Multimodal Semi-supervised Learning for Sentiment Analysis of Image Macros

Student
Pham Thai Hoang Tung
Nguyen Tan Viet
Ngo Tien Anh

Supervisor:
Assoc. Prof. Phan Duy Hung

Content

INTRODUCTION

1. Problems

2. Related works

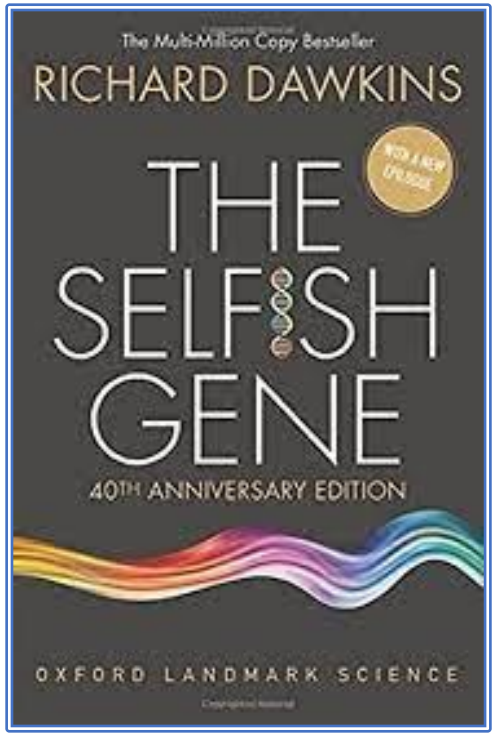
3. Motivation

4. Contribution

1. Problems

Basic Concept

the concept of "meme" first appeared in 1976 in the book The Selfish Gene (Figure 1) by Richard Dawkins (a British author



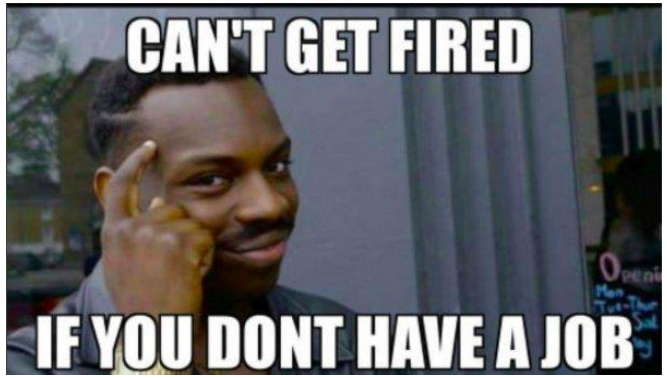
Basic Concept : Meme

1. Problems

Basic Concept

An "Internet meme," often abbreviated as "meme," is an idea, a famous saying, a trend, or a behavior that is spread on the internet [1]

Meme: Gift



Meme: Image + text

Meme: only Image

(1) L. K. Börzsei, “Makes a meme instead,” *Sel. Works Linda Börzsei*, pp. 1–28, 2013.

1. Problems

Basic Concept

"Image macro" is the most common form of internet meme; it consists of an image and a short piece of text overlaid on top of the background. [2]

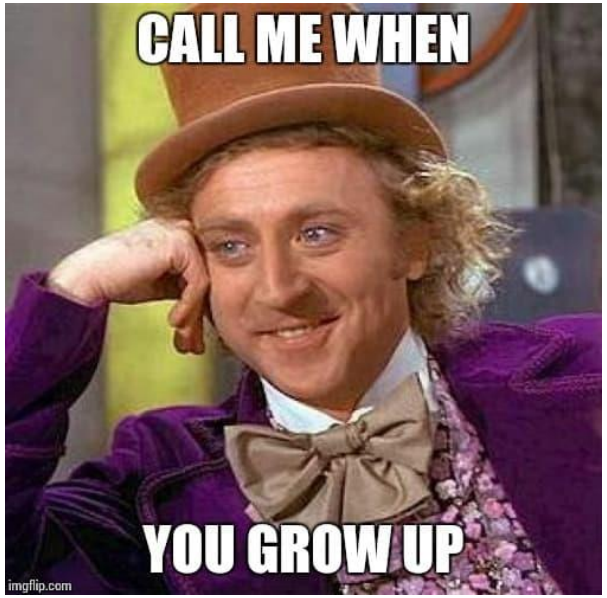
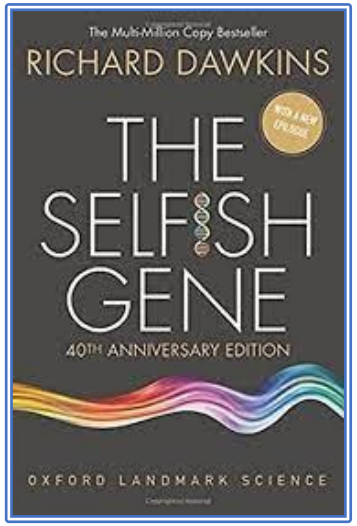


Image macro

[2] J. Lugea, “The pragma-stylistics of ‘image macro’internet memes,” *Contemp. Media Stylist.*, pp. 81–106, 2019

1. Problems

Basic Concept



Meme

Meme Internet
(top-down: meme GIF, meme Image, meme Image + Text)

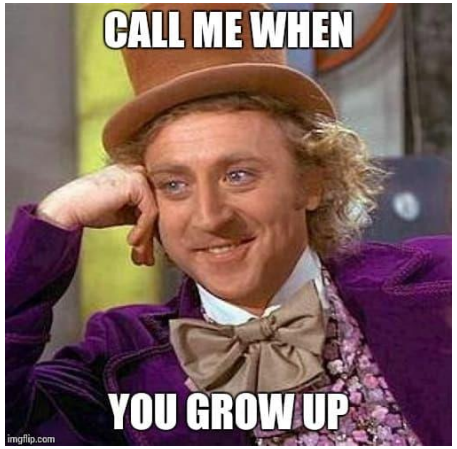


Image macro

1. Problems

Issues surrounding hateful content on social media

+ We have to understand and combine the visual and textual content of the meme

Data manually labeled by humans:
+ not comprehensive
+ subjective
+ conflicts and arguments [1]

[1] R. Gomez, J. Gibert, L. Gomez, and D. Karatzas, “Exploring hate speech detection in multimodal publications,” in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2020, pp. 1470–1478.

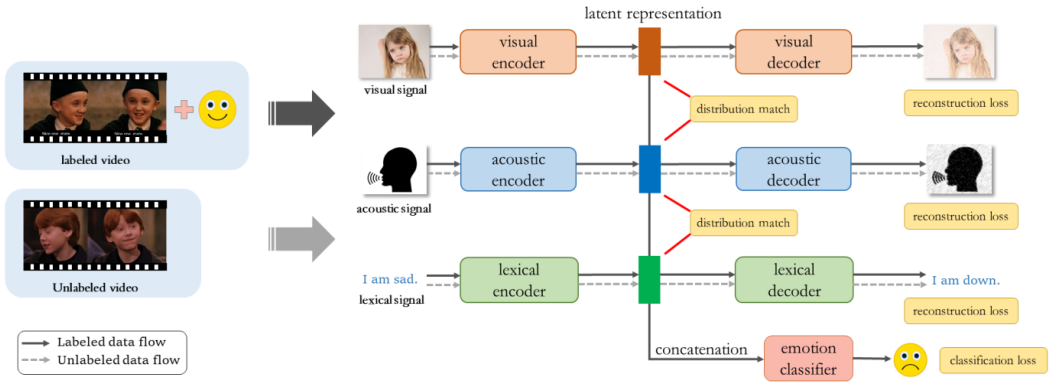
2. Related works

CLIP

VisualBERT

Hateful Meme Challenge

There are some competitions or datasets for meme sentiment analysis. Most of the top solutions are supervised learning, based on CLIP or VisualBERT.

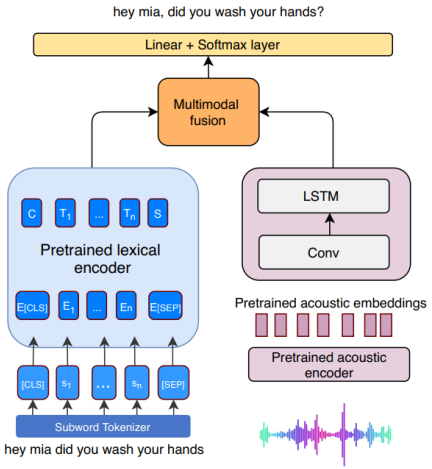


Employs auto-encoder to train end-to-end SSL on emotion recognition. [1]

Some studies about multimodal semi-supervised learning but on different tasks.

Current SOTA on multimodal semi-supervised classification on text and image: Comprehensive Semi-Supervised Multi-Modal Learning. [3]

CMML



Pre-train lexical and acoustic encoders on large-scale unlabeled data, then supervised finetune to predict punctuation on speech. [2]

[1] J. Liang, R. Li, and Q. Jin, "Semi-supervised multi-modal emotion recognition with cross-modal distribution matching,"
[2] M. Sunkara, S. Ronanki, D. Bekal, S. Bodapati, and K. Kirchhoff, "Multimodal semi-supervised learning framework for punctuation prediction in conversational speech,"
[3] Y. Yang, K.-T. Wang, D.-C. Zhan, H. Xiong, and Y. Jiang, "Comprehensive Semi-Supervised Multi-Modal Learning.,"

3. Motivation

- Firstly, the data collected on the meme to serve the research of combining images and text is still small; on the other hand, the cost of hiring labor to re-label is relatively high.
- Second, there are now studies on this issue, mainly results from competitions. However, most research groups go toward supervised learning.

