Domain-aware Self-supervised Pre-training for Label-Efficient Meme Analysis

Shivam Sharma[1,4], Mohd Khizir Siddiqui[3], Md. Shad Akhtar[1] and Tanmoy Chakraborty[2]
[1]Indraprastha Institute of Information Technology Delhi, India
[2]Indian Institute of Technology Delhi, India
[3]Birla Institute of Technology and Science, Goa, India
[4]Wipro AI Labs, India
{shivams, shad.akhtar}@iiitd.ac.in, mdkhizirsiddiqui@gmail.com, tanchak@ee.iitd.ac.in

Abstract

Existing self-supervised learning strategies are constrained to either a limited set of objectives or generic downstream tasks that predominantly target uni-modal applications. This has isolated progress for imperative multi-modal applications that are diverse in terms of complexity and domain-affinity, such as meme analysis. Here, we introduce two self-supervised pre-training methods, namely Ext-PIE-Net and MM-SimCLR that (i) employ off-the-shelf multi-modal hate-speech data during pre-training and (ii) perform self-supervised learning by incorporating multiple specialized pretext tasks, effectively catering to the required complex multi-modal representation learning for meme analysis.

We experiment with different self-supervision strategies, including potential variants that could help learn rich cross-modality representations and evaluate using popular linear probing on the Hateful Memes task. The proposed solutions strongly compete with the fully supervised baseline via label-efficient training while distinctly outperforming them on all three tasks of the Memotion challenge with 0.18%, 23.64%, and 0.93% performance gain, respectively. Further, we demonstrate the generalizability of the proposed solutions by reporting competitive performance on the HarMeme task. Finally, we empirically establish the quality of the learned representations by analyzing task-specific learning, using fewer labeled training samples, and arguing that the complexity of the self-supervision strategy and downstream task at hand are correlated. Our efforts highlight the requirement of better multi-modal self-supervision methods involving specialized pretext tasks for efficient fine-tuning and generalizable performance.

1 Introduction

The overwhelming scale of digital mutation constantly transpiring over the web is “creating the illusion of reality, addressing the viewer, and representing a convoluted space" (Manovich, 2001). Almost every social activity affects or is affected by an online entity, sometimes even disturbing social harmony, influenced by a prominent surge of multi-modal harmful, abusive and hateful online content. Therefore, it is imperative to explore solutions towards automatic mediation of online activities that pre-dominantly involve multi-modality. Recently, there has been a defining resurgence of advancements in multi-modal AI, albeit slowly.

Existing self-supervision strategies for visual-linguistic applications involve different *pretext* tasks like Masked Language Modeling (MLM) (Devlin et al., 2019), Masked Region Modeling (MRM) (Chen et al., 2020b), Word-Region Alignment (WRA) (Gupta et al., 2017), and Image-Text Matching (ITM) (Li et al., 2019a; Radford et al., 2021), which inherently presume visual-linguistic grounding (Karpathy and Fei-Fei, 2017). As a consequence, the large-scale datasets like MS COCO (Lin et al., 2014), Conceptual Captions (CC) (Sharma et al., 2018), Wikipedia-based Image Text (WIT) (Srinivasan et al., 2021) and LAION-400M (Birhane et al., 2021), curated towards the required pre-training, are either mostly generic in nature or represent a greater degree of visual-semantic association between the image and text pairs. Moreover, the required multi-modal datasets are rather challenging to create, as they often require multi-dimensional and fine-grained manual annotations for a large volume of multi-modal data.

These frameworks have demonstrated impressive pre-training schemes for addressing downstream multi-modal tasks like Visual Question Answering (VQA), Image Captioning (IC), Visual Commonsense Reasoning (VCR), etc. (Mogadala et al., 2021). Still, there is significant room for improvement in terms of their generalizability. For instance, besides *masked language modelling* (MLM), state-of-the-art multi-modal models like

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 792–805
November 20–23, 2022. ©2022 Association for Computational Linguistics

Visual BERT, ViLBERT and LXMERT are pre-trained wrt pretext tasks like *sentence-image prediction* (Li et al., 2019b), *masked multi-modal learning, multi-modal alignment prediction* (Lu et al., 2019a) and *detected-label classification* (Tan and Bansal, 2019), which presume aspects like availability of multiple *semantically grounded* sentences corresponding to an image and visual-semantic object and pixel-level annotations for the images. These requirements constrain modeling aspects for multi-modal content like *memes*. Although such approaches address the issue of scale and cross-modal alignment in terms of *common-sense* reasoning extremely well, they tend to fall short on performance for complex multi-modal tasks like meme analysis (Chen et al., 2020a; Kiela et al., 2020). This is because memes *do not* represent strong visual-linguistic grounding and solicit sophisticated multi-modal fusion along with contextual knowledge integration.

This paper presents the design and evaluation of efficient multi-modal frameworks that do not rely upon large-scale dataset curation and annotation and can be pre-trained using the datasets from the wild. Also, the pre-training employed is optimally designed toward learning enriched multi-modal representations, which can be further used for addressing downstream tasks like meme analysis in a label-efficient manner. Our contributions, as enlisted below, are three-fold:

1. We propose two self-supervision-based multi-modal pre-training frameworks which learn semantically rich cross-modal features for meme analysis.
2. We empirically establish the efficacy of the proposed self-supervision frameworks towards adapting to downstream tasks using only a few labeled training samples.
3. We finally demonstrate the generalizability of the representations learned across tasks and datasets.[1]

2 Related Work

Self-supervised and Semi-supervised Learning: Self-supervised learning approaches are formulated to optimize training objectives that do not require an explicit set of labels. They incorporate pretext tasks to introduce pseudo-labels and learn embedding space rather than solving a specific downstream task. One of the prominent pretext tasks for pre-training language models is next word prediction using a part of the sentence (Peters et al., 2018). ALBERT (Lan et al., 2020) performs sentence order prediction (SOP) to achieve a similar objective.

Although self-supervision has taken long strides for NLP applications, it has taken a while to show promise for vision applications. A prominent series of work aims at optimizing the similarity between positive pairs of augmented representations while reducing it for negative pairs (Oord et al., 2018), (Chen et al., 2020a), also known as contrastive learning. A non-contrastive learning approach increases similarity with the previous versions of augmented views (Grill et al., 2020). Such works have long been attempting to solve problems about specific modalities only. We aim to learn multi-modal embedding space enriched to solve non-trivial downstream tasks.

Multi-modal Pre-training: Recently, Wang et al. (2021) proposed a simple yet effective multi-modal system with specialized convolution layers at the beginning of the encoder and a textual decoder as a follow-up. Other recent similar works include DALL-E (Ramesh et al., 2021), a zero-shot, generative scalable Transformer that models multi-modal information in an auto-regressive manner and is conditioned on a textual query. This is followed by CLIP, a contrastive learning-based model (Radford et al., 2021), which is pre-trained on 400 million image-text pairs collected from different web-based resources. The primary objective of such efforts is to learn multi-modal embedding space jointly. However, the datasets used to pre-train are too generic to capture complex semantics. In this work, we intend to examine such constraints and their impact on the performance of multi-modal systems.

