COBRA: Contrastive Bi-Modal Representation Learning

Vishaal Udandarao[1*], **Abhishek Maiti**[1*], **Deepak Srivatsav**[1*], **Suryatej Reddy Vyalla**[1*]
Yifang Yin[2], **Rajiv Ratn Shah**[1]
[1]IIIT-Delhi
[2]National University of Singapore
{vishaal16119, abhishek16005, deepak16030, suryatej16102, rajivratn}@iiitd.ac.in,
yifang@comp.nus.edu.sg

Abstract

There are a wide range of applications that involve multi-modal data, such as cross-modal retrieval, visual question-answering and image captioning. Such applications are primarily dependent on aligned distributions of the different constituent modalities. Existing approaches generate latent embeddings for each modality in a joint fashion by representing them in a common manifold. However these joint embedding spaces fail to sufficiently reduce the modality gap, which affects the performance in downstream tasks. We hypothesize that these embeddings retain the *intra*-class relationships but are unable to preserve the *inter*-class dynamics. In this paper, we present a novel framework COBRA that aims to train two modalities (*i.e.*, image and text) in a joint fashion inspired by the Contrastive Predictive Coding (CPC) and Noise Contrastive Estimation (NCE) paradigms which preserve both inter-class and intra-class relationships. We have conducted extensive experiments on two downstream tasks spanning across three benchmark cross-modal datasets. These show that our proposed framework achieves state-of-the-art results and outperforms existing work, as it generates a robust and task agnostic joint-embedding space.

1 Introduction

Systems built on multi-modal data have been shown to perform better than systems that solely use uni-modal data [Baltrusaitis *et al.*, 2019; Shah and Zimmermann, 2017]. Due to this, multi-modal data is widely used in and generated by different large-scale applications. These applications often utilize this multi-modal data for tasks such as information retrieval [Feng *et al.*, 2014], classification [Tran *et al.*, 2016], and question-answering [Liu *et al.*, 2019; Peng *et al.*, 2019]. It is therefore important to represent such multi-modal data in a meaningful and interpretable fashion to enhance the performance of these large-scale applications. In this work, we focus on learning the joint cross-modal representations for images and text. Learning meaningful representations for multi-modal data is challenging because there exists a distributional shift between different modalities [Peng and Qi, 2019; Hu *et al.*, 2019]. The lack of consistency in representations across modalities further magnifies this shift [Arya *et al.*, 2019]. Due to such difficulties, any similarity metric between the representations across modalities is intractable to compute [Peng and Qi, 2019]. The reduction of this distributional shift boils down to two challenges: (1) projecting the representations of data belonging to different modalities to a common manifold (also referred to as the joint embedding space), and (2) retaining their semantic relationship with other samples from the same class as well as different classes.

The need for a joint embedding space is emphasized by the inability of uni-modal representations to align well with each other. Over the last few years, literature [Peng *et al.*, 2016; Hu *et al.*, 2019; Mai *et al.*, 2019] has been presented where the representations were modeled in the joint embedding space, but existing methods perform less satisfactorily as significant semantic gap still exists among the learnt representations from different modalities.

We believe this is due to the fact that current work such as [Hu *et al.*, 2019; Peng *et al.*, 2016] have focused on conserving the semantic relationship only between *intra* cross-modal data, *i.e.*, belonging to the same class. We surmise that along with this, preserving *inter* cross-modal interactions will help the model learn a more discriminatory boundary between different classes.

Motivation: We posit that preserving the relationship between representations of samples belonging to different classes, in a modality invariant fashion, can improve the quality of joint cross-modal embedding spaces. We formulate this hypothesis as it introduces a contrastive proximity between data belonging to different semantic classes. This will allow the model to converge to a better generalizing decision boundary. Similar *contrastive learning paradigms* based on information gain have been performing very well in the self-supervised learning problem settings [van den Oord *et al.*, 2018; Tian *et al.*, 2019; Hénaff *et al.*, 2019]. *To the best of our knowledge, we are the first to propose a method to learn joint cross-modal embeddings based on contrastive learning paradigms*.

