R2D2 at SemEval-2022 Task 5: Attention is only as good as its Values! A multimodal system for identifying misogynist memes

Mayukh Sharma, Ilanthenral Kandasamy and **W.B. Vasantha**
School of Computer Science and Engineering
Vellore Institute of Technology
Vellore - 632014, Tamil Nadu, India
04mayukh@gmail.com, ilanthenral.k@vit.ac.in,
vasantha.wb@vit.ac.in

Abstract

This paper describes the multimodal deep learning system proposed for SemEval 2022 Task 5: MAMI - Multimedia Automatic Misogyny Identification. We participated in both Subtasks, i.e. Subtask A: Misogynous meme identification, and Subtask B: Identifying type of misogyny among potential overlapping categories (stereotype, shaming, objectification, violence). The proposed architecture uses pre-trained models as feature extractors for text and images. We use these features to learn multimodal representation using methods like concatenation and scaled dot product attention. Classification layers are used on fused features as per the subtask definition. We also performed experiments using unimodal models for setting up comparative baselines. Our best performing system achieved an F1 score of 0.757 and was ranked 3^{rd} in Subtask A. On Subtask B, our system performed well with an F1 score of 0.690 and was ranked 10^{th} on the leaderboard. We further show extensive experiments using combinations of different pre-trained models which will be helpful as baselines for future work.

1 Introduction

Internet and social media sites have played an integral role in bringing people together by providing a simple yet effective way of communication. Over recent times, internet memes have become a popular choice for sharing sentiment on the internet. A meme is an approach, concept, idea, or style that spreads through social media within a society, often to express a trend, topic, or significance represented by it (Peirson and Tolunay, 2018). Memes shared on the internet are often harmless and used to express humour; however, recent trends have increased their usage to spread hate or cause social unrest (Lippe et al., 2020). Hate speech and, in particular, hate against women has seen an exponential rise in social media platforms (Pamungkas et al., 2020). Misogyny, a subset of hate-speech (Safi Samghabadi et al., 2020), is defined as hate or prejudice against women, which can be manifested in numerous ways, including social exclusion, sex discrimination, hostility, patriarchy, male privilege, belittling of women, disenfranchisement of women, violence against women, and sexual objectification(Anzovino et al., 2018). Women have a strong presence online, particularly on Instagram and Twitter. Women use social media multiple times a day compared to men(Fersini et al., 2020). This makes it extremely important to identify and remove such content to make the internet safer for women.

Efforts have been made to identify misogynous textual content on social media (Anzovino et al., 2018) (Fersini et al., 2018b) (Pamungkas et al., 2020) (Hewitt et al., 2016), however, no efforts have been made to identify the misogynous content spanning multiple modalities like memes. Memes are uniquely multimodal and convey information using images and text. The multimodal nature of memes allows them to combine harmless texts/images into misogynist memes when used together. This poses an exciting challenge as memes require joint language and visual understanding to infer their true meaning. SemEval 2022 Task 5: MAMI (Fersini et al., 2022) draws attention to the problem of identifying misogynous memes and further identifying the type of misogyny. The task provides a dataset of misogynist memes and its type. Both images and the corresponding text was available as a source of information; the text content of the provided dataset was in English.

Our proposed system for both subtasks uses the late fusion of visual and textual features obtained from pre-trained models. We use separate pre-trained models for each modality, i.e. text and image, and fuse the features to learn multimodal representation for memes. We experimented with simple concatenation of image and

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 761 - 770

text features; and scaled dot product attention (Vaswani et al., 2017), followed by a convolution layer to learn multimodal features. Once multimodal features are learnt, we stack classification layers on them as per the subtask requirements. We also performed extensive experiments using a single modality (text/image) to set up comparative baselines. We used ViT (Vision Transformer) for image feature extraction. We experimented with various PLMs for textual feature extraction like Bidirectional Encoder Representations from Transformers(BERT), Robustly Optimized BERT Pretraining Approach(RoBERTa), MPNet, and Decoding-enhanced BERT with disentangled attention(DeBERTa).

