MMVAE at SemEval-2022 Task 5: A Multi-modal Multi-task VAE on Misogynous Meme Detection

Yimeng Gu
Queen Mary
University of London
United Kingdom
yimeng.gu@qmul.ac.uk

Ignacio Castro
Queen Mary
University of London
United Kingdom
i.castro@qmul.ac.uk

Gareth Tyson
Queen Mary
University of London
United Kingdom
g.tyson@qmul.ac.uk

Abstract

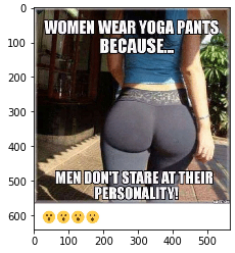
Memes have become quite common in day-to-day communications on social media platforms. They often appear to be amusing, evoking and attractive to audiences. However, some memes containing malicious content can be harmful to targeted groups. In this paper, we study misogynous meme detection, a shared task in SemEval 2022 - Multimedia Automatic Misogyny Identification (MAMI). The challenge of misogynous meme detection is to co-represent multi-modal features. To tackle with this challenge, we propose a Multi-modal Multi-task Variational AutoEncoder (MMVAE) to learn an effective co-representation of visual and textual features in the latent space. Our goal is to automatically determine if a meme contains misogynous information and then identify its fine-grained category. Our model achieves F_1 scores of 0.723 on the MAMI sub-task A and 0.634 on sub-task B. We carry out comprehensive experiments on our model's architecture and show that our approach significantly outperforms several strong uni-modal and multi-modal approaches. Our code is released on github[1].

1 Introduction

With the rapid development of social media, the use of image-based memes has been growing. People use memes for various purposes, such as to express humor (Velioglu and Rose, 2020), or to attract greater attention. Enabling this, simple websites that allow people to easily create new memes have further seen their proliferation.

However, this easy composition has allowed people to easily embed harmful messages within memes, often circumventing more traditional text-based moderation tools (Malmasi and Zampieri, 2017). This paper particularly focuses on misogynous memes (*i.e.* hatred towards women) — see Figure 1 for two examples. As some platforms may choose to limit the sharing of such material, we argue it is vital to build tools that can automatically identify misogynous memes at scale.

(a) A misogynous meme with an uncorrelated image.



(b) A misogynous meme with a correlated image.

Figure 1: Examples of misogynous memes. In (a), the embedded text is misogynous. In (b), both the image and the embedded text are misogynous.

Many prior works have sought to analyze the content of memes. Some have looked into meme emotion analysis (Sharma et al., 2020; Smitha et al., 2018), meme ecosystem measurements (Zannettou et al., 2018) and meme auto-generation (Vyalla and Udandarao, 2020). Past efforts have also shed light on hateful meme detection (Kiela et al., 2020) or offensive meme detection (Sabat et al., 2019) more generally.

This paper builds on these prior works to automatically detect *misogynous* meme content (Fersini et al., 2022). This comes with several key challenges. First, a meme usually comes with both a visual and a textual part. Sometimes the standalone image or text is not necessarily hateful or toxic, but when combined together, the semantic meaning becomes harmful. To effectively understand the semantic meaning of a meme, information encoded in both modalities should be considered. Second, different memes frequently have the same image, but are embedded with different text (and thus have different and even opposite meanings). This makes reliance on image-hash lists ineffective. It is also common that the image and text of a meme are unrelated, as in Figure 1a. In this case, finding an

[1] https://github.com/MMVAE-project/MMVAE

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 700 - 710

accurate co-representation of both visual and textual features is vital. Finally, malicious information contained in a meme can have granular labels and even belong to multiple categories. Thus, it is often necessary to devise a more nuanced taxonomy.

With the above challenges in mind, we propose MMVAE: a pipeline to determine if a meme contains misogynous information, and to identify its fine-grained hateful labels accordingly. Specifically, our contributions are as follows:

1. We propose a Multi-modal Multi-task Variational AutoEncoder (MMVAE) that effectively co-represents the textual and visual features of the meme. We use this to predict if a meme is misogynous/non-misogynous. We further expand our model to predict more fine-grained labels, *e.g.* shaming, violence, objectification.

2. We evaluate our model on SemEval competition Task 5: Multimedia Automatic Misogyny Identification (Fersini et al., 2022). We achieve an F_1 score of 0.723 on sub-task A and 0.634 on sub-task B.

3. We analyze the strengths and weaknesses of our multi-modal approach by presenting a text-based error analysis and case study. Our model is better at identifying misogynous memes with a high precision, yet less effective in determining the correct label for non-misogynous memes.