Contributions: Our contributions are as follows:

- We propose a novel joint cross-modal embedding

[*]Equal Contribution, Ordered Randomly

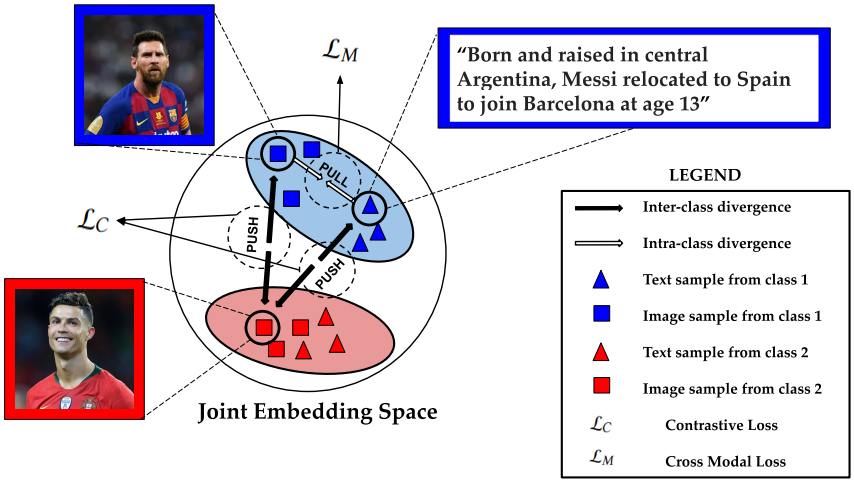


Figure 1: Visualization of the working of contrastive and cross modal losses. The contrastive and cross modal losses enforce divergence across samples of different classes but ensure that the samples of the same class are drawn together, regardless of their modality. This ascertains that the joint embedding space is both modality invariant and class discriminative.

framework called COBRA (COntrastive Bi-modal Representation Algorithm) which represents the data across different modalities (text and image in this study) in a common manifold.

- We formulate a combined loss function, which jointly preserves not only the relationship between different *intra* cross-modal data samples but also preserves the relationship between *inter* cross-modal data samples (refer Figure 1).
- We empirically validate our model by achieving state-of-the-art results on two downstream tasks: (1) finegrained multi-modal sentiment classification, and (2) multi-modal fake news detection.

2 Related Work

In this section, we discuss the topics that inspire the architecture and loss functions used in COBRA: Multi-modal Fusion and Contrastive Learning Paradigms.

2.1 Multi-modal Fusion

Significant amount of work in the domain of multimedia research has been based on fusion techniques for datasets of multiple modalities. The type of fusion affects the dynamics of the features produced. Early fusion techniques that are based on simple concatenation [Wöllmer *et al.*, 2013; Poria *et al.*, 2016] do not capture the intra modal relations well. Late fusion techniques [Nojavanasghari *et al.*, 2016; Kampman *et al.*, 2018] on the other hand prioritize intra modal learning abilities compromising on cross-modal differentiability. This is because these models make decisions on a weighted average score of individual modality features. To solve both these limitations, [Mai *et al.*, 2019] used a hierarchical graph neural network to capture multi-modal interactions. Fusion networks have also shown great performance in application specific tasks. [Ding *et al.*, 2019] proposed a fusion based DNN for predicting popularity on social media. However, literature suggests that cross modal tasks benefit more from learning a joint embedding space than employing multi-modal fusion techniques [Baltrusaitis *et al.*, 2019].

2.2 Contrastive Learning Paradigms

Contrastive Learning techniques have gained popularity recently because of their success in unsupervised settings. [van den Oord *et al.*, 2018] were one of the first to propose a Contrastive Predictive Coding (CPC) technique that could generate useful representations from high dimensional data universally in an unsupervised fashion. Further, [Tian *et al.*, 2019] developed a compact representation that maximized mutual information between different views of the same scene and hence improved performance on image and video unsupervised learning tasks. SimCLR [Chen *et al.*, 2020] eliminated the requirement of specialized architectures or memory banks for contrastive tasks and also gave state-of-the-art results on self-supervised classification tasks. All these techniques proposed so far have been employed only for single modality tasks.