The results of unimodal experiments showed that text-only models had a superior performance than image-only models. We experimented with feature concatenation and scaler dot product attention for multimodal models. Feature concatenation led to only minor performance gains. In case of scaler dot product attention, choice of query(Q) (Vaswani et al., 2017) turned out to be an essential factor. Our experiments showed that image features as query performed significantly well for all models and outperformed all unimodal and concatenation baselines. Our best performing model was a voting ensemble of attention based multimodal models and achieved an F1 score of 0.757 with a 3^{rd} rank on the official leaderboard for Subtask A. For Subtask B we used BERT and ViT based attention model, which performed well, attaining an F1 score of 0.690 and 10^{th} rank on the leaderboard. Our code available at GitHub[1] for method replicability.

2 Background

Identifying misogynous content is critical to making the internet accessible and safe for women. In recent times there has been an exponential rise in hateful content and, in particular, the phenomenon of hate against women on social media (Pamungkas et al., 2020) (Hewitt et al., 2016). There have been previous attempts to identify hate/toxic content on social media platforms ((Zampieri et al., 2020) (Zampieri et al., 2019) (Sharma et al., 2021a) (Pavlopoulos et al., 2021) but none deal specifically with identifying the hate against women. The first benchmark dataset to identify misogynous content was proposed in (Anzovino et al., 2018). The task and the papers of AMI@Evalita 2018 (Fersini et al., 2018a) and AMI@IberEval2018 (Fersini et al., 2018b) highlight the difficulties and barriers involved in automatically identifying misogynist content on social media. Workshop on Trolling, Aggression and Cyberbullying (TRAC) shared task (Kumar et al., 2020) contained a subtask to identify gender-based identification of hateful content.

Type	Misogynous	Not misogynous	Total
Train	4742	4758	9500
Validation	258	242	500
Test	500	500	1000

Table 1: Dataset Statistics for Subtask A

There has been a rise in multimodal content over the internet in the form of memes. Most efforts to identify toxic and misogynous content consider only the textual content. Due to the rapid increase of multimodal content, efforts have been made to analyse it. Memotion analysis (Sharma et al., 2020a) aimed to perform sentiment analysis on internet memes. It involved identifying offensive sentiment as part of its Subtasks. (Sharma et al., 2020b) used a feature fusion model using LSTMs, GRUs with attention and attained the best results in identifying offensive memes. Hateful memes challenge (Kiela et al., 2020) aimed to study and identify the hateful nature of internet memes. Analysing memes is an intrinsically difficult task as it requires multimodal reasoning capable of understanding textual and visual features. The common strategy used in analysing memes involved learning features for each modality and then fusing the features to represent joint features. Work done in (Pranesh and Shekhar, 2020) uses the simple concatenation of text and visual features for meme sentiment analysis. (Lippe et al., 2020) used early fusion models like LXMERT (Tan and Bansal, 2019), UNITER (Chen et al., 2020) uses Transformer (Vaswani et al., 2017) architecture and neural attention to learn joint representation.

Social media provides a platform for numerous people to express and share their thoughts. Misogyny which can be simplified as hate or prejudice against women, is on the rise in social media platforms. Hence, it is crucial to identify and remove such content from the internet. Attempts have been made to identify hateful/toxic and misogynist content on the internet, but none focuses on multimodal

[1] https://github.com/04mayukh/R2D2-at-SemEval-2022-Task-5-MAMI

Type	Shame	Stereotype	Objectification	Violence	Total
Train	1271	2810	2201	953	5000
Validation	60	141	105	47	250
Test	146	350	348	153	1000

Table 2: Dataset Statistics for Subtask B

content like memes. SemEval 2022 Task 5: MAMI - Multimedia Automatic Misogyny Identification (Fersini et al., 2022) aims to study the misogynist nature of memes and is divided into two subtasks which we define as:

Subtask A: Given a labelled dataset D of internet memes and their text, the objective of the task is to learn a classification function that can predict if a meme is misogynous or not.

Subtask B: Given a labelled dataset D of internet memes and their text, the objective of the task is to learn a multilabel classification function that can predict the type of misogyny M for a given misogynous meme, where $M \in \{$stereotype, shaming, objectification and violence$\}$.

Dataset Statistics: The dataset for the task consisted of internet memes and their textual content. For Subtask A, the memes were labelled as misogynous/non-misogynous. Misogynous memes from Subtask A were further labelled into the type of misogyny. Subtask B involved recognising the type of misogyny from overlapping categories like stereotype, shaming, objectification, and violence. We also split the provided dataset into train and validation sets for training and evaluating the models before using them to make predictions on the test set. Table 1 and Table 2 show the dataset statistics for Subtask A and Subtask B.