2 Background

Our approach leverages multi-modal learning, pre-trained model, variational autoencoder and multi-task learning. Below, we provide a brief overview of these concepts.

2.1 Multi-modal Feature Representation

We leverage multi-modal learning to co-represent features from both image and text. Common multi-modal features co-representation techniques (Zeng et al., 2021) include fusion mechanism (early at feature level (Su et al., 2020), or late at decision/scoring level (Poria et al., 2016)), tensor factorization (Zadeh et al., 2017; Mai et al., 2019) and complex attention mechanisms which can be further classified as dot-product attention (Yu et al., 2021), multi-head attention (Cao et al., 2021; Wu et al., 2021), hierarchical attention (Pramanick et al., 2021), attention on attention (Liu et al., 2021) among others. These techniques have proven to be effective in encoding multi-modal features. However, leveraging multi-modal features doesn't always enhance the task performance. Hence, another topic that has raised concern in multi-modal learning is the uni-modal contribution analysis. A straightforward approach (Hessel and Lee, 2020; Frank et al., 2021) is to ablate cross-modal inputs or interactions in order to evaluate the model's performance on uni-modal data. Zeng et al. (2021) demonstrated that multi-modal models might not achieve optimal performance because there are noises contained in each modality.

2.2 Pre-trained Model

We adopt pre-trained models to embed the text and image of a meme.

Usually trained on large dataset corpus, pre-trained models have achieved state-of-the-art (SOTA) performances on various Natural Language Processing (NLP) and Computer Vision (CV) benchmark tasks. Thus, they can be used as a powerful embedding tool or be easily fine-tuned on downstream tasks. Popular language pre-trained models such as BERT (Devlin et al., 2019), LASER (Artetxe and Schwenk, 2019) and LaBSE (Feng et al., 2020) have constantly updated the SOTA performance on downstream NLP tasks. Similarly, image pre-trained models including ResNet (He et al., 2016), VGG (Simonyan and Zisserman, 2015), and Inception (Szegedy et al., 2016) have proved their effectiveness in multiple tasks. Recently, there has been increasing interest in pre-training multi-modal models. Lu et al. (2019) proposed ViLBert, which extends BERT to multi-modal two-stream models that interacts through co-attentional transformer layers and can be easily transferred into performing multiple visual-and-language tasks. Similarly, LXMERT (Tan and Bansal, 2019) further included a cross-modality encoder that captures cross-modality relationships. Instead of implementing two stream models, Su et al. (2020) input the caption and image regions all together to the modified BERT model named VL-BERT. Radford et al. (2021) jointly trained an image encoder and a text encoder known as CLIP to match the image and its corresponding caption by contrastive learning.

2.3 VAE Overview

We later use Variational AutoEncoder (VAE) to fuse multi-modal features. Thus, here we introduce

the basic structure and loss calculation of a VAE. A VAE is an unsupervised approach for learning a lower-dimensional feature representation from unlabeled training data (Li et al., 2019a). The aim of finding a representation in the latent space is to capture meaningful factors of variation in the data (Li et al., 2019a). Typically, a VAE consists of an encoder and a decoder, which look "symmetrical" in the model's architecture. The encoder learns a latent variable space where a latent variable z is sampled from and input to the decoder for original data reconstruction. Figure 2 shows the structure of a simple VAE.

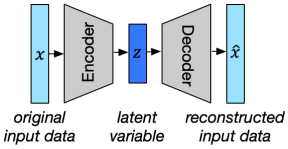


Figure 2: A simple VAE example.

Encoder. The encoder network of a VAE can have various structures. For textual input, the common structure is the Recurrent Neural Networks (RNN), *e.g.* Bi-LSTM (Cheng et al., 2020). For image input, the common encoder structure is the Convolutional Neural Networks (CNN). Mathematically, the encoder can be described as $q_{\phi}(\mathbf{z}|\mathbf{x})$, where ϕ is the parameters of the encoder network. $q_{\phi}(\mathbf{z}|\mathbf{x})$ stands for the probability distribution of latent variable z given x.

Decoder. Generally, the decoder network of a VAE is symmetrical to the encoder. It reconstructs the input by sampling from the learned latent variable space subject to a Gaussian distribution. The decoder can be described as $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$, where $\boldsymbol{\theta}$ is the parameters of the decoder network. $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ stands for the probability distribution of reconstructed x given z.