Studies on Memes: Although the recent past has witnessed an overwhelming amount of research related to memes, especially for topics like online hate, harm, offense, abuse, etc. (Kiela et al., 2020; Sharma et al., 2020), still, there are a wide array of meme related tasks, that are yet to be addressed. Kolawole (2015) explored the classification task on a small dataset and with a linear SVM on low-level descriptors, leveraging only visual information. Significant efforts have been invested towards meme generation by representing the meme image and the catchphrase in the same vector space

[1] The source codes are uploaded as supplementary material.

using a deep neural network (Kido Shimomoto et al., 2019), leveraging pre-trained Inception-v3 network-based feature extraction. This was further explored in (Peirson et al., 2018) for caption generation and rule-based classification. The human assessment in this study outperformed random choices. The quality, however, was below-par as compared to human-produced memes. Efforts are solicited wherein richer and more meaningful content modeling is achieved towards solving tasks that conventional multi-modal approaches cannot.

3 Dataset

Pretraining: To address generalizability towards an array of such topics, we employ the MMHS150K dataset (Gomez et al., 2020) as our primary data source for pre-training our proposed systems. It consists of 150K multi-modal (images + text) tweets spanning over four hate-inclined topics – *racism*, *sexism*, *homophobia*, and *religious extremism*. Moreover, the images in the dataset represent diversity with the presence of memes, morphed images, satirical art, etc.

Besides this, to ensure that our pre-training dataset reasonably represents the content type we would evaluate as part of downstream tasks, we also add the memes from the training split of the Facebook's Hateful Memes dataset (Kiela et al., 2020), that we reserve exclusively for our pre-training.

Training and Evaluation: We employ three datasets (Hateful Memes, Harm-P, and Memotion) and five different tasks (*hate detection*, *harmfulness detection*, *sentiment analysis*, *emotion classification*, and *emotion class quantification*) to demonstrate the efficacy of our proposed approaches. The Harm-P dataset belongs to the HarMeme task (Pramanick et al., 2021) and consists of 3552 memes annotated with two labels – *harmful or not-harmful*. The Memotion dataset (Sharma et al., 2020) has approx. 8K memes and defines three subtasks[2] – *sentiment analysis* (positive/negative), *emotion classification* (humour/sarcasm/offense/motivational), and *emotion class quantification* (slightly/mildly/very). Although these datasets are based on memes or multi-modal content, their objectives are different and have *varying* complexities. [3].

We leverage a dataset that represents the raw, unprocessed large-scale corpus of multi-modal information, specifically emphasizing different types of hate speech. We acknowledge that a labeling scheme initially accompanies the dataset (MMHS150K). However, we do not utilize that information either during the pre-training stage or during the task-specific fine-tuning stage. This is also represented in the form of proposed loss functions, which do not utilize source data labels but solely rely on the intermediate neural representations, hence self-supervised. Also, the underlying presumption for utilizing such a dataset (MMHS150K) in a self-supervised way is based on the fact that the original dataset owners collected it using a pre-defined set of database keywords (Gomez et al., 2020), and this is all that one would need to do to obtain such a dataset at scale towards pre-training the models proposed. Also, no explicit annotation process is required for pre-training MM-SimCLR and Ext-PIE-Net. Now, as for the task-specificity, we already showcase the performances of the fully supervised systems that utilize fine-tuning of the models, pre-trained using a generic dataset. We propose the frameworks that, if pre-trained using a "domain-oriented" dataset that can be easily obtained, without any special annotations, can quickly and in a label-efficient way adapt to related downstream tasks.

4 Proposed Solution

We propose two methods: MM-SimCLR and Ext-PIE-Net, that utilize adaptations of popular contrastive and triplet loss formulations for learning multi-modal embedding space. Proposed solutions also encapsulate specialized multi-modal pretext tasks suited toward joint multi-modal representation learning. Before describing the proposed solutions, we first review the two-loss formulations below.

- *SimCLR:* The SimCLR framework (Chen et al., 2020a), a popular self-supervision technique, learns representations for images by maximizing agreement between their augmented views in a latent space. The objective function is defined as:

$$\mathcal{L}^{\text{NT-Xent}}_{(i,j)} = -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq i]} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)} \quad (1)$$

where $\mathbb{1}_{[k\neq i]} \in \{0, 1\}$ is an indicator function; $\mathbf{z}_i$

[2] We use abbreviations SENT, EMOT and EMOT-Q for *sentiment analysis*, *emotion classification*, and *emotion class quantification*, respectively.

[3] We present further details like lexical characteristics and text-length comparison for the datasets used in App. B.

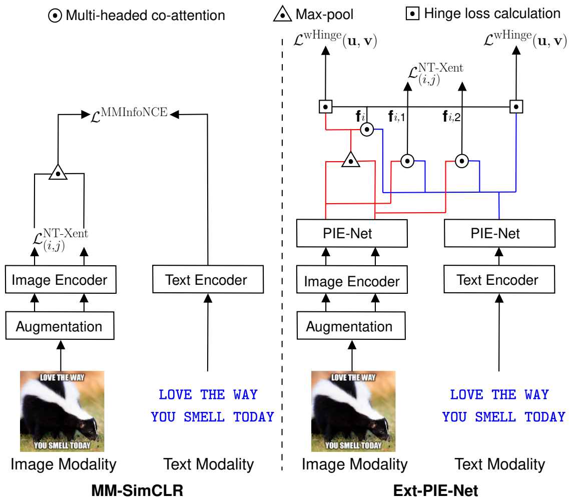


Figure 1: Solution architectures of multi-modal self-supervision for memes. MM-SimCLR: Multi-modal SimCLR (left); Ext-PIE-Net: Extended Pie-Net (right).

and $\mathbf{z}_j$ are the projections for augmented views i and j, respectively; and τ is temperature.

- *Hinge Loss:* Conventionally, hinge loss has been known to be applied to characterize optimization in uni-modal vector space (Rosasco et al., 2003). The formulation of the multi-modal hinge loss has been employed in (Faghri et al., 2018). For a two-modality system with u and v as modality-specific representations in common space, a multi-modal weighted hinge loss ($\mathcal{L}^{\text{wHinge}}$) is formulated using a cosine similarity function $s(\cdot)$. It assumes a margin of α and clamps the value with a ReLU function. Moreover, the individual terms are weighted by λ_{u2v} and λ_{v2u} before aggregation. This is expressed as follows:

$$\mathcal{L}^{\text{wHinge}}(\mathbf{u},\mathbf{v}) = \lambda_{u2v}\sum_{\hat{u}}\text{ReLU}\Big(\alpha - s(\mathbf{u},\mathbf{v}) + s(\hat{\mathbf{u}},\mathbf{v})\Big) + \lambda_{v2u}\sum_{\hat{v}}\text{ReLU}\Big(\alpha - s(\mathbf{u},\mathbf{v}) + s(\mathbf{u},\hat{\mathbf{v}})\Big) \quad (2)$$