3 System Overview

3.1 *Pre-trained Models:*

Finetuning (Qiu et al., 2020) pre-trained language models has become a popular approach in the deep learning community (Sharma et al., 2021b). It is a form of transfer learning that utilizes models trained on enormous amounts of unannotated text data to learn general-purpose representations. These models are then finetuned on downstream tasks. In recent times there has been a rapid rise in pre-trained models in Natural Language Processing (NLP). The knowledge from these models can be easily transferred to tasks where small amounts of data are present, making them extremely useful. The maximum of these PLMs like BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), MPNet (Song et al., 2020), DeBERTa (He et al., 2021) are based on transformers. Pre-trained models used in computer vision mostly rely on convolutional networks. Architectures like classic ResNet (He et al., 2016) have attained state of the art performance in large scale image recognition tasks (Kolesnikov et al., 2020) (Xie et al., 2020). Recently the transformer architecture has also been used in computer vision and performs at par with convolutional networks (Touvron et al., 2021) (Dosovitskiy et al., 2021) (Bao et al., 2022). The use of transformers in computer vision has also led to multimodal pre-trained models (Kim et al., 2021) (Li et al., 2019) (Tan and Bansal, 2019). Pre-trained models provide an efficient and scalable way to use large-scale learning to simple downstream tasks efficiently. Next, we describe the pre-trained models used in our multimodal system.

3.2 *Brief overview of Pre-trained models:*

Vision Transformer (ViT): ViT (Dosovitskiy et al., 2021) was proposed by Google and aimed to use the transformer architecture with minimal changes to computer vision tasks. Transformers use sequences to process data and cannot process grid-structured data. The images in the transformer were converted into smaller image patches and used as sequences. Trainable positional encodings were used to retain positional information of smaller image patches. These positional encodings help to learn the relationship between smaller image patches. Finally, the whole model is pre-trained as a classification task on the ImageNet dataset (Russakovsky et al., 2015).

BERT: It stands for Bidirectional Encoder Representations from Transformers (Devlin et al., 2019) is a PLM based on the transformer architecture. It is pre-trained on text corpus using Masked Language Modelling (MLM) and Next Sentence Prediction (NSP) objective.

RoBERTa: A Robustly Optimized BERT Pretraining Approach (Liu et al., 2019) was developed by Facebook. They used the BERT architecture

with few modifications and obtained better performance. They used dynamic masking in their pretraining and removed the NSP objective. They also trained the model using a larger batch size with more data for longer durations.

MPNet: MPNet (Song et al., 2020) was proposed by Microsoft. Most language models are pretrained using either MLM or permuted language modelling objectives. MPNet makes the best of both permuted language modelling and MLM. It proposed a unified view of MLM and permuted language modelling by splitting and rearranging the tokens into predicted and non-predicted parts. It uses MLM to see the positional information of complete sentences and permuted language modelling to model dependency among predicted tokens.

DeBERTa: DeBERTa stands for decoding-enhanced BERT with disentangled attention (He et al., 2021), was proposed by Microsoft and outperformed human performance on the SuperGlue benchmark (Wang et al., 2019). It used disentangled attention where two vectors are used, one to represent the content and the other to store the positional information. Attention weights are calculated using disentangled matrices on their content and relative positions. It also uses an enhanced mask decoder during pre-training to incorporate absolute positions while predicting masked tokens.

3.3 *Unimodal models (baselines):*

Meme classification is an intrinsically complex problem due to textual and visual cues. Memes can convey a message using image, text, or both, thus requiring textual, visual, and multimodal understanding. We first modelled misogyny detection using purely unimodal approaches to form comparative baselines as part of our experiments. BERT, RoBERTa, MPNet, DeBERTa were used for meme text classification, and ViT was used for meme image classification. The unimodal baselines were also helpful in understanding the predictive power of text vs image features. We experimented with unimodal models only for Subtask A.

3.4 *Multimodal models:*

The multimodal nature of memes makes it extremely difficult to understand their true meaning. They may contain a combination of completely different visual and textual content, which, when joined, turn out to be misogynist in nature. To understand this multimodal nature of memes, we used two different techniques to join visual and textual representations to learn multimodal features, which we will discuss next.