Loss. In order to compute the loss of VAE in our model, we first introduce how to compute the loss of a general VAE. Since VAE models the probabilistic generation of data $\{x^{(i)}\}$, the goal is to maximize the (log) data likelihood (Kingma and Welling, 2014):

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}) = D_{KL}(q_{\phi}(\mathbf{z}|\mathbf{x}^{(i)})||p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}^{(i)})) + \mathcal{L}(\boldsymbol{\theta}, \phi; \mathbf{x}^{(i)}) \quad (1)$$

where $\boldsymbol{\theta}$ and ϕ are parameters of the decoder and encoder network respectively. The first KL divergence term on the right hand side can't be computed explicitly but it is non-negative. The second term is called the *lower bound* on the marginal likelihood of datapoint i and can be rewritten as (Kingma and Welling, 2014):

$$\mathcal{L}(\boldsymbol{\theta}, \phi; \mathbf{x}^{(i)}) = -D_{KL}(q_{\phi}(\mathbf{z}|\mathbf{x}^{(i)})||p_{\boldsymbol{\theta}}(\mathbf{z})) + \mathbf{E}_{q_{\phi}(\mathbf{z}|\mathbf{x}^{(i)})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}|\mathbf{z})\right] \quad (2)$$

Hence, to maximize the log likelihood, we only need to maximize $\mathcal{L}(\boldsymbol{\theta}, \phi; \mathbf{x}^{(i)})$. To achieve that, the KL divergence term in eq. 2 should be minimized and the expected log likelihood of datapoint i should be maximized (equivalent to minimizing the expected reconstruction error).

2.4 Multi-task Learning

Our work leverages the concept of multi-task learning to learn fine-grained category labels of misogynistic memes. According to Ruder (2017), as long as a model is optimized by more than one loss functions, it is doing multi-task learning. When the tasks are relevant and the knowledge learned from one task could benefit the learning of other tasks, applying multi-task learning will have promising performances (Caruana, 1997). Multi-task learning has been successfully applied in many NLP tasks such as text classification (Cheng et al., 2020; Khattar et al., 2019), sentiment analysis (Majumder et al., 2019), neural machine translation (Niehues and Cho, 2017), etc. However, the most common issue of multi-task learning is the negative transfer when the performance on single task is undermined. Lee et al. (2016) avoids negative transfer by allowing asymmetric transfer between tasks. Wu et al. (2020) observes that misalignment between tasks can cause negative transfer. Wu et al. (2019) proposed a method to filter and select shared feature to prevent adverse features being integrated into certain tasks.

3 Problem Statement

The paper strives to build a classifier that can distinguish misogynous *vs.* non-misogynous memes. We break this down into two sub-tasks:

- Sub-task A: Given a meme's image I and its corresponding text transcription T, we must predict its binary (0/1) label on *misogyny* (Fersini et al., 2022).
- Sub-task B: Given a meme's image I and its corresponding text transcription T, we must

predict its binary (0/1) label on the following classes: *shaming*, *stereotype*, *objectification* and *violence* (Fersini et al., 2022), which are types of misogyny.

Note, the meme's image I has the embedded text on it, and T is transcribed from the meme's embedded text.

4 Our Approach

As illustrated in Figure 3, our MMVAE model consists of 3 components: (*i*) the Image/Text embedding Module, (*ii*) the Variational AutoEncoder Module and (*iii*) the Multi-Task Learning Module. The embedding module turns the inputs into unimodal vector representations. After that, the VAE module fuses multi-modal representations and generates a co-representation. Finally, the multi-task learning module gives the label prediction. In this section, we first introduce each part of our proposed model, and then demonstrate how to put them all together and jointly train the model.

4.1 Image/Text Embedding Module

In our pipeline, we first adopt two pre-trained models to embed the meme's text and image (illustrated in Figure 3 left part). We directly input the provided text transcription and meme's image to the pre-trained models to get embeddings as module output. However, when using ResNet-50 and BERT as pre-trained models, we apply data preprocessing beforehand.

Image Embeddings. To obtain the meme's image embedding, we have experimented with 2 different pre-trained models: ResNet-50 (He et al., 2016) pre-trained on ImageNet (Deng et al., 2009) and the multi-lingual version of OpenAI CLIP-ViT-B32 (Radford et al., 2021). When using ResNet-50 to embed images, we transform the input images by resizing it to 224×224, applying random rotation, random horizontal flip, random crop and normalization for data augmentation. We use the last fully connected layer's output as our image embedding. When using CLIP to embed images, although we have experimented with data transformation techniques, our optimal performance is achieved by directly embedding the raw image.

Text Embedding. To obtain the meme's text embeddings, we experiment with 4 different pre-trained models: BERT for sentence classification (Devlin et al., 2019), LASER (Artetxe and Schwenk, 2019), LaBSE (Feng et al., 2020) and the multi-lingual version of OpenAI CLIP-ViT-B32. We set the maximum input sentence length to 512 tokens when using BERT while we directly input the meme's text to other pre-trained models.