MM-SimCLR: In our first approach, MM-SimCLR, we integrate discriminative modeling capacity, which leverages contrastive learning in the latent space for images and a dedicated formulation for a multi-modal setup. This is motivated by (Zhang et al., 2020), which performs contrastive learning between the medical images and their associated texts. Their objective function $\mathcal{L}$ constitutes two terms ($\ell_i^{u\to v}$ and $\ell_i^{v\to u}$) to maximize association between image and text representations ($\mathbf{u}_i$ and $\mathbf{v}_i$). Both $\mathbf{u}_i$ and $\mathbf{v}_i$ are normalized to unit-vectors before being incorporated into the loss terms. τ is a scaling factor that controls the sensitivity of association, and λ controls the weight of the individual term in the final equation. This is given by:

$$\ell_i^{v\to u} = -\log\frac{\exp(\langle\mathbf{v}_i,\mathbf{u}_i\rangle/\tau)}{\sum_{k=1}^{N}\exp(\langle\mathbf{v}_i,\mathbf{u}_k\rangle/\tau)} \quad (3)$$

$$\ell_i^{u\to v} = -\log\frac{\exp(\langle\mathbf{u}_i,\mathbf{v}_i\rangle/\tau)}{\sum_{k=1}^{N}\exp(\langle\mathbf{u}_i,\mathbf{v}_k\rangle/\tau)} \quad (4)$$

We will refer to this objective function as Multi-modal InfoNCE loss in our work, given by:

$$\mathcal{L}^{\text{MMInfoNCE}} = -\frac{1}{N}\sum_{i=1}^{N}(\lambda\ell_i^{u\to v} + (1-\lambda)\ell_i^{v\to u}) \quad (5)$$

Finally, we formulate a new objective function for MM-SimCLR as the summation of SimCLR (Eq. 1) and Multi-modal InfoNCE (Eq. 5) losses. The overall process flow is shown in Fig. 1 (left).

$$\mathcal{L} = \mathcal{L}^{\text{MMInfoNCE}} + \sum_{i=1}^{N}\mathcal{L}_i^{\text{NT-Xent}} \quad (6)$$

Ext-PIE-Net: Inspired by PIE-Net (Song and Soleymani, 2019), which is a diversity-inducing visual-semantic embedding learning framework, we propose Ext-PIE-Net, which optimizes an *augmented* multi-modal objective function (in Eq. 7). PIE-Net leverages a representation learning scheme to cater to the lexical diversity within languages via symmetric cross-modal loss formulations. On the other hand, we augment such a formulation by factoring in an additional loss term due to image-specific contrastive loss. It essentially has three major components – SimCLR $\mathcal{L}^{\text{NT-Xent}}$ (Eq. 1) and a pair of weighted hinge losses $\mathcal{L}^{\text{wHinge}}$ (Eq. 2). $\mathcal{L}^{\text{NT-Xent}}$ optimizes the agreement between the augmented multi-modal representations $\mathbf{f}_{i,1}$ and $\mathbf{f}_{i,2}$. We compute these multi-modal representations using multi-headed co-attention between the textual and visual representations. The intuition is to leverage the contrasting representations of the visual and textual modalities.

We then fuse image views via max-pooling and subsequently with the textual representation using multi-headed co-attention. The obtained multi-modal representation helps in computing modality-reinforcing weighted hinge losses, $\mathcal{L}^{\text{wHinge}}(\mathbf{i}_i,\mathbf{f}_i)$ and $\mathcal{L}^{\text{wHinge}}(\mathbf{t}_i,\mathbf{f}_i)$, *w.r.t.* the image ($\mathbf{i}_i$) and text ($\mathbf{t}_i$) representations, respectively. The losses are weighted by λ_{f2f} ($=0.6$), λ_{f2i} ($=0.2$) and λ_{f2t}

($= 0.2$) to compute the final loss $\mathcal{L}$. Fig. 1 (right) shows the Ext-PIE-Net framework.

$$\mathcal{L} = \sum_{i}^{N} \Big[\lambda_{f2f} \cdot \mathcal{L}^{\text{NT-Xent}}(\mathbf{f}_{i,1}, \mathbf{f}_{i,2}) + \lambda_{f2i} \cdot \mathcal{L}^{\text{wHinge}}(\mathbf{i}_i, \mathbf{f}_i) + \lambda_{f2t} \cdot \mathcal{L}^{\text{wHinge}}(\mathbf{t}_i, \mathbf{f}_i) \Big] \quad (7)$$

5 Experiments and Results

This section presents the evaluation strategy, description of systems examined, results of experiments on self-supervision, and downstream evaluation. We first experiment with various self-supervision strategies and then evaluate the representations learned from best-performing systems by evaluating different downstream tasks for label-efficient supervised learning.[4,5]

To evaluate the representations learned through pre-training, we employ the linear evaluation strategy (Oord et al., 2018), which trains a linear classifier with frozen base network parameters. This is a popular strategy for assessing the quality of the representations learned with a minimal predictive modeling setup that facilitates a fair assessment of the resulting inductive bias. The performance on the test set implies the quality of the representations learned. Since the primary focus of our work is self-supervision for multi-modal applications, we emphasize our investigation and compare mainly with the multi-modal state-of-the-art setups. Also, as we motivate in the Introduction section, standardized large-scale multi-modal datasets like MS-COCO, CC, etc., used towards pre-training visual-linguistic models like ViLBERT (Lu et al., 2019a) and Visual BERT (Li et al., 2019b) incur significant development cost, we mostly restrict our SSL+FT comparison either to the setups that can conveniently leverage raw datasets like MMHS150K (Gomez et al., 2020), which are conveniently accessible via web (*one of the primary motivations for this work*), or pre-trained and fine-tuned versions of ViLBERT and Visual BERT. For comparison, we comply with the respective works and compute accuracy values for the Hateful Memes task and Macro-F1 scores for the Memotion and HarMeme tasks and report all the results by taking the average across *five* independent runs.

5.1 Self-supervised Learning and Linear Evaluation

Systems: We experiment with a few existing related approaches and different uni-modal and multi-modal variants and compare self-supervised and supervised learning frameworks for a comprehensive assessment. We do not consider explicit pre-training of models like Visual BERT and ViLBERT within the scope of the current study because their pre-training strategies are designed for explicitly modeling visual-linguistic grounding. This can constrain the self-supervised learning based upon *domain-aware* pre-training, using a dataset from the wild (WWW), which is a crucial aspect of our study. However, we do compare the SSL+FT systems with completely fine-tuned and pre-trained checkpoints of Visual BERT (MS-COCO) and ViLBERT (CC) systems. The details of these systems are enlisted as follows: • SimCLR (Chen et al., 2020a): The framework focuses on incentivizing the agreement between similar image views. • VSE++ (Faghri et al., 2018): It focuses on mining hard negatives to heavily penalize for dissimilarity with the anchor images through a hinge-like loss. • Modified SimCLR: We try to extend the loss proposed in SimCLR to text modality via augmentation. We do so using WordNet (Fellbaum, 1998) synonyms replacement and through back-translation (Sennrich et al., 2016) approaches.