Feature concatenation: To learn jointly from image and text features, we used late fusion to concatenate the features learnt by image and text pretrained models. Concatenated features are then fed to the final classification layer to label the meme as misogynistic (Subtask A) or identify the type of misogyny (Subtask B). The image and text pretrained models are jointly finetuned with the classification layer using the classification objective.

Attention-based feature fusion: Attention has been widely used in various NLP tasks. It forms the critical component of transformer architecture. The main idea behind attention is to learn representation for a given feature based on its relative relevance with respect to other features. We use the same idea to learn joint image-text features using attention mechanism. We use scaler dot product attention (Vaswani et al., 2017) which uses concept of queries(Q), keys(K), and values(V) and is defined as:

$$Attention(Q, K, V) = softmax(QK^T/\sqrt{d_k})V$$

where dimension of Q and K is (N, d_k), dimension of V is (N, d_v), and N is sequence length. The output of the attention step has the dimension of (N, d_v), which represent the context vectors. The language and vision models we used both utilised transformer architecture. As we know, transformers work with data sequences, which helped us get highly localised feature representations, allowing us to use attention to fuse visual and text features. The features from pre-trained text models represent the information corresponding to input tokens. The visual features from ViT represent information corresponding to $N \times N$ patches from the input image. We use the dot product attention defined above on these set of features to learn the multimodal representation.

Another essential aspect of scaler dot product attention is the choice of query and key during attention. In transformers, self-attention is calculated, which makes this choice trivial. However, while using attention on visual and text features, it is essential to note that attention is not commutative. Therefore, we performed experiments using both visual and textual features as the query parameter in attention. The context vectors obtained from the attention layer were then passed through a one-dimensional convolution layer to learn final

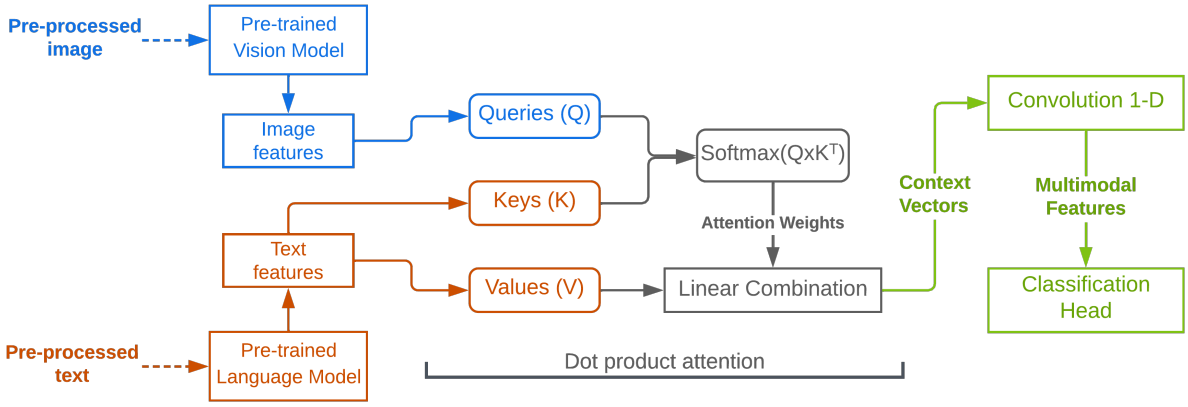

Figure 1: Architecture of our model using attention with image features as query.

multimodal features. These features were finally flattened and fed to the final classification layers. Figure 1 shows the architecture diagram of our attention based multimodal model using image features as query.

3.5 *Classification layers and finetuning:*

We finetuned the unimodal text models by stacking a simple dense and batch normalisation layer followed by a one-neuron classifier on top of features learnt by pre-trained models. We used the features from [CLS] token in the case of BERT and RoBERTa and start the token(<s>) in the case of MPNet and DeBERTa. For ViT we used the [class] token, which is taken as image representation. We pass the joint visual and text features through a batch normalisation and simple dense layer followed by a one-neuron classifier for multimodal models. Visual and text models along with concatenation/attention layer were wrapped into a single model for finetuning multimodal models.

4 Experimental setup

4.1 *Pre-processing images and texts*

Text: We used the ekphrasis (Baziotis et al., 2017) library for text pre-processing. It normalises time, date, numbers to a standard format and corrects misspelt words. Chatwords are commonly used in memes, so we converted them into their full forms. PLMs need text to be tokenised before it can be fed to them. We used Hugging Face's (Wolf et al., 2020) implementation of Fast tokenisers[2] for each pre-trained model.