We will use the **MAMI (Multimedia Automatic Misogyny Identification)** dataset [4] to identify and identify memes with inappropriate and sexist content against women

[4] E. Fersini *et al.*, "SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification," in *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, Seattle, United States, Jul. 2022, pp. 533–549. doi: 10.18653/v1/2022.semeval-1.74.

BACKGROUND

1. Transformer architecture
2. Language models
3. Vision Transformer
4. Multimodal
5. Loss Functions
6. Semi-supervised Learning (SSL)
7. Auto-Encoder
8. Activation function

1. Transformer architecture

+ Transformer block was originally introduced in “Attention is all you need” paper.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)\text{V}$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(head_1, \dots, head_h)W^O$$

$$\text{where head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

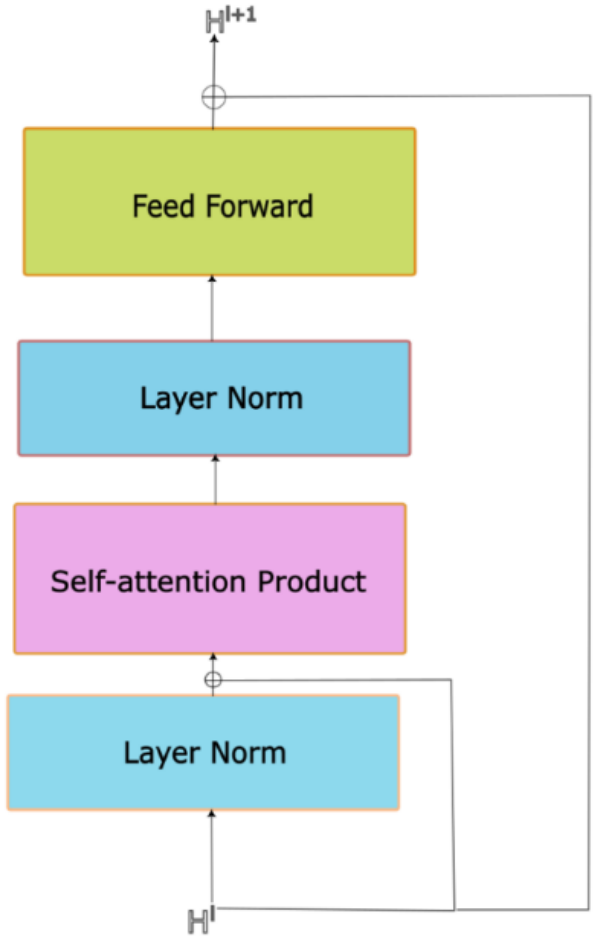


Illustration of transformer block

2. Language Models

First introduced in paper BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

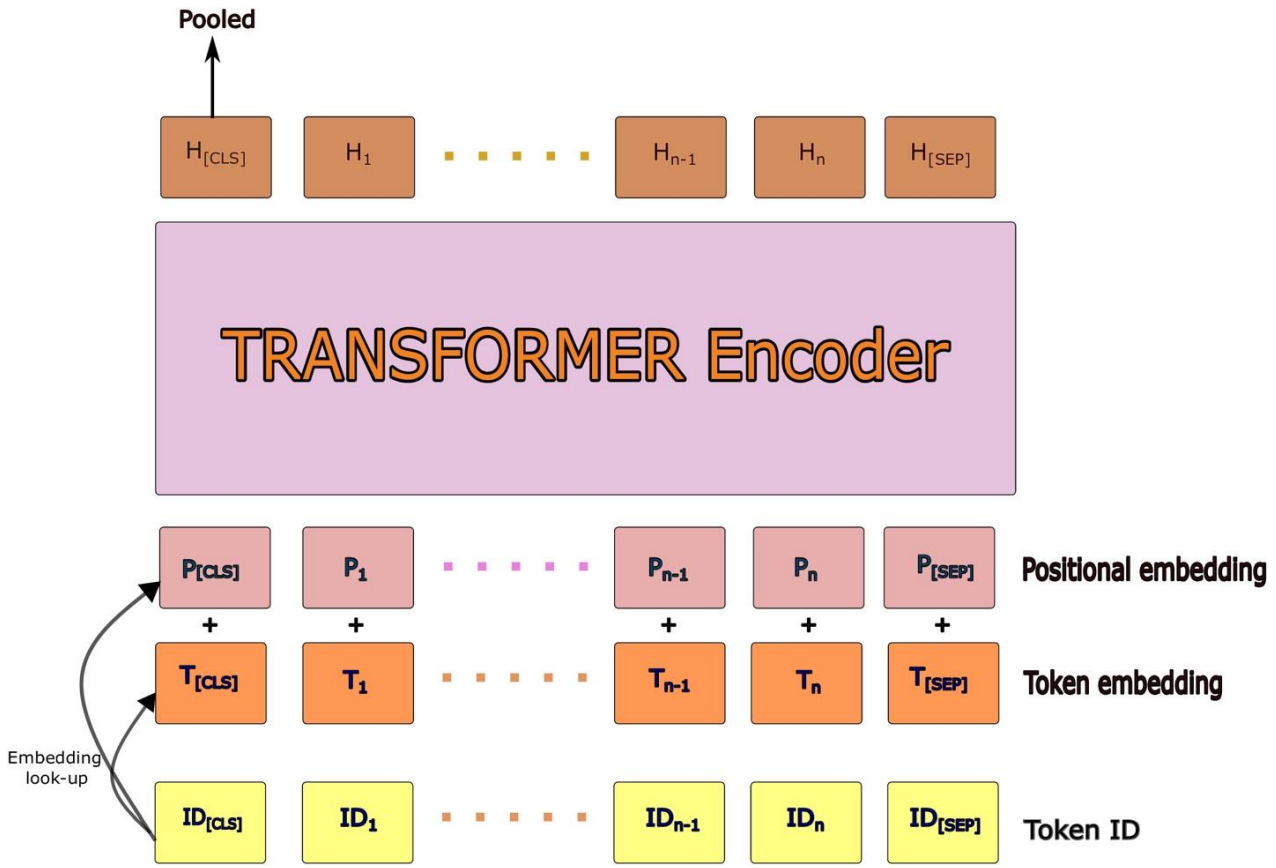
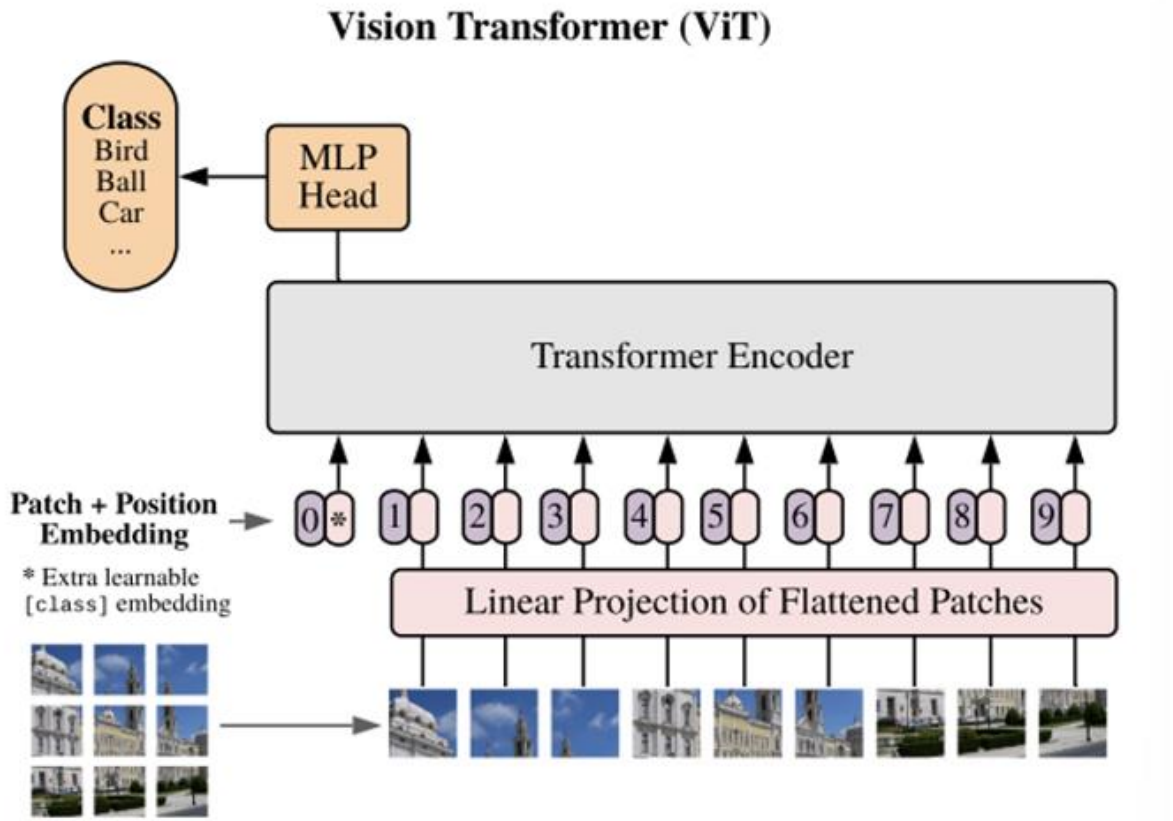


Illustration of language models

3. Vision Transformer



ViT Architecture

CLIP

Paper: Learning Transferable Visual Models From Natural Language Supervision.