We also compare state-of-the-art multi-modal systems for better task-specific assessment. These are: • Late fusion: Averages prediction scores of ResNet-152 and BERT. • Concat BERT: Concatenates representations from ResNet-152 and BERT, using a perceptron as a classifier. • MMBT: Multimodal Bitransformer (Kiela et al., 2019), capturing the intra/inter-modal dynamics. • ViLBERT CC: Vision and Language BERT (Lu et al., 2019b), trained on an intermediate multi-modal objective (conceptual captions) (Sharma et al., 2018), comprises of task-independent joint representation multi-modal framework. • Visual BERT COCO: Pre-trained (Li et al., 2019b) using MS-COCO dataset (Lin et al., 2014).

Results: We first examine representations learnt by SimCLR (Chen et al., 2020a) and evaluate them by fine-tuning on Hateful Memes task. As shown in Table 1, this results in a meagre accuracy of 0.50 – a difference of only 0.67% against the image-only *fully supervised* baseline (accuracy 0.5067). Moving forward, our initial attempt toward mod-

[4] We use abbreviations SL, SSL and FT for supervised, self-supervised learning, and fine-tuning, respectively.

[5] Additional details of experiments, along-with hyperparameters explored are included as part of App. A.

Type	Model	Acc.
SL	Image-Grid (image-only)	0.507
	ViLBERT	0.631
	ViLBERT CC	0.661
	Visual BERT	0.650
	Visual BERT COCO	0.659
	alfred lab	**0.732**
SSL	SimCLR (image-only)	0.500
	Mod. SimCLR-WN	0.481
	Mod. SimCLR-BT	0.450
	VSE	0.501
	VSE++†	0.536
	MM-SimCLR	0.551
	Ext-PIE-Net*	**0.600**
	$\Delta_{(\star\text{-}\dagger)\times 100}(\%)$	↑ 6.42%

Table 1: Comparison between the proposed SSL method and baselines on the Hateful Memes dataset. † represents SSL baseline and ⋆ is for the proposed approach.

eling multi-modality involves evaluating a VSE++ (Faghri et al., 2018) setup, which leverages *hard-negative* sampling to distinguish similar and dissimilar representations. Due to the factoring of hard-negatives in VSE++, the mutual information between the representations of semantically close image-text pairs is regulated and yields an improved accuracy of 0.53. Our attempt to extend SimCLR for textual modality results in low accuracy values of 0.48 and 0.45, respectively. The low performances are possible due to the changes in the textual semantics that augmentation techniques could induce, effectively reducing potential harmfulness modeling affinity.

In comparison, MM-SimCLR enhances the performance, yielding an accuracy of 0.5508. Ext-PIE-Net is observed to further enhance it to 0.5998 – a gain of $+9.98\%$ over the image-only SimCLR framework, whereas $+9.84\%$ and $+6.42\%$ over the multi-modal VSE and VSE++ systems respectively (Table 1). One of the characteristic changes that the proposed solutions incorporate in contrast to the other frameworks is the combined consideration of multiple image views and a single textual representation toward modeling a specialized multi-modal contrastive learning setup. This is likely responsible for the cross-modal efficacy observed in the performance. Although the performances of the proposed models fall behind that of their fully-supervised counterparts, they perform reasonably better than the strong self-supervised methods.

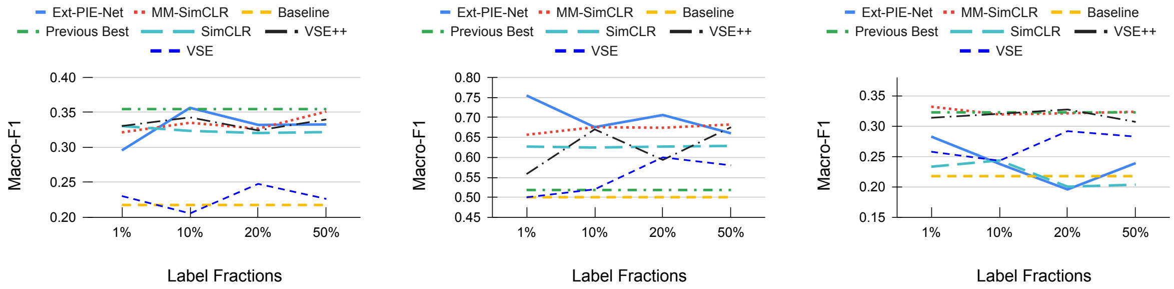
Type	Systems	Task-wise Macro-F1 scores		
		SENT	EMOT	EMOT-Q
SL	Baseline	0.218	0.500	0.301
	Visual BERT	0.320	-	-
	ViLBERT	0.335	-	-
	Previous Best‡	**0.355**	**0.518**	**0.323**
SSL	SimCLR (image-only)	0.330	0.629	0.244
	VSE	0.248	0.580	0.292
	VSE++†	0.343	0.675	0.327
	Ext-PIE-Net*	**0.357**	**0.755**	0.283
	MM-SimCLR*	0.351	0.682	**0.332**
	$\Delta_{(\star\text{-}\dagger)\times 100}(\%)$	↑ 1.37%	↑ 7.93%	↑ 0.46%

Table 2: Comparison of SSL+FT with previous best and baseline for Memotion tasks. † represents SSL baseline and ⋆ is for the proposed approach and ‡ (Previous best): best scores for the corresponding tasks.

5.2 Label-Efficient Training on Downstream Tasks

We evaluate the representations learned via linear classification using a *subset* of labeled samples following self-supervised pre-training to assess label efficiency during adaption. A classification head consisting of a linear layer brings the modalities into the same dimension (we use 512). Furthermore, a shallow, fully connected network classifies the obtained multi-modal representation into target labels. We opt for the Memotion and HarMeme tasks for this paradigm. Based on the results obtained from the evaluation of self-supervision strategies, we evaluate the pre-training performance on these downstream tasks.

Results on Memotion Analysis: Due to the complex nature of the dataset and the tasks involved, the baselines and the leader-board for Memotion task (Sharma et al., 2020) reflect the resulting non-triviality – with SOTA results as 0.354, 0.518, and 0.32 Macro-F1 for SENT, EMOT, and EMOT-Q tasks, respectively. Moreover, the complexity of the tasks can be further ascertained via the baseline's Macro-F1 scores of 0.217, 0.500, and 0.300 for the three tasks – the baseline systems are trivial early fusion (for SENT task), and late fusion-based (for EMOT and EMOT-Q tasks) approaches on top of CNN and RNN based image and text encoding mechanisms. The previous best systems involve a word2vec (Mikolov et al., 2013b,a) based feed-forward neural network for SENT (Keswani et al., 2020), a multi-modal multi-tasking based setup for EMOT (Vlad et al., 2020), and a feature-based ensembling approach for the EMOT-Q task (Guo et al., 2020). These results solicit improvement in multi-modal systems.