Image: Images need to be pre-processed before being fed to ViT. We first resized the image to 224 × 224. We also divided the pixel values by 255 to bring them within a range of 0-1. Finally, the images were normalised using the mean and standard deviation of 0.5 across all channels.

4.2 *Task-wise model definition:*

Subtask A: We experimented with unimodal and multimodal techniques. Our unimodal baselines used BERT, RoBERTa, MPNet, DeBERTa for text and ViT for images. We extracted the features from these models and passed them through a dense layer with 32 neurons, followed by a batch normalisation layer. Finally, we used a classification layer with a single neuron and sigmoid activation to label the input as misogynous/not misogynous. For multimodal models, we experimented with concatenation as well as attention using BERT and ViT initially. We passed the fused features through a batch normalisation and dense layer consisting of 64 neurons, followed by a single neuron classification and sigmoid activation. Further, we experimented using a combination of RoBERTa, MPNet, and DeBERTa with ViT using attention mechanism (image features as query). The one-dimensional convolution layer in multimodal models used 32 filters with a kernel size of 30 and stride 15.

Subtask B: We used only the multimodal models using concatenation and attention to learn multimodal features. Subtask B was a multilabel task where a misogynous meme could belong to more than one category: stereotype, shaming, objectification, and violence. We trained a single model to classify the memes into the given categories. We created a multi-branch model where each branch tries to predict if meme belongs to one of the given categories. The branch takes the independent text

[2]Hugging Face's Fast Tokenizers

and visual features fused using simple concatenation or a combination of attention and 1-D Convolution. The fused multimodal features for each branch are then passed through a 32-neuron dense layer, batch normalisation layer and finally through a single neuron classification layer with sigmoid activation.

4.3 *Hyperparameters and training:*

We developed our models using TensorFlow[3], Keras [4] (Chollet et al., 2015) and Hugging Face's [5] implementation of transformer[6] (Wolf et al., 2020) models. The models were trained using GPU/TPU on Google Colab. We fixed the sequence length to 80 tokens across both subtasks for text modality. Sequences greater or shorter than 80 tokens were accordingly truncated or padded. We used their base versions for all PLMs; for ViT we used the base model with the patch size of 16 $\times$ 16 and input image size of 224$\times$ 224. Finetuning was performed using Adam (Kingma and Ba, 2015) optimiser against a binary cross-entropy loss. We experimented with learning rates ranging from 2e-5 to 5e-5 with a batch size of 128 for TPU and 16 for GPU. Finetuning was done for ten epochs, and weights corresponding to the best results on the validation set were used to make predictions on the test set. For Subtask A, we finetuned the entire dataset containing misogynous as well as non-misogynous memes. For Subtask B, we trained the model only on misogynous memes to identify the type of misogyny. For evaluation on the test set, we used a hierarchical approach where we made predictions only on samples predicted as misogynous from the best performing model on Subtask A.

4.4 *Evaluation metric:*

Subtask A used the macro-averaged F1 score to evaluate the model's performance. F1 scores were calculated individually for each class and then averaged to give the Macro F1 score. For Subtask B, the weighted-average F1 measure is used as the evaluation metric. F1 scores are computed for each label, and then the weighted average is computed based on true instances belonging to each label category.

[3] https://www.tensorflow.org/
[4] https://keras.io
[5] https://huggingface.co
[6] https://huggingface.co/transformers

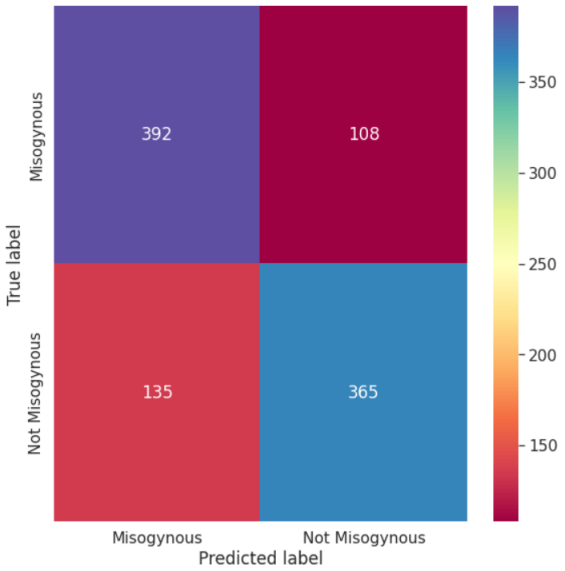

Figure 2: Confusion Matrix for Subtask A (ensemble model)