3 Methodology

In this section, we first explain the formulation of our problem statement in terms of the data we use. We then introduce and explain the architecture of our model, along with the loss functions used. We finally explain our optimization and training strategy.

3.1 Problem Formulation

Let us assume that we have two modalities, *i.e.* text and image, we denote the j-th image sample as $x_I^j \in R^{d_I}$ and the j-th text sample as $x_T^j \in R^{d_T}$. Here, d_I and d_T represent the dimensionality of the image and text samples respectively. We denote the image dataset as $X_I = \{x_I^0, x_I^1, ..., x_I^{n_I-1}\}$ and the text dataset as $X_T = \{x_T^0, x_T^1, ..., x_T^{n_T-1}\}$, where n_I and n_T denote the total number of data samples in the image and

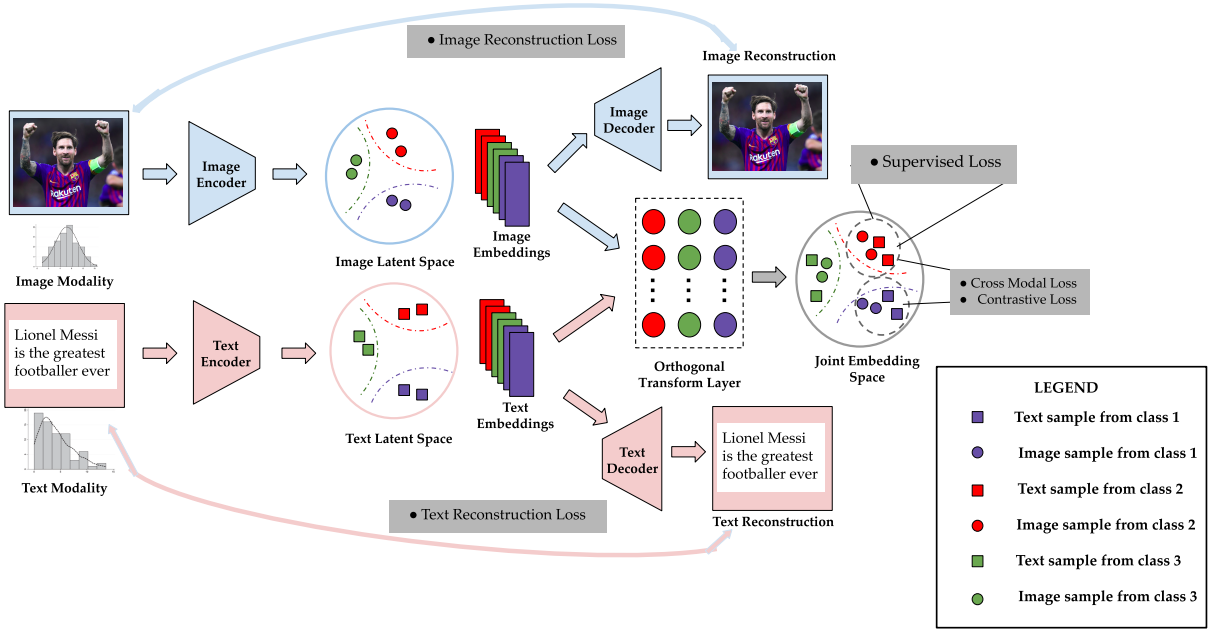


Figure 2: The general architecture of our proposed COBRA model. The different shapes are present to help visualize the structure of the joint embedding space with respect to the orthogonal projections of both image and text samples. The losses of our model are highlighted in the grey boxes.

text datasets respectively. The corresponding labels for the image and text modalities are represented as follows: $Y_I = [y_I^0, y_I^1, ..., y_I^{n_I-1}]$ and $Y_T = [y_T^0, y_T^1, ..., y_T^{n_T-1}]$. Assuming there are C distinct semantic classes in our multi-modal dataset, the labels are: $y_I^{j_I}, y_T^{j_T} \in \{0, 1, ..., C-1\} \forall j_I \in \{0, 1, ..., n_I - 1\}, j_T \in \{0, 1, ..., n_T - 1\}$.