For image feature extractor: an Image with size HxW will be divided into n patches P with size pxp.

$$P \in \mathbb{R}^{P \times P}$$

$$P^{\oplus n} \xrightarrow{embedding} E_{image}{}^{\oplus n} \xrightarrow{encoder} H_{image}{}^{\oplus n}$$

$$F_{image} = H_{image}{}^{\oplus n}_{[0]}$$

For text feature extractor: an input sequence of words S = $H_{text}{}^{\oplus m}$ {w1,w2,..,wm} will produce .

$$F_{text} = H_{text}{}^{\oplus m}_{[EOS]}$$

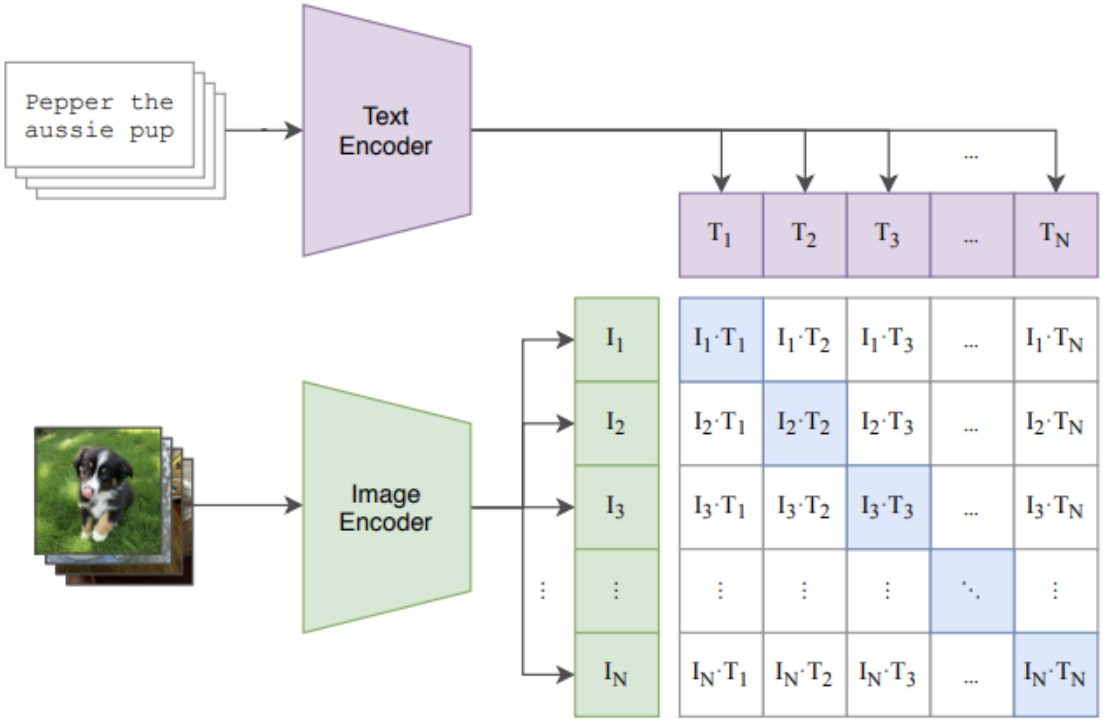


Illustration of CLIP

5. Loss Functions

BCE loss:

$$BCE = -\frac{1}{N}\sum_{i=0}^{N} y_i \cdot \log(p) + (1 - y_i) \cdot \log(1 - p)$$

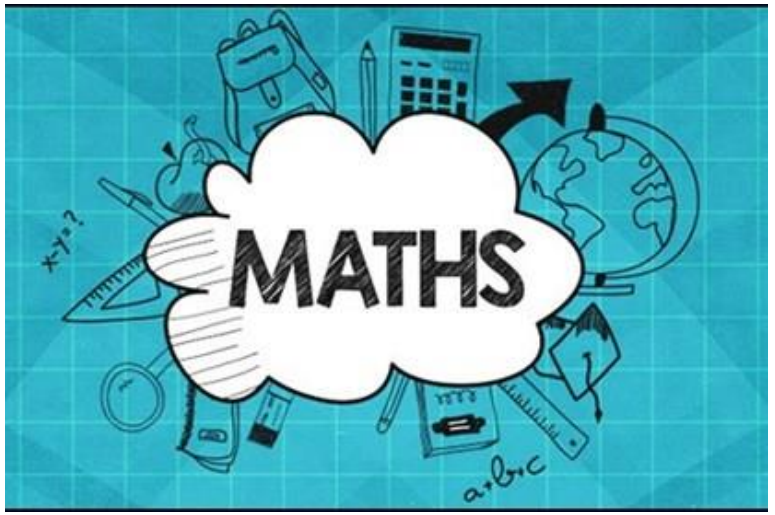
Re-balance distribution focal loss (paper: Distribution-Balanced Loss for Multi-Label Classification in Long-Tailed Datasets).

$$P_i^C(x^k) = \frac{1}{C}\frac{1}{n_i}$$

$$P^I(x^k) = \frac{1}{C}\sum_{y_i^k=1}\frac{1}{n_i}$$

$$r_i^k = \frac{P_i^C(x^k)}{P^I(x^k)}$$

$$\hat{r} = \alpha + \frac{1}{1 + \exp(-\beta \times (r - \mu))}$$



5. Loss Functions

We have distribution balanced loss (DB loss).

$$\mathcal{L}_{DB}(x^k, y^k) = \frac{1}{C}\sum_{i=0}^{C} \hat{r}_i^k \left[y_i^k \log\left(1 + e^{-(z_i^k - v_i)}\right) + \frac{1}{\lambda}\left(1 - y_i^k\right) \log\left(1 + e^{\lambda\left(z_i^k - v_i\right)}\right)\right]$$

We set p_+ as prediction of positive logits and p_- as prediction of negative logits.

$$p_+ = \frac{1}{1 + e^{-(z_i^k - v_i)}} \qquad p_- = \frac{1}{1 + e^{-\lambda(z_i^k - v_i)}}$$

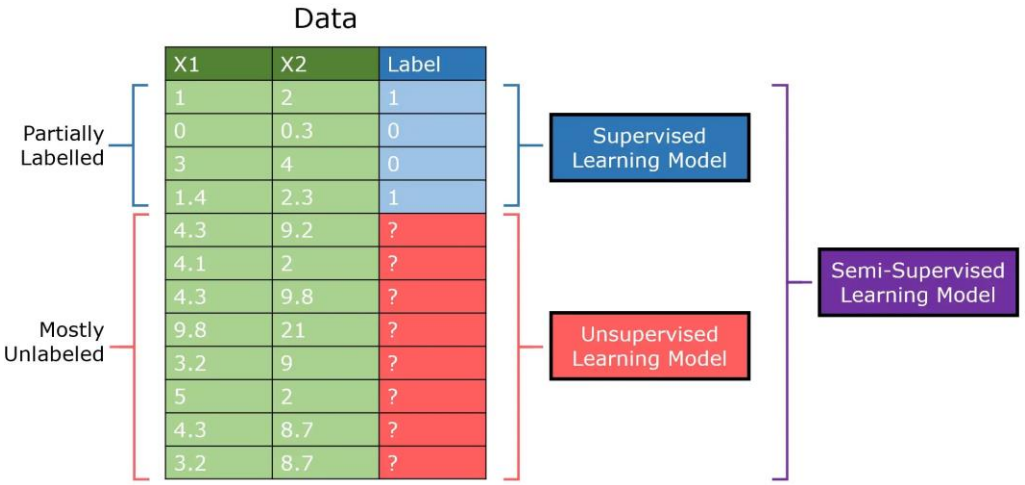
DB-loss can be re-written as:

$$\mathcal{L}_{\mathrm{DB}}(x^k, y^k) = -\frac{1}{C}\sum_{i=0}^{C} \hat{r}_i^k \left[y_i^k \log(p_+) + \frac{1}{\lambda}\left(1 - y_i^k\right) \log(1 - p_-)\right]$$

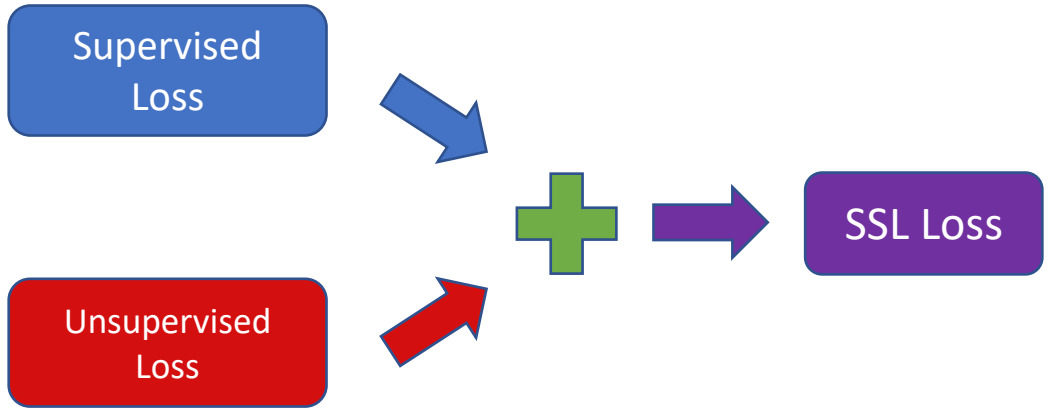
For focal version of DB loss can be written as:

$$\mathcal{L}_{DB-focal}(x^k, y^k) = -\frac{1}{C}\sum_{i=0}^{C} \hat{r}_i^k \left[(1 - p_+)^\gamma y_i^k \log(p_+) + \frac{1}{\lambda}(p_-)^\gamma \left(1 - y_i^k\right) \log(1 - p_-)\right]$$