4.2 Variational AutoEncoder (VAE) Module

The next module in our pipeline takes the embeddings as an input and finds a multi-modal co-representation. We assume that there is an effective multi-modal co-representation in the latent space which can better capture the inter-relationship between text and image data.

In our pipeline, we leverage VAE to learn the multi-modal co-representation (illustrated in Figure 3 middle part). The input to this module are the embeddings generated from the pre-trained models, and the output is the reconstructed embeddings. Yet what we need for later meme detection classifiers is the latent variable generated from the VAE encoders.

Encoders. We build a text encoder and an image encoder to first learn the latent variables of the text embedding and image embedding separately. In our case, since the input embeddings are already semantically meaningful, there is less need to further extract semantic information using complex and deep layers. Therefore, we decide to use 1 fully connected layer for each modality as the encoder structure. And then we concatenate the learned latent variables to form a multi-modal co-representation z in the latent space.

Decoders. Accordingly, we build a text decoder and an image decoder to reconstruct the text embedding and image embedding from the learned multi-modal co-representation z. Our decoders still consists of 1 fully connected layer, with the number of output channels equal to the number of input channels in the encoder network.

Losses. We calculate the reconstructed image embedding loss $\mathcal{L}_{img}$, the reconstructed text embedding loss $\mathcal{L}_{txt}$ (corresponding to the 2nd part of eq. 2) and the KL loss KLD (corresponding to the 1st part of eq. 2) from the sampled latent variable z for gradient descent. The reconstruction loss is calculated by L_2 loss function (squared error). The KL loss is calculated using the formula in Appendix B of Kingma and Welling (2014).

4.3 Multi-task Learning

We leverage multi-tasking learning because we expect that learning the fine-grained hateful class labels will benefit the misogyny meme detection and

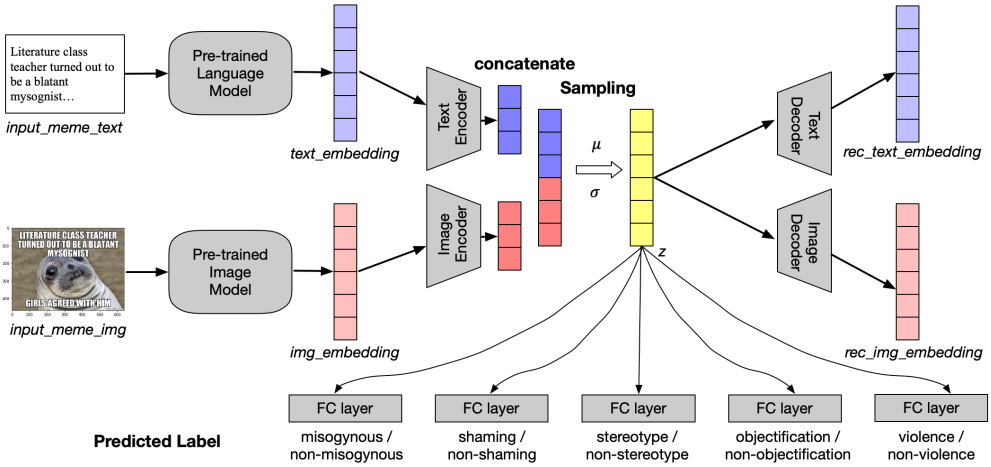


Figure 3: The architecture of our proposed Multi-modal Multi-task VAE (MMVAE) model. The upper part illustrates the image/text embedding module and the VAE module from left to right. The lower part shows the multi-task learning module.

vice versa. The input to this module is the latent variable z (colored yellow in Figure 3) learned in Section 4.2 and the output is label predictions on each class.

Our model learns 5 tasks at the same time: misogyny detection, shaming detection, stereotype detection, objectification detection and violence detection. The latter 4 classes are more specific types of misogyny. Note, if a meme is misogynous, it could fall into more than one specific misogyny classes.

The architectures are the same across all the sub-networks in our model as shown in Figure 3: 1 fully connected layer followed by the softmax binary classifier. Yet the sub-networks are independent - the parameters are not shared.

Here the cross entropy loss function is used to calculate the loss for each task:

$$\mathcal{L}_t = -\mathbf{E}_{y\sim Y_t}\left[y \log y_t + (1-y)\log(1-y_t)\right] \tag{3}$$

where y is the ground-truth label of data point x in t detection (e.g. misogyny detection) and y_t is the predicted probability of x belonging to class t (e.g. misogyny).

4.4 Putting It All Together

We have introduced each module in our MMVAE and how to calculate their standalone losses. Next, we introduce how to jointly train the model by putting three modules together.