3.2 Model Architecture

The overall architecture for our model is given in Figure 2. Our goal is to represent the data in a common manifold, such that the class-wise representations are modality invariant and discriminatory. To this end, we use an autoencoder for each modality to generate representations that are high fidelity in nature. We utilize an orthogonal transform layer, which takes as input the hidden space representations from the encoders of each modality, and projects these representations into a joint space that is modality invariant and discriminates between classes well.

We denote the encoded representation as $z_i^j = f_i(x_i^j, \Theta_i)$ and the reconstructed sample as $\hat{x}_i^j = g_i(z_i^j, \Phi_i)$ where $j \in \{0, 1, ..., n_T - 1\}$ and $j \in \{0, 1, ..., n_I - 1\}$ for text and image respectively, and where $i \in \{T, I\}$ for text and image respectively. f_i denotes the encoder of the i-th modality parameterised by Θ_i. Similarly g_i denotes the decoder of the i-th modality parameterised by Φ_i. Given the representations z_T^j and z_I^j, which have dimensions Z_T and Z_I, we project the representations to a joint subspace such that the representation of each semantic class is orthogonal to each other [Hu *et al.*, 2019]. We call these projections O_T^j and O_I^j, both of which have dimension Z.

We define the loss function in COBRA as a weighted sum of the reconstruction loss, cross-modal loss, supervised loss and contrastive loss, the details of which are introduced below. To preserve the inter-class dynamics, we innovatively introduce the *Contrastive Loss* that has never been used in representing multi-modal data.

Reconstruction Loss

Given the decoder output $\hat{x}_i^j$ and the input x_i^j, we define the reconstruction loss shown in Eq. 1 as:

$$\mathcal{L}_R = \sum_{i \in \{I,T\}} \sum_{j=0}^{n_i - 1} \|\hat{x}_i^j - x_i^j\|_2^2 \tag{1}$$

Cross-Modal Loss

The projected representations O_I^j and O_T^j align class representations within each modality. The cross-modal loss aims to align representations of the same class across different modalities. Given the projected representations O_I^j and O_T^i, we define the cross-modal loss shown in Eq. 2 as:

$$\mathcal{L}_M = \sum_{j=0}^{\min\{n_T, n_I\} - 1} \|O_T^j - O_I^j\|_2^2 \tag{2}$$

We use the $\min$ function because the dataset may not have equal text and image samples. We only take those pairs in which the corresponding text and image samples are present.

Supervised Loss

As we try to model an orthogonal latent space having the joint embeddings, we utilize the one-hot labels of the data samples to reinforce those samples belonging to the same class but different modalities to be grouped together in the same subspace. Let $\hat{y}_i^j$ be the one-hot encoded label for the j-th sample of the i-th modality, and O_i^j be the projected representation, we then define the supervised loss shown in Eq. 3 as:

$$\mathcal{L}_S = \sum_{i \in \{I,T\}} \sum_{j=0}^{n_i - 1} \| O_i^j - \hat{y}_i^j \|_2^2 \tag{3}$$

Contrastive Loss

As stated in recent literature [Tian *et al.*, 2019; Arora *et al.*, 2019], to implement the contrastive loss [Gutmann and Hyvärinen, 2010; Sohn, 2016], the definitions of positive samples and negative samples of representations are of utmost importance. We will first define the positive and negative samples pertaining to our model. Given the projected representations O_I^j and O_T^j, a positive pair is defined as the representations of data samples belonging to the same modality and class. A negative pair is defined as the representations of two data samples belonging to same or different modality of different classes. To define the contrastive loss, a scoring function is required, which yields high values for positive samples and low values for negative values. Here we define the scoring function by taking the dot product of the representations in the joint embedding space. Following recent works [van den Oord *et al.*, 2018; Chen *et al.*, 2020], our loss function enforces the model to select the positive sample from a fixed sized set $S = \{p, n_1, n_2, ..., n_N\}$ containing one positive and N negative samples. Thereafter we formulate our contrastive loss shown in Eq. 4 as:

$$\mathcal{L}_C = -E_S \left[\log \frac{a^T p}{a^T p + \sum_{i=1}^{N} a^T n_i} \right] \tag{4}$$

where a is the anchor point, p is its corresponding positive sample, E is an expectation operator over all possible permutations of S and n_i iterates over all the negative samples. The anchor, positive and negative samples are randomly drawn from each mini-batch. We minimize the above expectation running over all samples. Since fetching negative samples from the entire dataset is computationally infeasible, we sample the negative points only from each mini-batch locally.

Since, we sample only a finite sized set of negative samples, the model can miss out on characteristics of the distribution of the joint embeddings. To avoid this, we implement another loss called the Noise Contrastive Estimation (NCE) loss [Gutmann and Hyvärinen, 2010], which is an effective method for estimating unnormalized models. NCE helps to model the distribution of the negative samples by leveraging a proxy noise distribution. It does so by estimating the probability of a sample coming from a joint distribution rather than it coming from a noise distribution. The noise distribution is assumed to be a uniform distribution. Denoting the joint distribution of positive samples as p_J, the noise distribution as p_N, the anchor sample as a and every other sample (can be either positive or negative) as s, the probability of data sample s coming from the joint distribution p_J is:

$$P(X = 1|s; a) = \frac{p_J(s|a)}{p_J(s|a) + N p_N(s|a)} \tag{5}$$

where N is the number of samples from the noise distribution. Instead of using Eq. 4, now we can estimate the contrastive loss more accurately based on Eq. 6 as follows:

$$\begin{aligned} \mathcal{L}_C = -E_a \{ E_{s \sim p_J(\bullet|s)} [[P(X = 1|s; a)] \\ + N \times E_{s \sim p_N(\bullet|s)} [1 - P(X = 1|s; a)] \} \end{aligned} \tag{6}$$

where E_a is an expectation over all possible anchor samples, $E_{s \sim p_J}$ is an expectation over all possible positive samples (corresponding to anchor a) from the joint distribution p_J, and $E_{s \sim p_N}$ is an expectation over all samples from the noise distribution p_N.

3.3 Optimization and Training Strategy

The overall loss of our network is defined to be a weighted sum of the reconstruction loss, cross-modal loss, supervised loss and contrastive loss. The weights are treated as hyperparameters.

$$\mathcal{L} = \lambda_R \mathcal{L}_R + \lambda_S \mathcal{L}_S + \lambda_M \mathcal{L}_M + \lambda_C \mathcal{L}_C \tag{7}$$

The objective function in Eq. 7 is optimized using stochastic gradient descent. The loss is summed over all modalities, and the corresponding gradient is propagated through all the components in the model. We adopted the PyTorch framework for implementation, and trained all our models for 200 epochs on an Nvidia RTX 2080Ti GPU.

4 Experiments

To evaluate our proposed method, we test our model on two tasks, namely, multi-modal fake news detection and multi-modal sentiment classification. We compare the performance of our model against state-of-the-art models of corresponding tasks.

In the following sections, we describe the datasets and evaluation metrics adopted, followed by the results achieved on each downstream task mentioned above.

4.1 Multi-modal Fake News Detection

In the task of multi-modal fake news detection, we use COBRA to determine whether a given bi-modal query (text and image) corresponds to a real or fake news sample.

Datasets

For this multi-modal task, we utilize the FakeNewsNet Repository [Shu, 2019]. This repository contains two datasets, namely, Politifact and Gossipcop. These datasets contain news content, social context, and dynamic information. We pre-process the data similar to Singhal *et al.* 2020. Each dataset contains two semantic classes, namely, Real and Fake.

