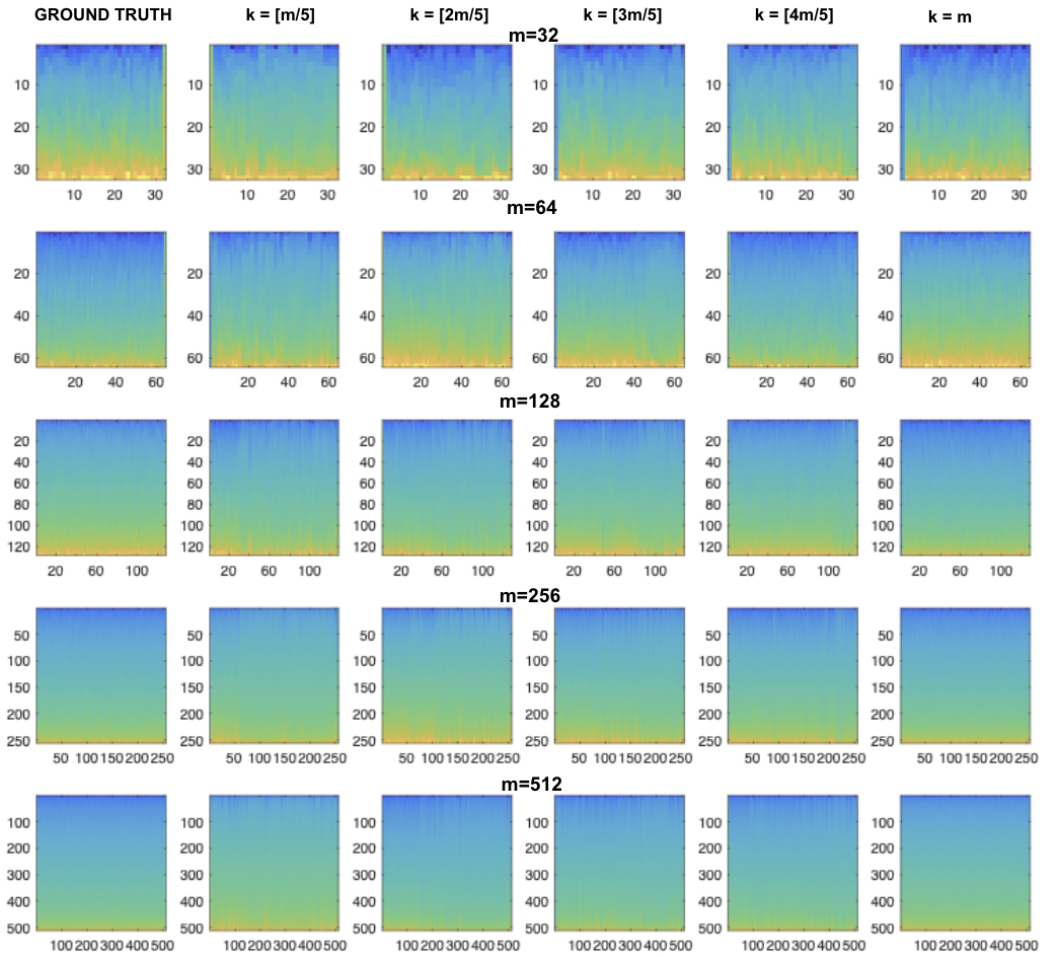

Figure 9: Comparison of eigen-vectors (sorted for visualization purposes) of the ground truth and different dimensions of fusion embeddings, demonstrating that when k=m, Linear Fusion Modules can adequately reconstruct the ground truth.

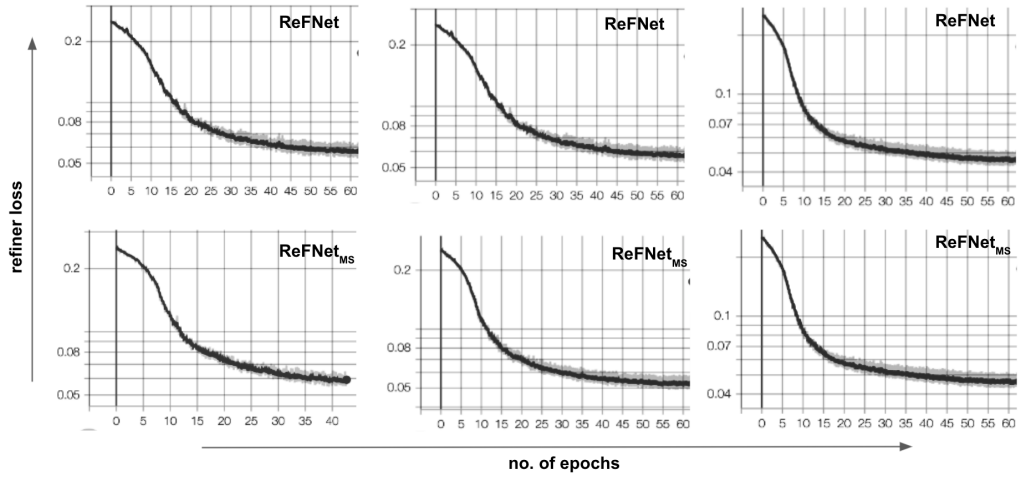

Figure 10: Convergence of ReFNet and ReFNet$_{\text{MS}}$ losses across three different datasets (from left - 10%, 20% and 100% of the training data.

the initial value. This reduction is better for larger m, and for a given m, larger d (though the curves are fairly insensitive to d).

In addition, we show in Fig. 9 that the salient features of the reconstructed adjacency matrix are close to the actual adjacency matrix when k nears m. The figure shows eigenvectors of the matrix (sorted) for visualization of the differences in reconstruction.

A.7 Effect on Training Time Increase

The increase in training time when adding the REFINER module for ReFNet is between 3-5%, and when using ReFNet$_{\text{MS}}$ is between 30-35%. The higher train time when using the Multi-Similarity loss comes from looking at different pairs of similar and dissimilar labels within a batch. If training time is a bottleneck, the gains from the REFINER module alone can be leveraged. As shown in Figure 10, the increase in train time of ReFNet$_{\text{MS}}$ also corresponds to a smaller loss and faster initial rate of convergence. The loss converges faster with larger datasets because each epoch corresponds to a larger set of samples.

A.8 Unimodal retrieval from Multimodal Embeddings

model	Top-1	Top-2	Top-5	Top-10
baseline	23.30	34.26	49.45	59.26
ReFNet	23.66	36.40	53.05	63.82

Table 7: Text-based retrieval of the Multimodal embeddings from the fusion module using unimodal text encoder as the query, and with a weight of 0.2 for the modality refiners.

We performed a study to retrieve target using just the features from the text encoder. The multimodal embeddings were obtained in two settings - (i) using the baseline fusion module as the encoder, and (ii) using a ReFNet. The query for retrieval were the unimodal text encoder features (padded with zeros to project onto the embedding space). Key and value pair for the retrieval were the fused embeddings and the target vectors respectively. Table 7 summarizes the results for the retrieval using Top-1, Top-2, Top-5 and Top-10 k Nearest Neighbors, demonstrating that the REFINER mod-

ule consistently outperforms the baseline fusion module. This further strengthens our claim that REFINER provides better unimodal representation in the fused embedding.