6. Semi-supervised Learning (SSL)



Some Learning Methods (*)



SSL optimizes both Supervised and Unsupervised Loss

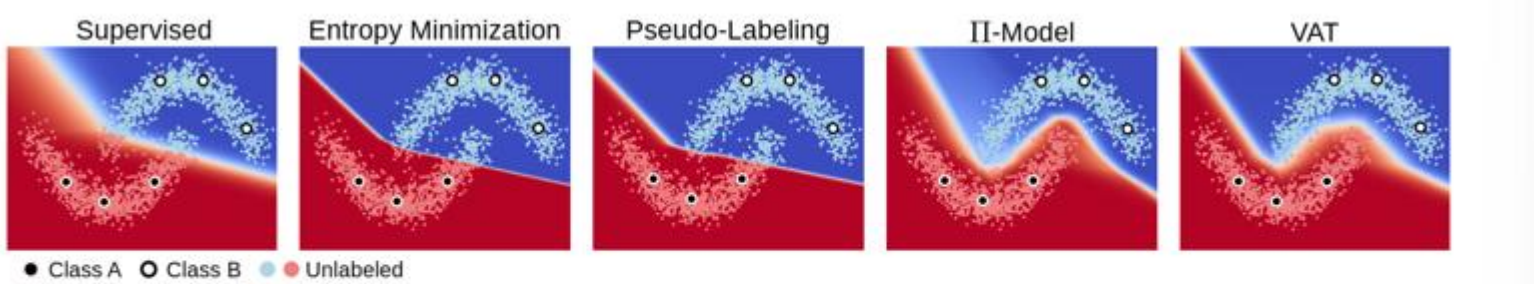
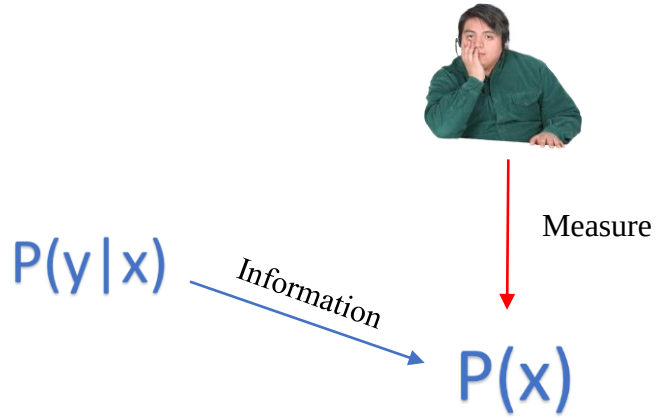


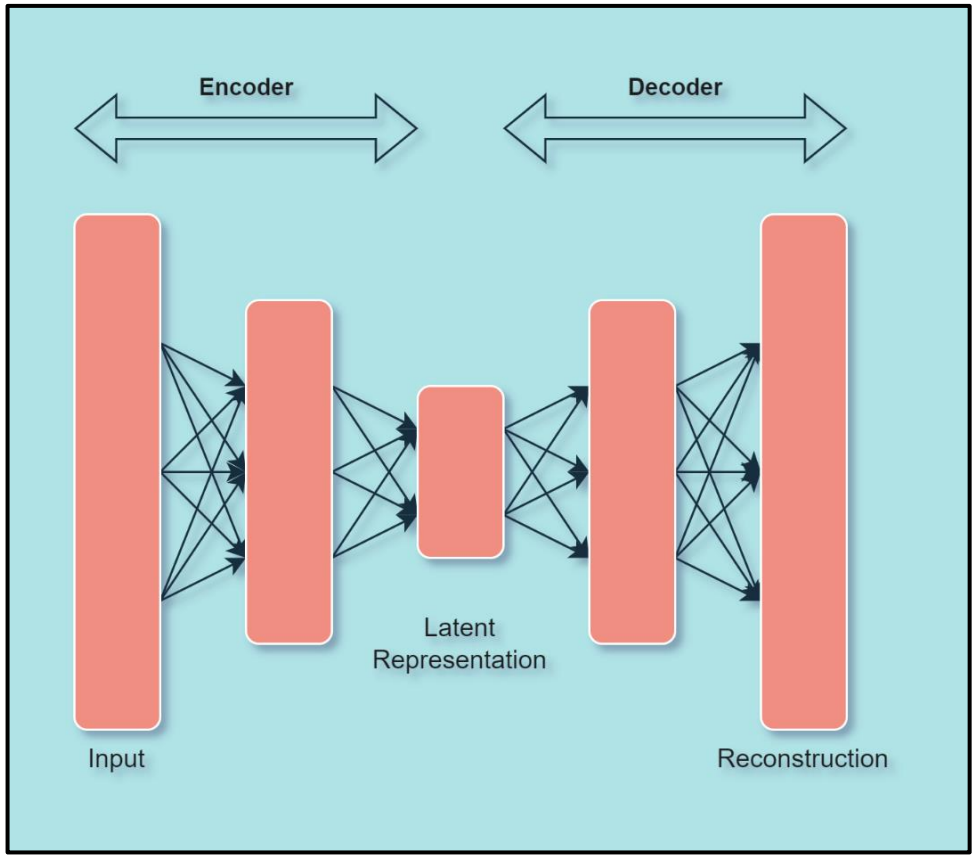
Figure 7. The decision boundaries of supervised and different SSL algorithms on a two-moons shape dataset, with 6 labeled samples, 3 for each class, and the remaining points as unlabeled data. [15]



P(x) might contain information about P(y|x). One can use unlabeled data to measure P(x), thereby P(y|x)

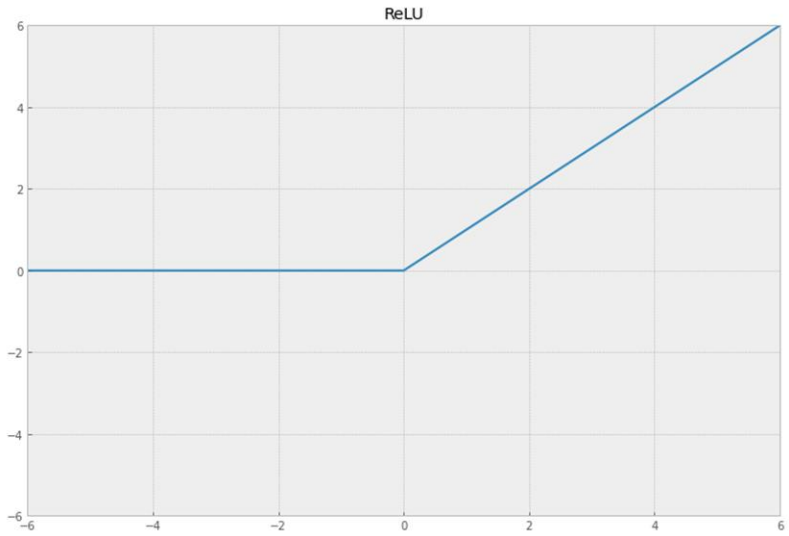
(*) Image from: https://jagan-singhh.medium.com/semi-supervised-learning-19e431be16e

7. Auto-Encoder

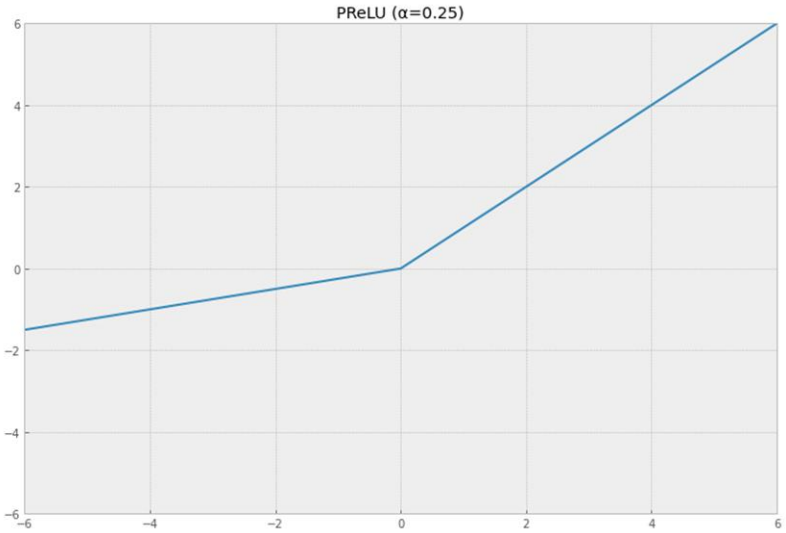


Autoencoder Architecture

8. Activation Functions



$$ReLU(x) = \begin{cases} x, & if\ x \geq 0 \\ 0, & otherwise \end{cases}$$



$$PReLU(x) = \begin{cases} x, & if\ x \geq 0 \\ \alpha x, & otherwise \end{cases}$$

Where α is a learnable parameter

If the nodes are converted to 0, then it will not make sense for the linear activation step in the next layer. The problem is called "Dying ReLU". Here comes PReLU (*).