Table 1: Accuracy on the FakeNewsNet dataset

Method	Politifact (%)	Gossipcop (%)
Wang *et al.* [2018]	74	86
Khattar *et al.* [2019]	67.3	77.5
Singhal *et al.* [2019]	72.1	80.7
Singhal *et al.* [2020]	84.6	85.6
COBRA[1]	**86**	**86.7**

- The Politifact dataset contains 1056 text-image pairs. We get 321 Real and 164 Fake text-image pairs after pre-processing. We use a training, validation and test set of 381, 50 and 54 text-image pairs [Singhal *et al.*, 2020] respectively.
- The Gossipcop dataset contains 22140 text-image pairs. We get 10259 Real and 2581 Fake text-image pairs after pre-processing. We use a training, validation and test set of 10010, 1830 and 1000 text-image pairs Singhal *et al.* [2020] respectively.

Evaluation metrics

We compare our performance against existing state-of-the-art models based on number of correctly classified queries (accuracy). For the purpose of our evaluation, we ensure that we use the same features that were used across other existing state-of-the-art models.
To visualize the purity of the joint embedding space for different classes and modality samples, we plot the joint embeddings of COBRA trained on both the Gossipcop and Poltifact datasets. We plot the embeddings (Figure 3a and 3b) by employing the t-SNE transformation to reduce the high dimensional joint embeddings (O_I and O_T) to 2 dimensional data points. The figures clearly exhibit the high discrimination between samples of different classes in the joint embedding space. This provides further empirical validation for the high class divergence across the joint embedding space, irrespective of the modalities of the data points.

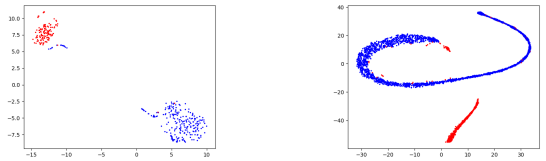
Results

We achieve a 1.4% and a 1.1% improvement over the previous state-of-the-art (SpotFake+ [Singhal *et al.*, 2020]) on the Politifact and Gossipcop dataset respectively (Table 1). On observing the t-SNE plots in Figure 3, we discern a high intra-class variability in the Gossipcop dataset. We believe that there is only a small improvement because of the high class imbalance in these two datasets.

4.2 Multi-modal Fine-grained Sentiment Classification

In the task of multi-modal fine-grained sentiment classification, we use COBRA to perform ten tasks of classifying a given bi-modal query (text and image) into a sentiment category.

[1]Comparison with contemporary work [Zhou *et al.*, 2020] is left for future work as their current results are on a different data split.



(a) Politifact (b) Gossipcop

Figure 3: t-SNE visualizations of the joint embedding spaces of the models trained on Gossipcop and Politifact datasets. The different colours correspond to the various class labels in the dataset.

Datasets

For this task, we analyze the performance of our model on the MeTooMA dataset [Gautam *et al.*, 2019]. This dataset contains 9973 tweets that have been manually annotated into 10 classes. We use a training, validation and test set of 4500, 1000 and 1000 text-image [Gautam *et al.*, 2019] respectively, across all models that we test.

Evaluation Metrics

We report the number of correctly classified queries (accuracy). To the best of our knowledge, we are the first to test a multi-modal classification model on this dataset. To this end, we evaluate our model against a Text-only and Image-only baseline, and Early Fusion. For the baselines, we use a Fully Connected network.

Results

We obtain an average classification accuracy of 88.32% across all classes on the MeTooMA Dataset. This is a 1.2% improvement over Early Fusion (Table 2). We observe a low increase in Text only and Image only informative tasks due to the fact that 53.2% of our training data had text-image pairs with conflicting labels, *i.e.*, from a given text-image pair, the text may be labelled as "relevant" whereas the corresponding image may be labelled as "irrelevant". Furthermore, for classes under the Hate Speech, Sarcasm, and Dialogue Acts categories, we observe that there are less than 600 samples for each class. In categories such as Stance, where the 'Support' class has over 3000 samples, we observe much larger improvements in performance.