(a) SENT task

(b) EMOT task

(c) EMOT-Q task

Figure 2: Comparison between the proposed method and baselines on Memotion tasks. X-axis signifies the incremental supervision during fine-tuning.

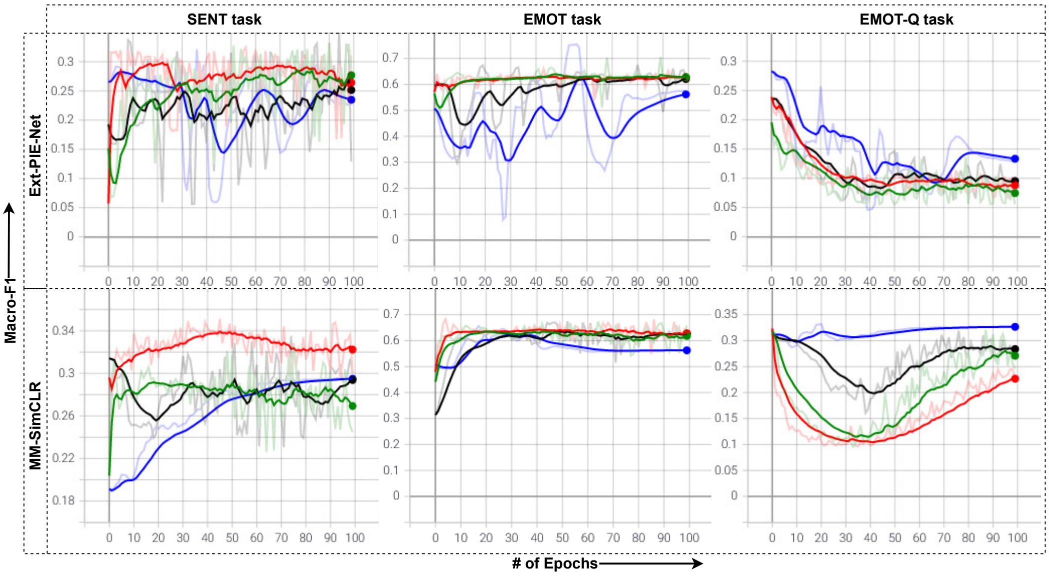


Figure 3: Training performance comparison for different label fractions [**1 %** – **10 %** – **20 %** – **50 %**] for Ext-PIE-Net (top row) and MM-SimCLR (bottom row) on Memotion tasks. Dominant curves are *smoothed* depiction of the actual curves in the background.

Type	Systems	Macro-F1
SL	Late Fusion	0.7850
	Concat BERT	0.7638
	MMBT	0.8023
	ViLBERT CC	0.8603
	Visual BERT COCO	0.8607
	MOMENTA	**0.8826**
SSL	SimCLR (image-only)	0.6328
	VSE	0.6569
	VSE++†	0.7912
	Ext-PIE-Net	0.5717
	MM-SimCLR*	**0.8140**
$\Delta_{(\star-\dagger)\times 100}(\%)$		↑ 2.28%

Table 3: Comparison of SSL+FT with previous best and baseline for HarMeme task.

We showcase the results on the same tasks by our proposed approaches in Table 2. Ext-PIE-Net outperforms Late-fusion baseline, Visual BERT, ViLBERT, the previous best (amongst SL), and uni-modal, multi-modal, and MM-SimCLR (amongst SSL) systems in the SENT and EMOT tasks. It reports an improvement of 1.37% in SENT but a significant 7.93% increment over that from VSE++ (best SSL) in EMOT at 0.3565 and 0.7547 Macro-F1 scores, respectively. In comparison, the performance in EMOT-Q is non-convincing at 0.2827 Macro-F1 score – this could be due to the multi-class and multi-label nature of the task. Whereas, since SENT and EMOT tasks are formulated by aggregating data samples for the higher level of categorical consideration, they are relatively complex due to the resulting data imbalance. Although MM-SimCLR performs better on EMOT-Q task and overall, at-par or better than the baseline, it still lags by a small margin for SENT task and significantly for Task B compared to Ext-PIE-Net. Also, Ext-PIE-Net setup has a relatively more significant number of trainable parameters than MM-SimCLR, facilitating better modeling capacity for SENT and EMOT tasks. Conversely, MM-SimCLR performs better on EMOT-Q task due to better compatibility of the modeling capacity and task. The overall results signify the efficacy of proposed SSL strategies on complex downstream multi-modal tasks. These results highlight the task-specific peculiarities that modeling needs to factor in for optimal performance.

Results on Harmful Memes: The transferability of the representations learned through pre-training is examined by fine-tuning on another meme dataset, i.e., Harm-P. We report the results in Table 3. The fully supervised models, such as VilBERT CC (Pramanick et al., 2021), Visual BERT COCO (Pramanick et al., 2021), and MOMENTA (Pramanick et al., 2021), obtain Macro-F1 scores of 0.8603, 0.8607, and 0.8826, respectively. In comparison, MM-SimCLR in a label-efficient setup records a convincing performance of 0.8140 Macro-F1. One of our proposed approaches Ext-PIE-Net performs poorly with 0.5717 F1 against an impressive F1 score of 0.8140 by MM-SimCLR. Like its performance on Memotion task, MM-SimCLR is observed to perform better on a relatively more straightforward HarMeme task. Even though MM-SimCLR lags behind by 4.6% from strong SL baselines ViLBERT CC and Visual BERT COCO, and MOMENTA by 7.02%, it distinctly outperforms other competitive multi-modal baselines (supervised) like Late Fusion, Concat BERT and MMBT by 2.9%, 5.02% and 1.87%, respectively. MM-SimCLR also leads SimCLR (0.6328) by 18.12%, and SSL multi-modal baselines VSE (0.6569), VSE++ (0.7912) and Ext-PIE-Net (0.5717) by 15.71%, 2.28% and 24.2%, respectively on the HarMeme task.

It is also worth highlighting that the performances of strong multi-modal models like Visual BERT and ViLBERT can be inconsistent, depending upon the task being addressed. This is primarily due to the fact that the corresponding pre-training involved leverages strong visual-linguistic grounding, which based on downstream task complexity, can give varying results as observed for Memotion (c.f. Table. 2) and HarMeme (c.f. Table 3). This suggests the scope of enhancement towards the pre-training objectives and frameworks within the existing multi-modal systems.

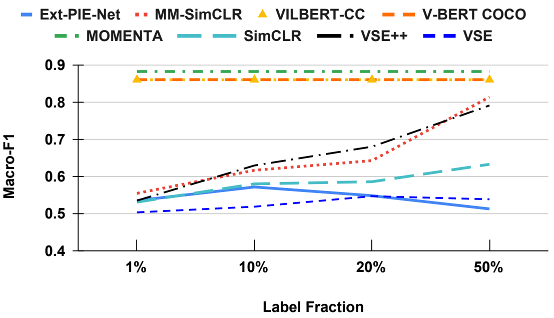


Figure 4: Comparison b/w the proposed method and baselines on HarMeme tasks. X-axis signifies the incremental supervision during fine-tuning.