(*) PReLU was introduced in 2015 in the paper “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification” as the first time a machine learning model beats humans on the image classification task.

DATA

1. Overview Data

2. Data Preprocessing

1. Overview Data

The contest: Multimedia Automated Misogyny Identification (MAMI) [1]

Use both available images and text to identify inappropriate memes for women. The contest task for their MAMI dataset consists of two main functions:

Task A: Identify memes with hateful content. The memes will be classified as either hate women or not hate women. The task is equivalent to a binary classification task.

Task B: Identify memes with misogynistic content by incorporating identification of categories such as stereotyping, shame, protest, and violence. The task is equivalent to a multi-label classification task with 4 binary labels.

[1] E. Fersini *et al.*, "SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification," in *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, Seattle, United States, Jul. 2022, pp. 533–549. doi: 10.18653/v1/2022.semeval-1.74.

1. Overview Data

Dataset MAMI

Sample

Train: 10 000
Test: 1000

Train: 8000
Val: 2000
Test: 1000

Training datasets

Labeled data: 2000
Unlabeled data: 6000

	file_name	Text Transcription	misogynous	shaming	stereotype	objectification	violence
0	5532.jpg	My Mom telling me behave or I'll end up like t...	0	0	0	0	0
1	3144.jpg	1511 BestDemotivationalPosters.com CORN DOG? Y...	1	0	1	1	0
2	7755.jpg	A PROSTITUTE GAVE ME ALL MONEY SHE MADE FOR TH...	0	0	0	0	0
3	4959.jpg	IM GONNA SLAP ÎŸÎ¤Î— SAN ANICE PACK ON RIHANNA...	1	0	0	0	1

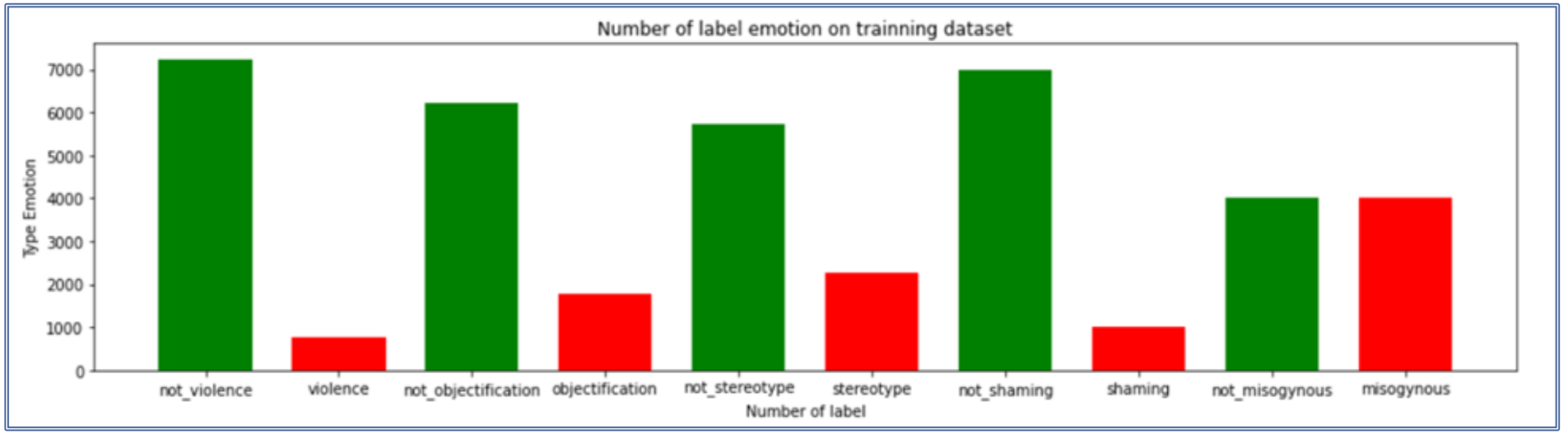
Content: dataset MAMI

1. Overview Data

Column Name	Meaning
file_name	Name of image file
Text Transcription	Extracted OCR Text in Meme
misogynous	Memes are classified as misogynous ("1") or not misogynous ("0"). Used for task A.
shaming	Memes are classified as shaming ("1") or not shaming ("0"). Used for task B.
stereotype	Memes are classified as a stereotype ("1") or not stereotype ("0"). Used for task B.
objectification	Memes are classified as objectification ("1") or not objectification ("0"). Used for task B.
violence	Memes are classified as violent ("1") or not violent ("0"). Used for task B.

Table 1: Meaning columns in metadata of MAMI dataset

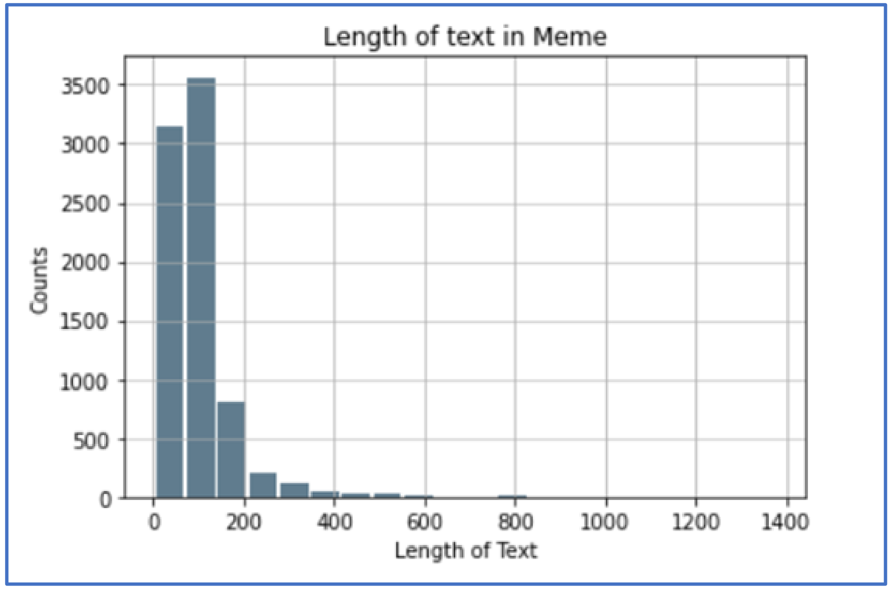
1. Overview Data



Number of samples by category

2. Data Preprocessing

Example meme with short text in MAMI



Histogram length of Text

Word cloud in the training dataset

2. Data Preprocessing

Text

+ Remove URL mixed-in text
+ Remove non-ASCII characters
+ Convert all characters to lowercase
+ Remove punctuation.

Image

+ Resized to a square image with a size of 768x768 to fit the size of the pretrain model