The total loss used to compute gradient descent is composed of the reconstructed image error L_{img}, reconstructed text error L_{txt}, KL divergence KLD, and multi-task cross entropy losses L_t, where $t \in \{\textit{misogyny}, \textit{shaming}, \textit{stereotype}, \textit{objectification}, \textit{violence}\}$. This can be expressed as:

$$\mathcal{L}_{total} = \lambda_i \mathcal{L}_{img} + \lambda_t \mathcal{L}_{txt} + \lambda_{kl} KLD + \sum_t \lambda_t \mathcal{L}_t \tag{4}$$

where λs are used to adjust the learning focus, so that we can direct the focus of learning to the task we care more about. We calculate gradient descent on $\mathcal{L}_{total}$ and back propagate it to update model's parameters.

5 Experimental Setup

In this section, we first introduce the Multimedia Automatic Misogyny Identification (MAMI) dataset released by the task organizer (Fersini et al., 2022) and then discuss our experimental settings.

5.1 Dataset

The MAMI dataset contains 10,000 memes in the training set and 1,000 memes in the test set. Each meme has an associated text transcription in English and labels on 5 classes. The label distributions of given classes are not balanced and are summarized in Table 1. Only in the misogyny class, the labels are totally balanced. Detailed task descriptions can be found in Section 3.

5.2 Experimental Settings

We have experimented with different latent variable (z) sizes of 256, 512, and 1024. 512 has the best

Class	# Positive	# Negative
Misogyny	5,000	5,000
Shaming	1,274	8,726
Stereotype	2,810	7,190
Objectification	2,202	7,798
Violence	953	9,047

Table 1: Number of positive labels and negative labels in each class.

performance on the validation set. Hence, we set the latent variable size to 512 during the evaluation phase.

Different pre-trained models generate embeddings with different dimensions, D. In both encoders, we map the input embeddings' dimension from D to half of z's dimension.

For the optimizer, we use the Adam optimizer with weight decay set to 1e-5. The batch size is set to 64. In addition, the initial learning rate is 1e-4 and divided by 4 every 8 training epochs. The model is trained for 30 epochs in total with early stopping.

We split the training set into the training and validation set with a ratio of 9:1. We train our model on 9,000 memes from the original training set and test it on the rest of 1,000 memes to evaluate our model's performance.

5.3 Baselines

BERT (Devlin et al., 2019). As the text-only baseline, we use BERT for sequence classification. Here, only the text transcription is used for misogyny detection. BERT has achieved SOTA performances on most NLP benchmark tasks. Therefore, we consider it as a robust textual baseline model for this task.

ResNet-50 (He et al., 2016). As the image-only baseline, we use ResNet-50. Here, only the image is used for misogyny detection. ResNet-50 has achieved SOTA performances on ImageNet, a benchmark task in CV. Therefore, we consider it as a robust image baseline model for this task.

CNN-Based VAE. We build an image-only VAE model whose encoder and decoder are both composed of 5 CNN layers. Here, only the meme image is used for misogyny detection. This image-only input model is used to compare with our MMVAE: both of them are constructed based on VAE.

Note, we only experiment with BERT and ResNet-50 with sub-task A because we directly use the pre-trained model's architecture without incorporating multi-task learning into them.

6 Results and Analysis

In this section, we present our model's performances on both tasks, and give further analysis on its strengths and weaknesses.

6.1 Evaluation Metrics

The main evaluation metric for both tasks is F_1 score (macro-F_1 for sub-task B), which calculates the average of F_1 for both labels:

$$F_1 = \frac{pos\_F_1 + neg\_F_1}{2} \quad (5)$$

We also extend it to include *precision* and *recall* by calculating the average score similarly.

6.2 Results

Table 2 summarizes the performances of our model and the baselines. The first group of models are unimodal. The second group of models share the same architecture, MMVAE, but with different embedding methods. Furthermore, MMVAEs used in the third group come from the best performed model of the second group, *i.e.* $\text{MMVAE}_{\text{LASER+CLIP}}$. In the third group, we have tried a number of techniques to mitigate the overfitting problem: adding different dropout rates, concatenating the word embeddings generated by LASER and LaBSE, adding one more liner layer to the text VAE encoder, and introducing image transforms.

For sub-task A, the optimal F_1 performance is 0.723, which is achieved by applying batch normalization layers and set dropout = 0.2 after each linear layer in the encoders and decoders of the VAE. Among all the tweaks we apply to reduce overfitting, introducing dropout is the most effective by which we get our top two F_1 scores (0.723 and 0.714). Instead of dropping more parameters, keeping 80% of them produces a better result (0.723).This might because the number of parameters in our model is not large, so excluding more will harm the learning capability. We also see a performance improvement to 0.712 by concatenating text embeddings from LASER and LaBSE, which introduces more information, thereby reducing overfitting. The other two attempts fail to improve the performance, the reason might be that adding more layers doesn't make the model more generalizable and image transforms is not effective when applying pre-trained models to embed.