6 Impact of Label-Efficient Supervision During Fine-tuning

Towards assessing the label-efficient setup, we compare the performances over incremental supervision. We also analyze their temporal training behavior.

As can be observed from Fig. 2a, Ext-PIE-Net converges efficiently to 0.3565 F1 score with just 10% (600) training samples, as compared to MM-SimCLR which converges to 0.3511 F1 score after learning from 50% (3000) of the labeled samples. This highlights the capacity of a sophisticated SSL regime to learn better representations for a complex setup for the SENT task compared to a slightly simpler model MM-SimCLR. A similar pattern can be observed for EMOT task in Fig. 2b. Ext-PIE-Net is observed to achieve an overall better F1 score of 0.7547, which is better than MM-SimCLR and outperforms all other results.

Although the optimal performance of SimCLR is reasonably at-par or even better for SENT and EMOT tasks compared to the baseline and the previous best results, there is barely any active convergence visible within the plots depicted in Fig. 2 for it. This is obvious considering the incomplete information that an image-only based uni-modal system would learn for the downstream task. VSE is observed to yield 3.02% and 7.98% improvement over the SL baseline. Still, it fails to register an impressive performance compared to the increment of 12.52% and 17.52% for the two tasks, respectively, by VSE++.

These observations can also be correlated with the training performance (c.f. Fig. 3), wherein the performance curves are depicted for a total of 100 epochs across four different label-efficiency scenarios. For primary assessment, we showcase *smoothed* curves overlaid on *unsmoothed* ones towards observing global and local trends. [6]

Fig. 3 presents a clear depiction of progressive learning for all the supervision configurations evaluated in case of Ext-PIE-Net for the SENT and EMOT tasks (c.f. Fig. 3) is given. On the other hand, the training curves for MM-SimCLR show saturated learning for tasks SENT and EMOT respectively (c.f. Fig. 3).

Delineating on the performance trend observed in the EMOT-Q task earlier, neither Ext-PIE-Net nor SimCLR shows definite convergence, as we consider the incremental supervision depicted in Fig. 2c. Whereas, MM-SimCLR is observed to show stable, yet non-incremental growth in performance reporting the best overall F1 score of 0.3318 (c.f. Table 2). This task entails a relatively balanced training set (Sharma et al., 2020), and MM-SimCLR is observed to offer just the required simplicity for solving such a task. The training characteristics observed for this task, are found to be contrasting for Ext-PIE-Net and MM-SimCLR (c.f. Fig. 3, last figures from *first and second rows, respectively*). MM-SimCLR indicates overall progressive learning. On the other hand, Ext-PIE-Net depicts a consistently regressive trend. This corroborates the optimal convergence demonstrated by a simple multi-modal contrastive loss-based self-supervision for a more straightforward task formulation.

For HarMeme task, the incremental supervision (c.f. Fig. 4) exhibits incremental performance with the increase in the amount of supervision during fine-tuning. Notably, the final F1 score of 0.814 obtained by the MM-SimCLR model is on just 50 % (1510) of the actual training set. This demonstrates the efficacy and generalizability of the pre-training via strategies adopted in this work. Also, the progressive convergence observed at 50% supervision, as shown in Fig. 4 for MM-SimCLR, demonstrates the generalizability of the proposed approach. This also suggests the importance of having smaller architectures with sophisticated fusion strategies to solve the task at hand effectively.

7 Discussion

The observations made from the results obtained for the downstream evaluation suggest interesting trends. Since Memotion dataset involves multi-class, multi-label and multi-level hierarchical granularity due to the natural distribution of such realistic dataset, either ensembling-based approaches are observed to yield better results or, there are strong variations observed in the performance trends across the three Memotion tasks (Sharma et al., 2020). The results reported as part of Table 1, 2 and 3 exhibit insights correlating the task complexity with that of the modelling solutions required. This is further corroborated by the results on HarMeme task. To this end, we have highlighted the performances and drawn comparisons for two models that we empirically examined as part of this investigation.

8 Conclusion

This work empirically examined various self-supervision strategies to learn effective representations that help solve multiple multi-modal downstream tasks in a label-efficient setting. We propose two strategies for this – (i) MM-SimCLR: a multi-modal contrastive loss formulation that factors in the loss terms for image modality and the multi-modality in a joint manner, and (ii) Ext-PIE-Net: a joint formulation of weighted modality-specific hinge loss terms, combined with the contrastive loss that is computed between a pair of representations, obtained using symmetric multi-modal fusion. Extensive analysis over 2 datasets and 5 tasks demonstrate how domain-aware self-supervised pre-training, using a multi-modal dataset, that can be directly obtained from the wild (WWW) in raw form, can be leveraged to perform label-efficient multi-modal adaptation, leading to competitive, even superior performance gains for some scenarios.

The performances observed for the proposed methods indicate *task-dependent* efficacies. MM-SimCLR being a lighter model is observed to perform better on EMOT-Q and HarMeme tasks, having a lower level of granularity to be modeled. Whereas Ext-PIE-Net performs better on SENT and EMOT tasks, which require modeling a higher abstraction level for the target categories. Despite exhibiting interesting performance within label-efficient evaluation settings, the objectives addressed in this work can further benefit from extensive analysis and evaluation towards obtaining a broader understanding of the generalizability of the proposed methodology.

[6] For further reference, *unsmoothed* training curves are also included and discussed separately in App. C.

Acknowledgments

The work was supported by Wipro research grant.

References

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes.

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. 2020a. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607.

Y. C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu. 2020b. Uniter: Universal image-text representation learning. In *ECCV*, pages 104–120.

Victor G Turrisi da Costa, Enrico Fini, Moin Nabi, Nicu Sebe, and Elisa Ricci. 2021. Solo-learn: A library of self-supervised methods for visual representation learning. *arXiv preprint arXiv:2108.01775*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. Vse++: Improving visual-semantic embeddings with hard negatives. In *BMVC*.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, pages 249–256.

Raul Gomez, Jaume Gibert, Lluís Gómez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *WACV*, pages 1459–1467.

J. B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, et al. 2020. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*.

Yingmei Guo, Jinfa Huang, Yanlong Dong, and Mingxing Xu. 2020. Guoym at SemEval-2020 task 8: Ensemble-based classification of visuo-lingual metaphor in memes. In *SemEval-2020*, pages 1120–1125, Barcelona.

T. Gupta, K. J. Shih, S. Singh, and D. Hoiem. 2017. Aligned image-word representations improve inductive transfer across vision-language tasks. In *ICCV*, pages 4223–4232.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. *CVPR*, pages 770–778.

Andrej Karpathy and Li Fei-Fei. 2017. Deep visual-semantic alignments for generating image descriptions. *IEEE TPAMI*, 39(4):664–676.

Vishal Keswani, Sakshi Singh, Suryansh Agarwal, and Ashutosh Modi. 2020. IITK at SemEval-2020 task 8: Unimodal and bimodal sentiment analysis of Internet memes. In *SemEval-2020*, pages 1135–1140, Barcelona.