Some examples of memes that used in the MAMI dataset with different size

METHODOLOGY

1. Overview

2. Cross Modality Auto Encoder (CROM-AE)

3. Raw and Cooked Features Classification Model (RAW-N-COOK)

1. Overview

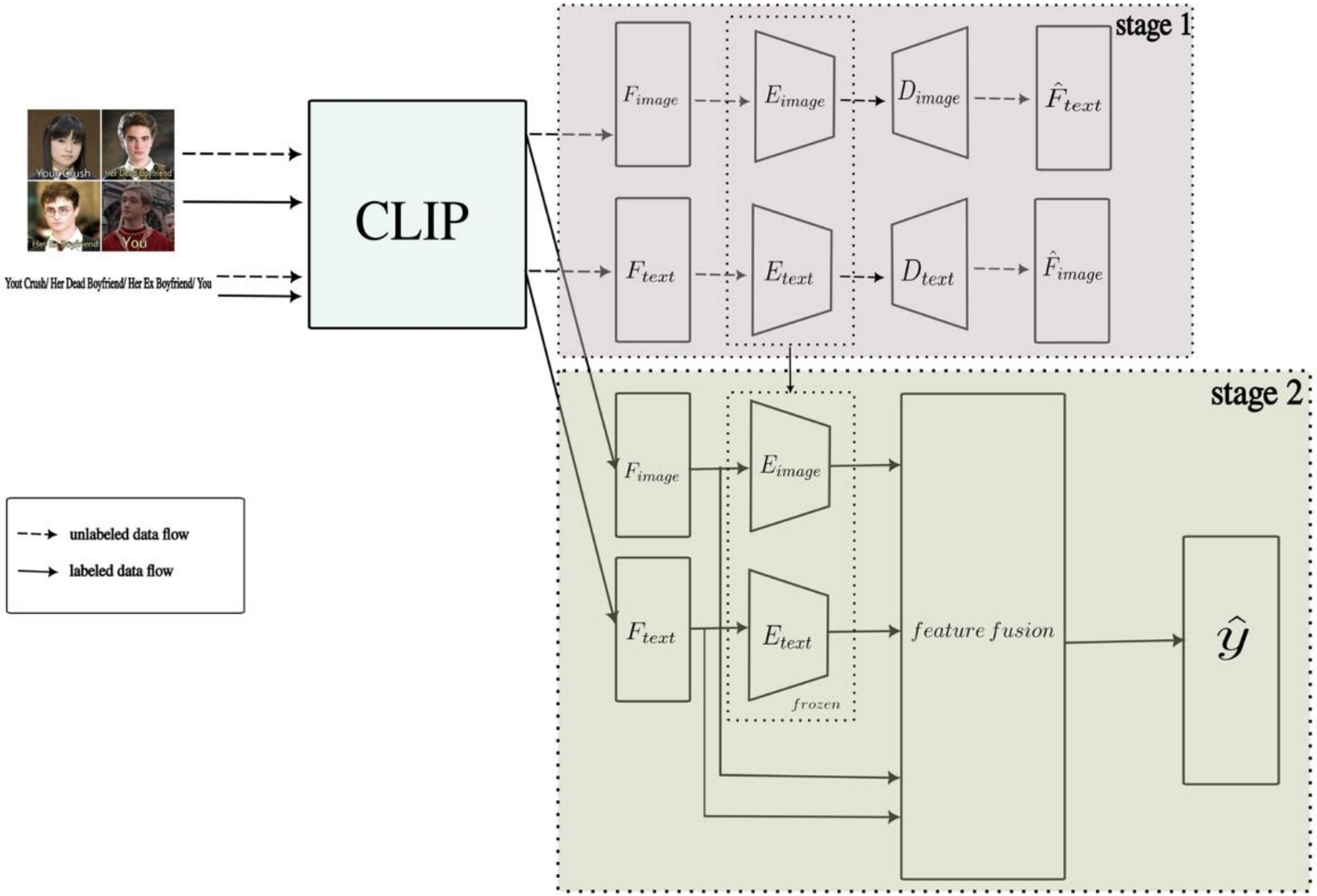


Figure 18 Overall our SSL methods

2. Cross Modality Auto Encoder (CROM-AE)

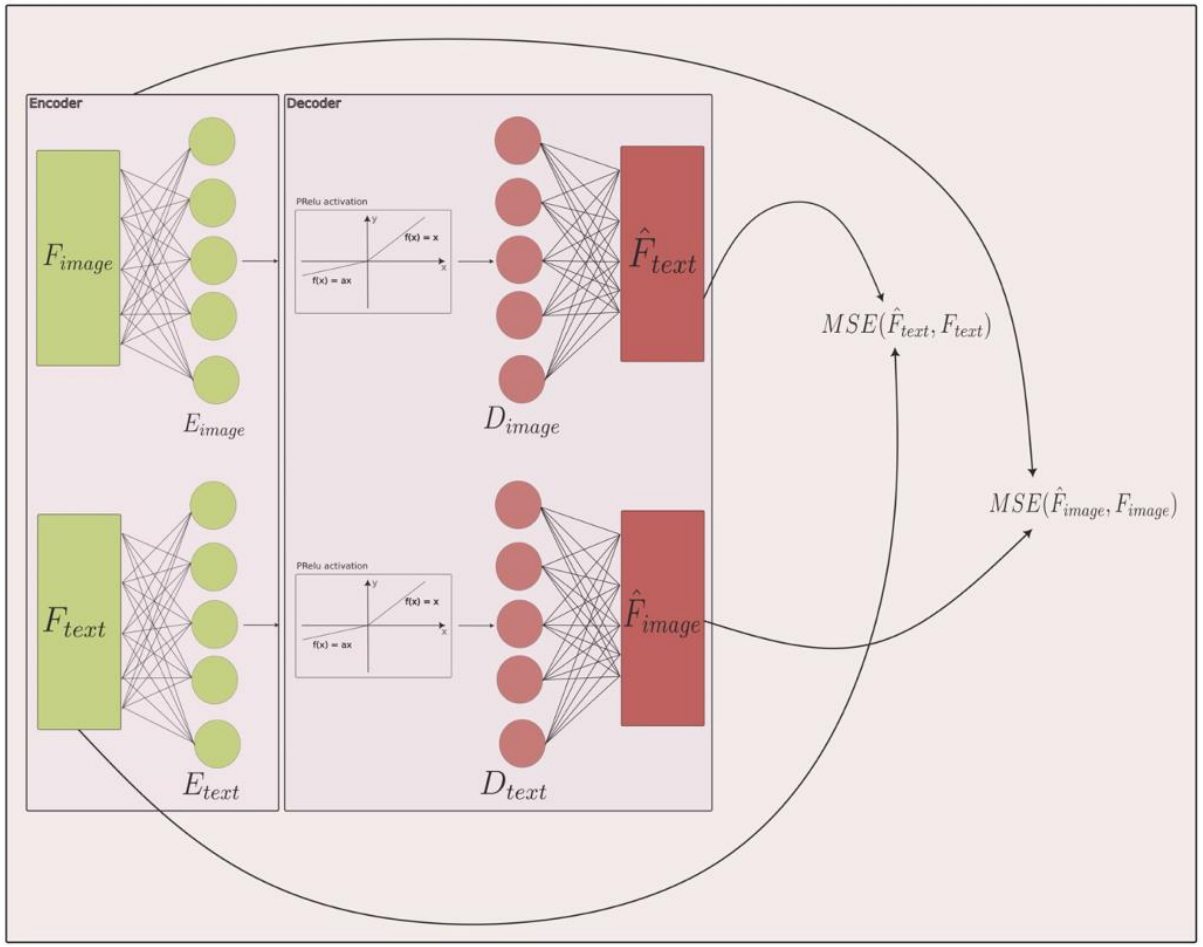


Figure 19 The pipeline of CROM-AE

$$\hat{F}_{text} = AE_{image}(F_{image})$$

$$\hat{F}_{image} = AE_{text}(F_{text})$$

$$\mathcal{L}_{AE_{image}} = MSE(\hat{F}_{text}, F_{text})$$

$$\mathcal{L}_{AE_{text}} = MSE(\hat{F}_{image}, F_{image})$$

Encoder	Input
	Linear
Decoder	PReLU
	Linear
	Output

Detail Architecture

3. Raw and Cooked Features Classification Model (RAW-N-COOK)

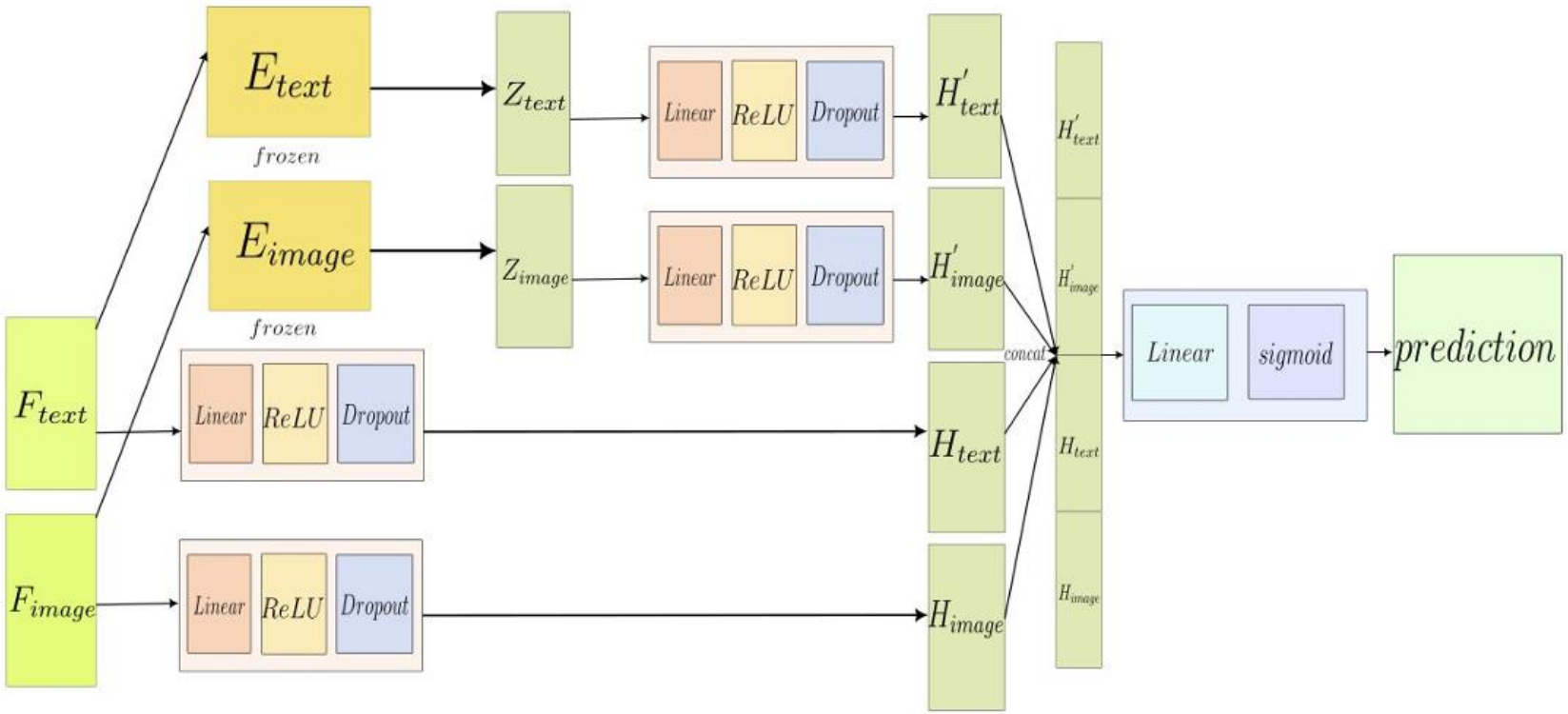


Figure 20 The architecture of our proposed finetune model

RAW-N-COOK fuses latent representation from encoders (cooked features) and original CLIP features (raw features). The model is trained with distribution balanced loss.