Model	Sub-task A			Sub-task B		
	Precision	Recall	F1	Precision	Recall	macro-F1
BERT	0.608	0.632	0.589	-	-	-
ResNet-50	0.635	0.656	0.622	-	-	-
CNN-VAE	0.526	0.550	0.462	0.514	0.545	0.469
$\text{MMVAE}_{\text{BERT+ResNet}}$	0.640	0.653	0.632	0.543	0.590	0.532
$\text{MMVAE}_{\text{BERT+CLIP}}$	0.707	0.752	0.693	0.586	0.633	0.589
$\text{MMVAE}_{\text{LASER+CLIP}}$	0.721	0.756	**0.711**	0.594	0.648	0.600
$\text{MMVAE}_{\text{LaBSE+CLIP}}$	0.707	0.751	0.694	0.578	0.686	0.575
$\text{MMVAE}_{\text{CLIP+CLIP}}$	0.712	0.760	0.698	0.587	0.658	0.592
$\text{MMVAE}_{\text{+dropout=0.5}}$	0.724	0.759	0.714	0.606	0.656	0.616
$\text{MMVAE}_{\text{+dropout=0.2}}$	0.730	0.756	**0.723**	0.613	0.647	0.622
$\text{MMVAE}_{\text{+concat}}$	0.721	0.751	0.712	0.602	0.657	0.609
$\text{MMVAE}_{\text{+more layers}}$	0.710	0.750	0.698	0.631	0.649	**0.634**
$\text{MMVAE}_{\text{+img transform}}$	0.710	0.756	0.696	0.605	0.651	0.615

Table 2: Performance of our MMVAE model and variants on the test set.

For sub-task B, the optimal performance in F_1 is 0.634, which is achieved by applying batch normalization layers after each linear layer in the encoders and decoders, and adding one more linear layer to the text encoder in the VAE.

Our MMVAE's best performance on sub-task A is significantly higher than that of the uni-modal baselines: 22.8% higher than the text-only BERT baseline, and 16.2% higher than the image-only ResNet-50 baseline. This confirms that our proposed model has effectively learned from both textual and visual features.

Our constructed image-only CNN-based VAE produces the least F_1 score, probably because CNN layers are less effective in capturing complex semantic information contained in meme images.

6.3 Text-Based Error Analysis

We observe that on our randomly selected validation set (10% of memes from the training set), MMVAE obtains nearly 0.87 as F_1 score on sub-task A. But on the test set, we see a 16.7% performance drop. Thus, there is a large gap between our model's performance on the validation set and the test set. We speculate that multiple factors could cause the misclassifications, for instance, similarities between images that are associated with different text. For simplicity, here we only investigate the impact of hateful text on the classification results. We delay further investigations to future work.

To start, we compute the confusion matrix of our best model: $\text{MMVAE}_{\text{+dropout=0.2}}$, which is displayed in Figure 4. There are 500 misogynous memes and 500 non-misogynous memes in the test set. As a result, the number of false negative (56) is much lower compared to the false positives (214). Our model correctly classifies 88.8% of misogynous memes yet only 57.2% of non-misogynous memes.

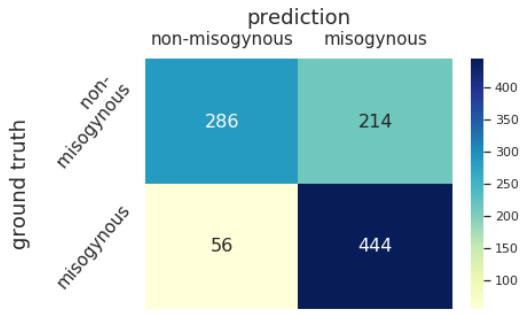


Figure 4: Confusion matrix of $\text{MMVAE}_{\text{+dropout=0.2}}$'s prediction.

We conjecture that issues with performance may be driven by the nature of hateful text. To analyze this, we calculate the hateful scores on the memes' text using the sentiment analysis toolkit in (Pérez et al., 2021). Figure 5 presents boxplots showing the scores. We observe that for true positives (TP) and false negatives (FN), the hateful score distributions are similar, although the former's 2nd quartile is much higher. In contrast, we find that true negatives (TN) contain less hateful text in memes compared to false positives (FP), and the mean score value is significantly different. Apparently, non-misogynous memes tend to have lower hateful scores for their text. Based on our analysis, we

infer that our model is more confident in assigning the correct non-misogynous label to a meme with less hateful text content. Yet, when assigning misogynous labels (although FNs have comparably lower hateful scores), our model is less accurate. As such, we cannot tell the decision boundary between FN and TP by simply looking at the hateful messages contained in meme's text.