Erica Kido Shimomoto, Lincon Souza, Bernardo Gatto, and Kazuhiro Fukui. 2019. News2meme: An automatic content generator from news based on word subspaces from text and image. In *MVA*, pages 1–6.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text. In *Proceedings of the NeurIPS Workshop on Visually Grounded Interaction and Language*, ViGIL '19, Vancouver, Canada.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *arXiv preprint arXiv:2005.04790*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Olamide Temitayo Kolawole. 2015. Classification of internet memes.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. *ArXiv*, abs/1909.11942.

Kunpeng Li, Yulun Zhang, K. Li, Yuanyuan Li, and Yun Raymond Fu. 2019a. Visual semantic reasoning for image-text matching. *ICCV*, pages 4653–4661.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019b. Visualbert: A simple and performant baseline for vision and language. In *Arxiv*.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019a. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019b. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the Conference on Neural Information Processing Systems*, NeurIPS '19, pages 13–23, Vancouver, Canada.

L. Manovich. 2001. *The Language of New Media*. Leonardo (Series) (Cambridge, Mass.). MIT Press.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *ICLR*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *NeurIPS*, volume 26.

Aditya Mogadala, Marimuthu Kalimuthu, and Dietrich Klakow. 2021. Trends in integration of vision and language research: A survey of tasks, datasets, and methods. *JAIR*, 71:1183–1317.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

V Peirson, L Abel, and E Meltem Tolunay. 2018. Dank learning: Generating memes using deep neural networks. *arXiv preprint arXiv:1806.04510*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL*.

S. Pramanick, S. Sharma, D. Dimitrov, M. S. Akhtar, P. Nakov, and T. Chakraborty. 2021. MOMENTA: A multimodal framework for detecting harmful memes and their targets. In *EMNLP-Findings*, pages 4439–4455, Punta Cana, Dominican Republic.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *ICML*, pages 8821–8831.

L. Rosasco, E. De, Vito A. Caponnetto, M. Piana, and A. Verri. 2003. Are loss functions all the same. *Neural Computation*, 15:2004.

R. Sennrich, B. Haddow, and A. Birch. 2016. Improving neural machine translation models with monolingual data. In *ACL*, pages 86–96, Berlin, Germany.

C. Sharma, D. Bhageria, W. Scott, S. PYKL, A. Das, T. Chakraborty, V. Pulabaigari, and B. Gambäck. 2020. SemEval-2020 task 8: Memotion analysis-the visuo-lingual metaphor! In *SemEval*, pages 759–773.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*.

Yale Song and Mohammad Soleymani. 2019. Polysemous visual-semantic embedding for cross-modal retrieval. In *CVPR*, pages 1979–1988.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. *WIT: Wikipedia-Based Image Text Dataset for Multimodal Multilingual Machine Learning*, page 2443–2449. Association for Computing Machinery, New York, NY, USA.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proc. of the 57th Annual Meeting of the Assoc. for Computational Linguistics*, ACL '19, pages 3645–3650, Florence, Italy.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

G. A. Vlad, G. E. Zaharia, D. C. Cercel, C. Chiru, and S. Trausan-Matu. 2020. UPB at SemEval-2020 task 8: Joint textual and visual modeling in a multi-task learning architecture for memotion analysis. In *SemEval-2020*, pages 1208–1214, Barcelona.

Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao. 2021. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv 2108.10904*.

Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz. 2020. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*.

Type	Name	BS	Epochs	LR	Image Encoder	Text Encoder
SSL	SimCLR	32	150	0.1	ResNet-50	-
	VSE++		100	0.0001	ResNet-18	distilbert-base-uncased
	Mod. SimCLR					
	MM-SimCLR					
	Ext-PIE-Net					
SL	SimCLR	512	100	0.0001	ResNet-50	-
	MM-SimCLR	256		0.0005	ResNet-18	distilbert-base-uncased
	Ext-PIE-Net					

Table 4: Hyperparameter values for the experiments.

A Experimental setup and Hyperparameters:

We train all our experiments using Pytorch on an NVIDIA Tesla P4 with 8 GB dedicated memory. We use VISSL, an open-source library (da Costa et al., 2021) to evaluate SimCLR, a uni-modal image-only setup for memes. For the multi-modal setups, we initialize the networks with weights of pre-trained models available for image encoders with PyTorch library and the text models with weights available from `transformers` package from hugging face library[7].

The image encoder is a ResNet-18 (He et al., 2016) architecture and the text encoder is a `distilbert-base-uncased` in all our multi-modal experiments. After self-supervised pre-training, we freeze the text and image encoder weights and discard the projection heads attached. As part of the classification head, a new set of layers are added to perform supervised learning using fewer labeled samples. We initialize the layers using Xavier initialization (Glorot and Bengio, 2010) and set the bias to zero. We train all the models using the Adam optimizer (Kingma and Ba, 2015) and a cross-entropy loss as the objective function for supervision for all the tasks evaluated in this work. We perform multi-modal self-supervision experiments keeping a batch size of 32 for 100 epochs at a learning rate of 0.0001. The SimCLR experiment in self-supervision is carried out for 150 epochs with a batch size of 32 and a learning rate of 0.1 using a ResNet-50 backbone. The encoder weights are frozen during the label-efficient training, and the classification heads are used, allowing 256 batch-size in multi-modal experiments and 512 for uni-modal SimCLR experiment. The SimCLR-based label-efficient setup is trained with 0.0001 learning rate, while the other multi-modal experiments are trained with 0.0005 learning rate. We also present these details in Table 4.

B Statistical Analysis of Datasets

The datasets used in this work have been either created synthetically using specific hate topics or downloaded from social media platforms using generic and domain-specific hate keywords (Kiela et al., 2020; Gomez et al., 2020; Pramanick et al., 2021). The top-5 hate and non-hate keywords ranked as per the tf-idf scores of their occurrences within the accompanying texts are shown in Table 5. This table shows that the hateful lexicon for MMHS150K represents extreme urban parlance, depicting realistic social media communication, whereas in the Hateful Memes dataset, hate keywords are canonical and topic-oriented. To counter the potential keyword bias within the datasets, the categorical representation of these keywords was explicitly balanced by introducing confounders or considering contrastive examples for the exact hate keywords.

The accompanying texts from all datasets used have a mean length of 8 (c.f. Fig. 5). The distribution observed for MMHS150K in Fig. 5a is almost uniform, with most of the posts having lengths of less than 30 words, primarily due to the 280-character limit on tweets. Hateful Memes, on the other hand, is created with reasonable variation, having examples with lengths greater than 30 as well. Their confounding effect is also clearly visible within these histogram plots, where hateful content with larger corresponding text could also be present in some samples (Fig. 5b), as against the general trend where the variation in the length is confined. Finally, Harm-P reflects the distribution of the accompanying textual contents over social media. Hence the variation depicted in Fig. 5c.