EXPERIMENT

1. Evaluation metric
2. Experiment setting
3. Benchmark
4. Experiment result
5. Analysis
6. Ablation study

1. Evaluation metric

F1 weighted score: weighted average
The sum of the F1 score.

$$F1(class = 1) = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

$$F1(class = 0) = \frac{TN}{TN + \frac{1}{2}(FP + FN)}$$

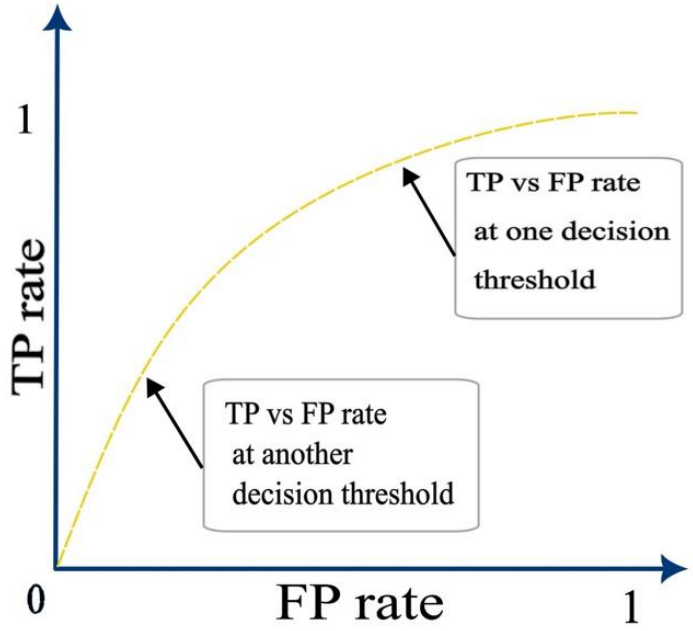
$$F1 - macro = \frac{F1(class = 0) + F1(class = 1)}{2}$$

$$F1 - weighted = \frac{1}{\sum_i support_i} \sum_{i=1}^{n} F1 - macro_i \cdot support_i$$

AUC score: is the are under the ROC curve.

$$TPR = \frac{TP}{TP + FN} \quad (23)$$

$$FPR = \frac{FP}{FP + TN} \quad (24)$$



2. Experiment setting

We benchmark on the MAMI dataset and compare it with 7 methods: CMML-CLIP, CMML-ORIGIN,TIB-VA, Visual Bert,Extpienet, Vilt, Lxmert.

Tuning gamma for different label ratio data setting:

Ratio	γ (StepLR)	Epochs	Learning rate	Batch size	Optimizer
0.3	0.85	50	1e-4	40	Adam
0.1	0.9	50	1e-4	40	Adam
0.05	0.93	50	1e-4	40	Adam

Table 2 Hyperparameters gamma for different settings

3. Benchmark

Ratio	Ours	CMML origin	CMML-CLIP	TIB-VA	VisualBERT	Extpienet	ViLT	Lxmert
0.3	0.7433	0.5673	0.7265	0.7083	0.6337	0.4422	0.6010	0.6005
0.1	0.7184	0.5496	0.6872	0.5885	0.5583	0.4139	0.5566	0.5597
0.05	0.6792	0.5256	0.6496	0.4150	0.5174	0.4139	0.5456	0.5303

Table 3 Weighted-average F1-Measure on Test Set

Ratio	Ours	CMML origin	CMML-CLIP	TIB-VA	VisualBERT	Extpienet	ViLT	Lxmert
0.3	0.8310	0.6234	0.8289	0.8333	0.6825	0.5043	0.6948	0.6684
0.1	0.8145	0.5794	0.7956	0.7901	0.6119	0.4833	0.6312	0.6163
0.05	0.7989	0.5647	0.7664	0.7093	0.5604	0.4761	0.5978	0.5986

Table 4 AUC Measure on Test Set

4. Experiment result

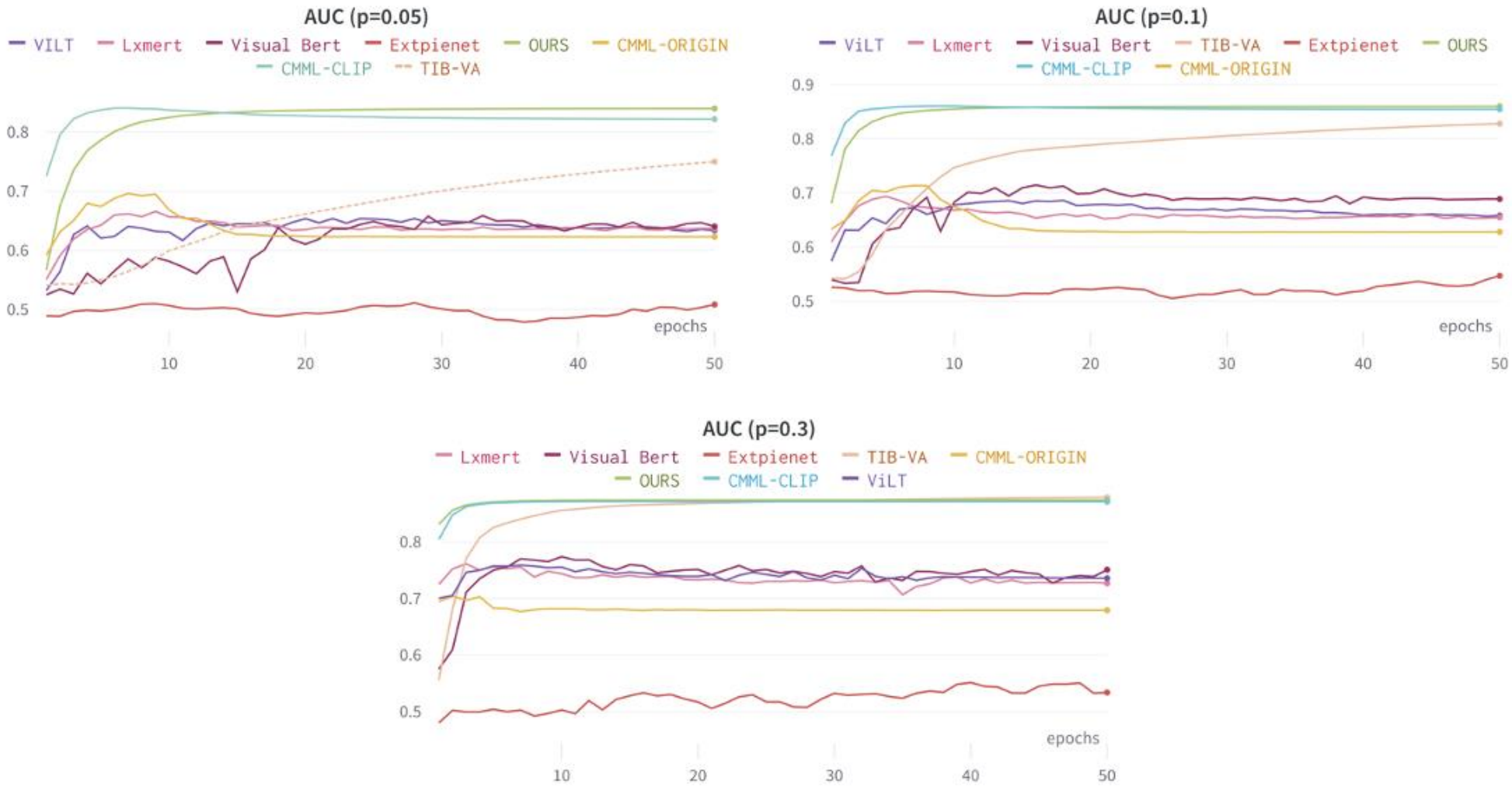


Figure 22 AUC scores on validation set benchmarked on 8 methods with 50 epochs training