4.3 Ablation Study

Role of the Loss Functions.

To robustly evaluate the importance of each loss function in our training objective, we conduct a simple ablation study. In Table 3, we take our best performing models trained on the Politifact and the Gossipcop datasets and dissect each of the loss functions. To ensure an unbiased setup, we decouple the task in this ablation study from the previous tasks. We conduct a simple cross-modal retrieval task on the two datasets, wherein given a text query we retrieve an image and vice-versa, and use mAP (mean average precision) as our evaluation metric. It is immediately evident that supervision plays a very important role for the model to perform well. We further see that the contrastive and cross-modal losses help us gain

Table 2: Accuracy on the MeTooMA Dataset

Label	**COBRA** (%)	Text-only baseline (%)	Image-only baseline (%)	Early Fusion (%)
Text only informative	**73.77**	73.43	63.39	72.15
Image only informative	67.36	63.21	**67.74**	66.97
Directed Hate	**96.43**	95.12	94.67	95.85
Generalized Hate	**97.77**	96.19	95.89	96.88
Sarcasm	**98.55**	96.94	96.45	97.16
Allegation	**93.75**	92.67	92.40	93.19
Justification	**98.44**	96.23	95.66	97.34
Refutation	**98.54**	96.90	96.81	97.37
Support	**66.29**	61.60	59.93	63.28
Opposition	**92.3**	90.1	89.5	91.1
Average	**88.32**	86.23	85.24	87.12

the best performing models. Further, we see that the model is especially sensitive to the reconstruction loss without which performance drops significantly.

Table 3: Ablation Study of loss functions. (-) indicates that particular loss is removed from the overall loss function.

Model	Politifact	Gossipcop
COBRA	**79.07**	**76.793**
(-) $\mathcal{L}_C$	78.63	76.37
(-) $\mathcal{L}_M$	77.51	71.05
(-) $\mathcal{L}_R$	75.68	69.92
(-) $\mathcal{L}_S$	74.95	69.46

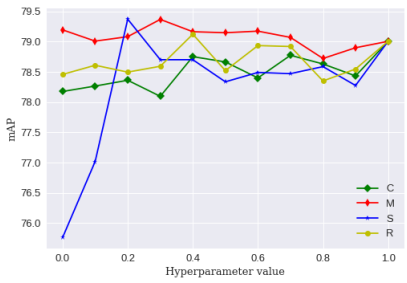


Figure 4: Hyperparameter Selection study on the Politifact dataset. The labels are: C - contrastive loss, M - cross-modal loss, S - supervised loss, R - reconstruction loss.

Hyperparameter Selection.

We evaluate the significance of the individual weights of each loss function in our overall objective. To ensure a consistent setting, for each loss term, we vary its corresponding weight from 0 to 1 with intervals of 0.1 while keeping the weights of the other loss terms as 1 (refer Figure 4). We ensure the same cross-modal retrieval task setup as in 4.3 to ensure fairness across experiments. It is clear that the contrastive, cross-modal and reconstruction loss terms are robust to large variations in their weight hyperparameter. However, we see that the supervised loss requires a higher weight (> 0.2) for the model to perform well. This again highlights the major significance of supervision required for our method to accomplish effective results.

5 Conclusion

In this paper, we propose a novel approach (COBRA) to jointly learn bi-modal representations in an orthogonal space. We show that our proposed method learns better representations which allows the model to generalize across tasks in a much more robust fashion. This enables us to achieve state-of-the-art results on two downstream tasks. The representations learnt are high-fidelity in nature, containing sufficient information for reconstruction as well as tasks such as retrieval and classification. Different from other models, COBRA, along with preserving the $intra$-class relationship of samples in the embedding space, also preserves the $inter$-class relationships. This ensures that the samples belonging to the same class are clustered together, and that the distance between clusters of samples belonging to different classes (irrespective of the modality) is maximized in the joint embedding space. In the future, we would like to extend our method to a self-supervised/semi-supervised problem setting, and to complex tasks such as image captioning.

References

Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019.