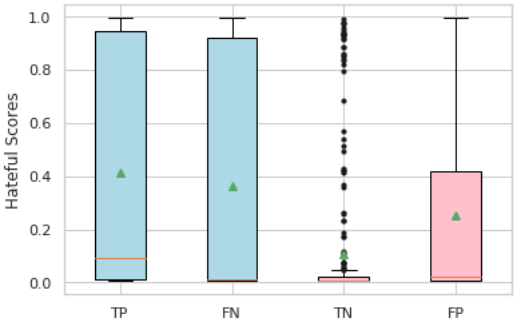


Figure 5: Boxplots of hateful scores on true positives (TP), false negatives (FN), true negatives (TN) and false positives (FP). Green triangle indicates the mean value of the given class. Blue and pink respectively refer to misogynous and non-misogynous memes in ground-truth.

Manual inspection on FPs suggests that the misclassification is potentially driven by several different types of meme. The misogynous text is crossed out in the left meme of Figure 6, but it's still included in the text transcription, which can be confusing for the classifier. Similarly, although the other meme in Figure 6 is not misogynous, it is related to women. We suspect that the presence of certain words associated with femininity e.g. woman, girl, is a another determinate in the prediction. We will test this hypothesis in our future work.

(a) Hateful score: 0.017

(b) Hateful score: 0.086

Figure 6: Two examples of FPs, *i.e.* the ground-truth labels are non-misogynous while the predicted labels are misogynous.

7 Related Work

The misogynous meme detection task is novel, but has similarities to other more general hateful meme detection tasks. Related datasets has been released on a number of shared hateful meme detection/classification tasks (Kiela et al., 2020; Mostafazadeh Davani et al., 2021). Most of the prize-winning models adopted visual-and-linguistic pre-trained models. Velioglu and Rose (2020) utilized pre-trained VisualBERT (Li et al., 2019b) to encode the meme image regions and caption all together. Sharif et al. (2021) and Zia et al. (2021) leveraged visual and textual pre-trained models to encode the meme image and the embedded text respectively, and then learned multi-modal co-representation through vector concatenation. Instead of concatenating the vectors, Pramanick et al. (2021) applied self-attention to learn intra-modality semantic alignments. Lippe et al. (2021) used an ensemble of existing multi-modal pre-trained models based on UNITER (Chen et al., 2020). Zhu (2020) showed that directly applying state-of-the-art multi-modal models on hateful meme classification won't get the optimal performance. They used various data pre-processing approaches to get sufficient features, *e.g.* entity tags, as additional inputs to the pre-trained models. Note, there have also been studies of misogyny in uni-modal platforms (Guest et al., 2021; Zeinert et al., 2021; Jiang et al., 2022).

Our work differs from the above. Many of these previous works directly leverage uni-modal embeddings produced by pre-trained models. They then build a relatively simple model afterwards, of which the learnt multi-modal features are not integrated. In contrast, we strive to overcome this limitation via the co-representation of both textual and image modalities. Moreover, we differ in that we are focusing on misogynous meme detection, rather than the broader topic of hateful meme detection.

8 Conclusion

The spread of hateful memes targeting certain groups has become an important problem on social media platforms. To mitigate the negative consequences brought by misogynous memes, we propose a Multi-modal Multi-task Variational AutoEncoder (VAE) to identify them and assign them more fine-grained labels. Our model consists of three main components: the Image/Text Embed-

ding Module, the Variational AutoEncoder Module, and the Multi-Task Learning Module. Our model's performance outperforms the state-of-the-art unimodal baselines by 22.8% and 16.2%. It effectively learns the co-representation of visual and textual features, and is jointly trained on multiple downstream classification tasks. In our future work, we plan to integrate attention mechanism into our model and carry out more comprehensive statistical analysis on model's results.

References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Meng Cao, Long Chen, Mike Zheng Shou, Can Zhang, and Yuexian Zou. 2021. On pursuit of designing multi-modal transformer for video grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9810–9823.

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

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Mingxi Cheng, Shahin Nazarian, and Paul Bogdan. 2020. Vroc: Variational autoencoder-aided multi-task rumor classifier based on text. In *Proceedings of The Web Conference 2020*, pages 2892–2898.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 Task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Stella Frank, Emanuele Bugliarello, and Desmond Elliott. 2021. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9847–9857.

Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. An expert annotated dataset for the detection of online misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350.

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

Jack Hessel and Lillian Lee. 2020. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 861–877.