4. Experiment result

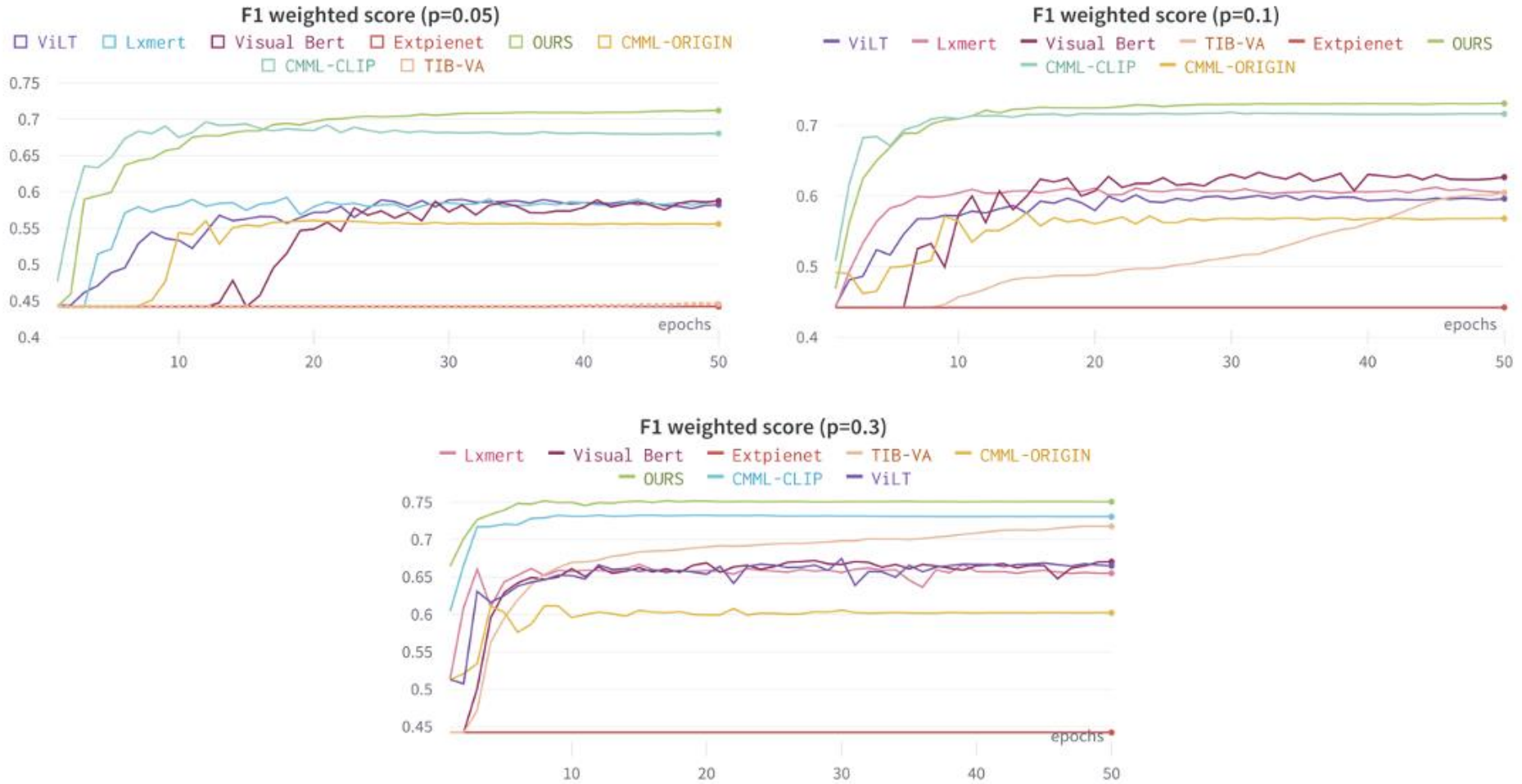
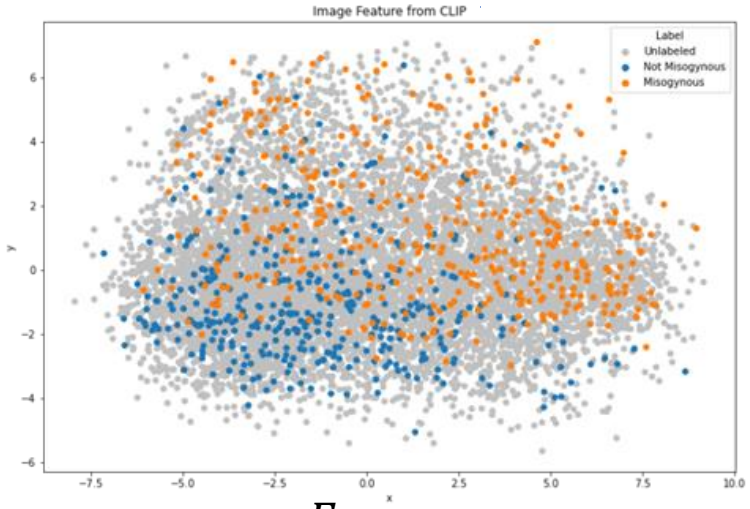
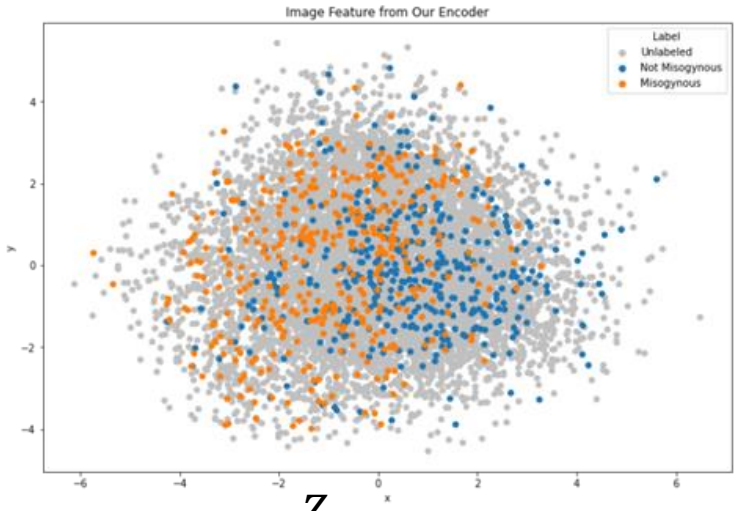


Figure 23 F1 weighted scores on validation set benchmarked on 8 methods with 50 epochs training

5. Analysis

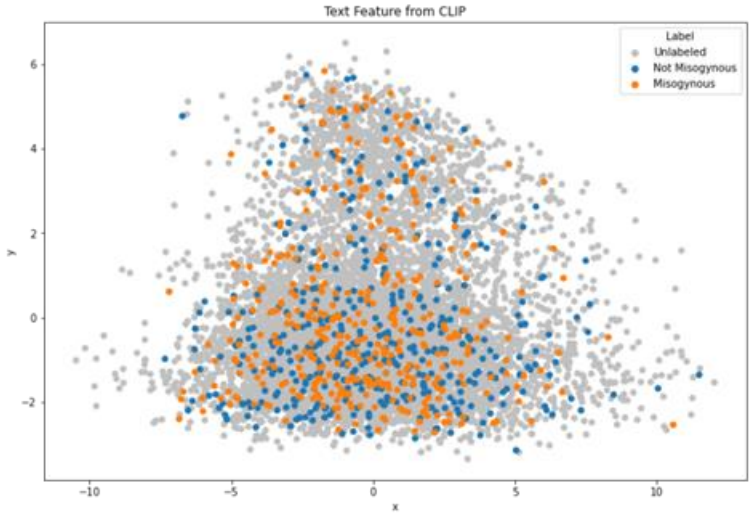


F_{image}

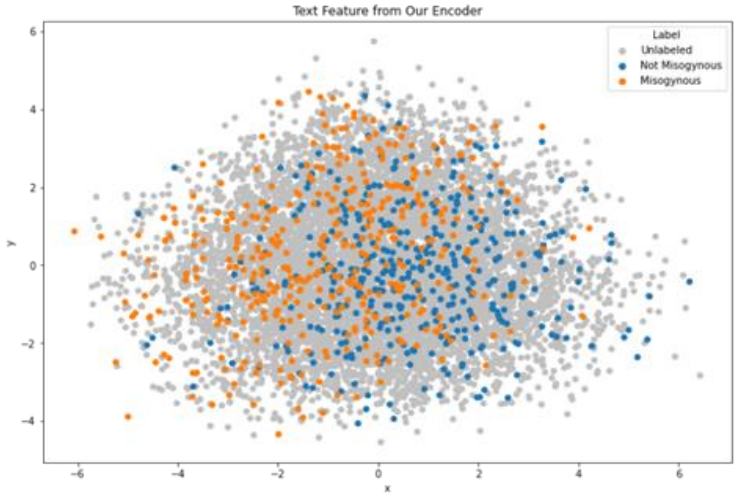


Z_{image}

Cooked feature Z_{image} keeps the disentanglements between classes, like raw features F_{image}.



F_{text}



Z_{text}

Cooked feature Z_{text} disentangles classes so much, even without seeing any labels, compare to raw features F_{text}.

6. Ablation study

Ratio	Ours	w/o CROM-AE encoders	w/o DB loss	w/o DB loss + CROM-AE encoders
0.3	0.7433	0.7365	0.7316	0.7074

Table 5 Ablation: Weighted-average F1-Measure on Test Set

CONCLUSION

1. Conclusion

+ Successfully applied semi-supervised learning for sentiment analysis of memes in the image macros form
+ Created a SOTA approach employing multiple improvements: CLIP features, unsupervised pre-train on unlabeled data, integrating pre-train models into the supervised model, and applying a balance loss function
+ Help to leverage a huge amount of unlabeled memes on the internet and solve the annotation process's pain points

2. Future works

+ Benchmark on other datasets
+ End-to-end approach

Thanks for
Your attention

Any questions please contact our via email

phamthaihoangtung@gmail.com
vietnguyen@gmail.com
ngotienanhmathk27@gmail.com

Q & A