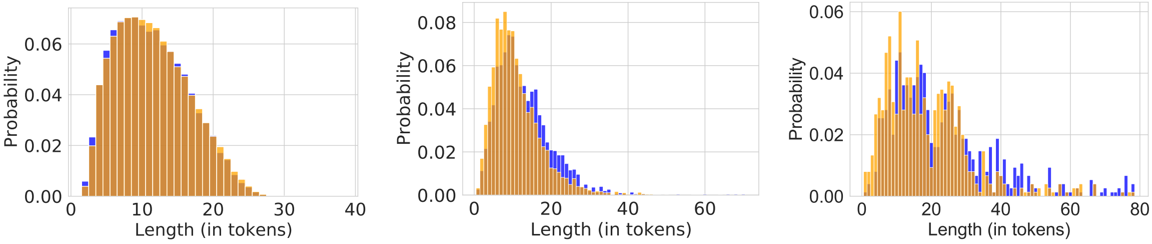
C Training Characteristics

The *unsmoothed* training curves, depicted in Fig. 6 reflects the trends observed with the *smoothed* depiction in Fig. 3. Besides significant fluctuations within the training curves across tasks, especially for SENT and EMOT-Q tasks, subtle temporal trends can be inferred. There is a gradual enhancement in the performances observed within early epochs (<60) for both SENT and EMOT tasks, for both Ext-PIE-Net and MM-SimCLR, with Ext-PIE-Net registering the best macro-f1, along with significant variation. But overall, the performances are reasonably similar. For SENT task, Ext-PIE-Net showcases consistent growth in the macro-f1 score for all the label-configuration

[7] https://huggingface.co

MMHS150K				Hateful Memes				Harm-P			
Hateful		Not-hateful		Hateful		Not-hateful		Harmful		Not-harmful	
Word	*tf-idf score*	*Word*	*tf-idf score*	*Word*	*tf-idf score*	*Word*	*tf-idf score*	*Word*	*tf-idf score*	*Word*	*tf-idf score*
faggot	0.0441	redneck	0.0099	black	0.0433	like	0.0337	photoshopped	0.0589	party	0.02514
cunt	0.0364	love	0.0098	white	0.0378	day	0.018	married	0.0343	debate	0.0151
nigger	0.0346	happy	0.0081	muslim	0.0321	got	0.0174	joe	0.0309	president	0.0139
retarded	0.0306	good	0.0074	jews	0.0239	time	0.0172	trump	0.0249	democratic	0.0111
trash	0.0214	hillbilly	0.0071	kill	0.0223	love	0.0138	nazis	0.0241	green	0.0086

Table 5: The top-5 most frequent words and their tf-idf scores in each class.



(a) MMHS150K (b) Hateful Memes (c) Harm-P

Figure 5: Distributions of the text's length. Blue: Hateful/Harmful; Orange: Not-hateful/harmful.

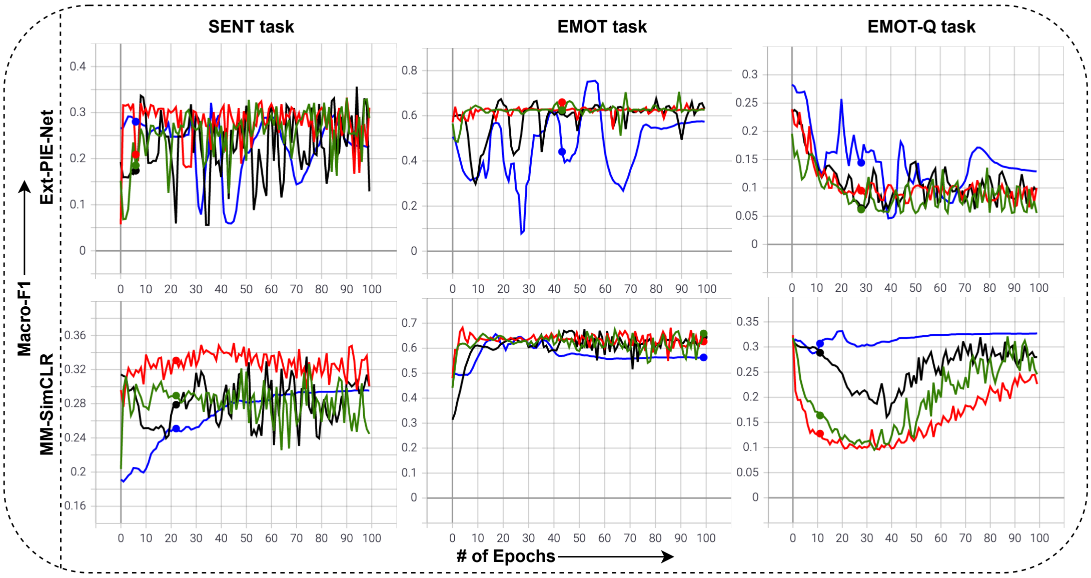


Figure 6: Training performance comparison (*unsmoothed*) for different label fractions [**1 %** – **10 %** – **20 %** – **50 %**] for Ext-PIE-Net (top row) and MM-SimCLR (bottom row) on Memotion tasks.

scenarios. In contrast, MM-SimCLR showcases progress for scenarios involving 1% and 50% labeled samples only. On the other hand, for EMOT-Q task, MM-SimCLR is observed to exhibit better convergence after 30^{th} epoch, as against that by Ext-PIE-Net, across label-configurations, suggesting better training behavior (c.f. Fig. 6).

D Ethics and Broader Impact

User Privacy. The meme content and the associated information do not include any personal information. Issues related to copyright are addressed as part of the dataset source.

Biases. Any biases found in the datasets (Gomez et al., 2020; Kiela et al., 2020; Pramanick et al., 2021) leveraged in this work are presumed to be unintentional, as per the attributions made in the respective sources, and we do not intend to cause harm to any group or individual. We acknowledge that detecting emotions and harmfulness can be subjective, and thus it is inevitable that there would be biases in gold-labeled data or the label distribution. The primary aim of this work is to contribute with a novel multi-modal framework that helps perform downstream-related tasks, utilizing the representations learned via self-supervised learning.

Misuse Potential. We find that the datasets used in this work can be potentially used for ill-intended purposes, like biased targeting of individuals/communities/organizations, etc., that may or may not be related to demographics and other information within the text. Any research activity would require intervention with human moderation to ensure this does not occur.

Intended Use. We use the existing dataset in our work in line with the intended usage prescribed by its creators and solely for research purposes. This applies in its entirety to its further use as well. We commit to releasing our dataset, aiming to encourage research in studying harmful targeting in memes on the web. We distribute the dataset for research purposes only, without a license for commercial use. We believe that it represents a valuable resource when used appropriately.

Environmental Impact. Finally, due to the requirement of GPUs/TPUs, large-scale Transformers require many computations, contributing to global warming (Strubell et al., 2019). However, in our case, we do not train such models from scratch; instead, we fine-tune them on relatively small datasets. Moreover, running on a CPU for inference, once the model has been fine-tuned, is perfectly feasible, and CPUs contribute much less to global warming.