Devanshu Arya, Stevan Rudinac, and Marcel Worring. Hyperlearn: A distributed approach for representation learning in datasets with many modalities. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2245–2253, 2019.

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2):423–443, February 2019.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020.

Keyan Ding, Ronggang Wang, and Shiqi Wang. Social media popularity prediction: A multiple feature fusion approach with deep neural networks. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 2682–2686, New York, NY, USA, 2019. Association for Computing Machinery.

Fangxiang Feng, Xiaojie Wang, and Ruifan Li. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the 22nd ACM International Conference on Multimedia*, MM '14, page 7–16, New York, NY, USA, 2014. Association for Computing Machinery.

Akash Gautam, Puneet Mathur, Rakesh Gosangi, Debanjan Mahata, Ramit Sawhney, and Rajiv Ratn Shah. #metooma: Multi-aspect annotations of tweets related to the metoo movement, 2019.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304, 2010.

Peng Hu, Liangli Zhen, Dezhong Peng, and Pei Liu. Scalable deep multimodal learning for cross-modal retrieval. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 635–644, New York, NY, USA, 2019. Association for Computing Machinery.

Olivier J. Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding, 2019.

Onno Kampman, Elham J. Barezi, Dario Bertero, and Pascale Fung. Investigating audio, video, and text fusion methods for end-to-end automatic personality prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 606–611, Melbourne, Australia, July 2018. Association for Computational Linguistics.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. Mvae: Multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference*, WWW '19, page 2915–2921, New York, NY, USA, 2019. Association for Computing Machinery.

Fei Liu, Jing Liu, Richang Hong, and Hanqing Lu. Erasing-based attention learning for visual question answering. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1175–1183, 2019.

Sijie Mai, Haifeng Hu, and Songlong Xing. Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion, 2019.

Behnaz Nojavanasghari, Deepak Gopinath, Jayanth Koushik, Tadas Baltrušaitis, and Louis-Philippe Morency. Deep multimodal fusion for persuasiveness prediction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI '16, page 284–288, New York, NY, USA, 2016. Association for Computing Machinery.

Yuxin Peng and Jinwei Qi. Cm-gans: Cross-modal generative adversarial networks for common representation learning. *ACM Trans. Multimedia Comput. Commun. Appl.*, 15(1), February 2019.

Yuxin Peng, Xin Huang, and Jinwei Qi. Cross-media shared representation by hierarchical learning with multiple deep networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, page 3846–3853. AAAI Press, 2016.

Liang Peng, Yang Yang, Zheng Wang, Xiao Wu, and Zi Huang. Cra-net: Composed relation attention network for visual question answering. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1202–1210, 2019.

S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain. Convolutional mkl based multimodal emotion recognition and sentiment analysis. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 439–448, 2016.

Rajiv Shah and Roger Zimmermann. *Multimodal analysis of user-generated multimedia content*. Springer, 2017.

Kai Shu. FakeNewsNet, 2019.

S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, and S. Satoh. Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, pages 39–47, 2019.

Shivangi Singhal, Anubha Kabra, Mohit Sharma, Rajiv Ratn Shah, Tanmoy Chakraborty, and Ponnurangam Kumaraguru. Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract). 2020.

Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in neural information processing systems*, pages 1857–1865, 2016.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding, 2019.

Thi Quynh Nhi Tran, Hervé Le Borgne, and Michel Crucianu. Cross-modal classification by completing unimodal representations. In *Proceedings of the 2016 ACM Workshop on Vision and Language Integration Meets Multimedia Fusion*, iV&;L-MM '16, page 17–25, New York, NY, USA, 2016. Association for Computing Machinery.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 849–857, New York, NY, USA, 2018. Association for Computing Machinery.

M. Wöllmer, F. Weninger, T. Knaup, B. Schuller, C. Sun, K. Sagae, and L. Morency. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems*, 28(3):46–53, 2013.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. Safe: Similarity-aware multi-modal fake news detection. *arXiv preprint arXiv:2003.04981*, 2020.