Aiqi Jiang, Xiaohan Yang, Yang Liu, and Arkaitz Zubiaga. 2022. Swsr: A chinese dataset and lexicon for online sexism detection. *Online Social Networks and Media*, 27:100182.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference*, pages 2915–2921.

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. *Advances in Neural Information Processing Systems*, 33:2611–2624.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Giwoong Lee, Eunho Yang, and Sung Hwang. 2016. Asymmetric multi-task learning based on task relatedness and loss. In *International conference on machine learning*, pages 230–238. PMLR.

Fei-Fei Li, Justin Johnson, and Serena Yeung. 2019a. Lecture notes of cs231n.

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. *arXiv preprint arXiv:1908.03557*.

Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, and Helen Yannakoudakis. 2021. A multimodal framework for the detection of hateful memes. *PMLR*, 133:344–360.

Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. 2021. Visual news: Benchmark and challenges in news image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6761–6771.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Sijie Mai, Haifeng Hu, and Songlong Xing. 2019. Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 481–492.

Navonil Majumder, Soujanya Poria, Haiyun Peng, Niyati Chhaya, Erik Cambria, and Alexander Gelbukh. 2019. Sentiment and sarcasm classification with multitask learning. *IEEE Intelligent Systems*, 34(3):38–43.

Shervin Malmasi and Marcos Zampieri. 2017. Detecting hate speech in social media. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 467–472.

Aida Mostafazadeh Davani, Douwe Kiela, Mathias Lambert, Bertie Vidgen, Vinodkumar Prabhakaran, and Zeerak Waseem, editors. 2021. *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*.

Jan Niehues and Eunah Cho. 2017. Exploiting linguistic resources for neural machine translation using multi-task learning. In *Proceedings of the Second Conference on Machine Translation*, pages 80–89.

Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. 2016. Convolutional mkl based multimodal emotion recognition and sentiment analysis. In *2016 IEEE 16th international conference on data mining (ICDM)*, pages 439–448. IEEE.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. MOMENTA: A multimodal framework for detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4439–4455.

Juan Manuel Pérez, Juan Carlos Giudici, and Franco Luque. 2021. pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks.

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 *Proceedings of the 38th International conference on machine learning*, volume 139. PMLR.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Benet Oriol Sabat, Cristian Canton Ferrer, and Xavier Giro-i Nieto. 2019. Hate speech in pixels: Detection of offensive memes towards automatic moderation. In *AI for Social Good workshop at NeurIPS*.

Omar Sharif, Eftekhar Hossain, and Mohammed Moshiul Hoque. 2021. NLP-CUET@DravidianLangTech-EACL2021: Offensive language detection from multilingual code-mixed text using transformers. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 255–261.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Bjorn Gamback. 2020. Semeval-2020 task 8: Memotion analysis–the visuolingual metaphor! In *Proceedings of the 14th International Workshop on Semantic Evaluation*, pages 759–773.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015*.

ES Smitha, Selvaraju Sendhilkumar, and GS Mahalaksmi. 2018. Meme classification using textual and visual features. In *Computational Vision and Bio Inspired Computing*, pages 1015–1031. Springer.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

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.

Riza Velioglu and Jewgeni Rose. 2020. Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge. In *The Hateful Memes Challenge at NeurIPS*.

Suryatej Reddy Vyalla and Vishaal Udandarao. 2020. Memeify: A large-scale meme generation system. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, pages 307–311.

Lianwei Wu, Yuan Rao, Haolin Jin, Ambreen Nazir, and Ling Sun. 2019. Different absorption from the same sharing: Sifted multi-task learning for fake news detection. *CoRR*, abs/1909.01720.

Sen Wu, Hongyang R. Zhang, and Christopher Ré. 2020. Understanding and improving information transfer in multi-task learning. In *International Conference on Learning Representations*.

Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. 2021. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2560–2569.

Tiezheng Yu, Wenliang Dai, Zihan Liu, and Pascale Fung. 2021. Vision guided generative pre-trained language models for multimodal abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3995–4007.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114.

Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018. On the origins of memes by means of fringe web communities. In *Proceedings of the Internet Measurement Conference 2018*, pages 188–202.

Philine Zeinert, Nanna Inie, and Leon Derczynski. 2021. Annotating online misogyny. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3181–3197.

Ying Zeng, Sijie Mai, and Haifeng Hu. 2021. Which is making the contribution: Modulating unimodal and cross-modal dynamics for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1262–1274.

Ron Zhu. 2020. Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution. *arXiv preprint arXiv:2012.08290*.

Haris Bin Zia, Ignacio Castro, and Gareth Tyson. 2021. Racist or sexist meme? classifying memes beyond hateful. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 215–219.