5 Results and Analysis

Table 3 and Table 4 contain the results of our models for Subtask A and Subtask B. In Subtask A our best performing model was a voting ensemble of attention (image as a query) models trained with different seed values attaining an F1 score of 0.757 and was ranked 3^{rd} on the leaderboard. Figure 2 shows the confusion matrix for our ensemble model. In Subtask B we used a hierarchical model. We made predictions only for samples classified as misogynous by our best performing ensemble model on Subtask A. Our BERT + ViT model using attention (image as a query) performed best with an F1 score of 0.690 and was ranked 10^{th} on the leaderboard.

Table 3 contains a comparative analysis of unimodal vs multimodal techniques we used as part of our experiments for Subtask A. If we compare unimodal models, we can see that text-based models have better performance than the image-based models on both development and test sets. It points to the possibility that text present in memes is a prominent factor in identifying misogynous content. DeBERTa outperforms other models by a considerable margin among text models, performing almost equivalent to the multimodal concatenation model. For multimodal models, we performed initial sets of experiments using BERT and ViT over different methods of fusing the modalities. The simple concatenation worked well and provided a slight improvement when compared to unimodal

Model	Type	Development					Test				
		Precision*	Recall*	Precision	Recall	F1	Precision*	Recall*	Precision	Recall	F1
BERT	TEXT ONLY	.813	.831	.829	.830	.830	.618	.786	.662	.650	.643
RoBERTa		.813	.792	.816	.815	.815	.657	.518	.630	.624	.620
MPNet		.821	.835	.835	.836	.836	.645	.712	.662	.660	.659
DeBERTa		.850	.792	.837	.834	.835	.665	.734	.684	.682	.6811
ViT	IMAGE ONLY	.792	.665	.744	.740	.737	.600	.820	.658	.637	.624
ViT + BERT	Concatenation	.860	.835	.858	.857	.857	.646	.852	.714	.692	.684
ViT + BERT	Attention(QUERY-Text)	.802	.659	.748	.743	.739	.618	.920	.731	.676	.655
ViT + BERT	Attention(QUERY-Image)	.921	.771	.857	.851	.848	.682	.836	.735	.723	.719
ViT + MPNet		.930	.725	.846	.833	.829	.700	.822	.742	.734	.732
ViT + RoBERTa		.928	.698	.836	.820	.814	.700	.822	.742	.735	.733
ViT + DeBERTa		.949	.725	.857	.842	.837	.683	.824	.731	.721	.718
ENSEMBLE	NA						.771	.730	.758	.757	.757

Table 3: Experimental results on development and test set for Subtask A. Metrics for misogynist class are represented using $*$ after the metric name.

Model	Type	F1 Stereotype		F1 Shaming		F1 Objectification		F1 Violence		F1 Weighted Avg.	
		Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test
BERT + ViT	Concatenation	.658	.652	.710	.636	.741	.695	.752	.722	.710	.672
BERT + ViT	Attention (QUERY – Image)	.648	.666	.663	.67	.745	.708	.747	.711	.696	.690

Table 4: Experimental results on development and test set for Subtask B.

techniques.

Our next set of experiments used scaler dot product attention using BERT and ViT. As we can see from the results, the choice of query(Q) played a crucial role in calculating the multimodal features using attention. BERT + ViT model using text features as query performed poorly when compared to concatenation and best performing textual models. However, when we use the same architecture and change the query(Q) term to image features, there is a significant gain that outperforms all other models. The attention mechanism uses queries(Q), keys(K) and values(V) to calculate the context vectors. We can observe from the formula defined in section 3 that while the query(Q) and key(K) is used to calculate attention weights, the final features are a linear combination of attention weights over the values(V). Our experiments using unimodal models showed that textual features perform better than visual features, almost as good as concatenation based multimodal models. Using image features as a query allows us to preserve the textual features which are used as values(V) while modelling the correlation between images and text using query-key matching. This might be one of the possible reasons for the better performance of attention-based models with image features as query. We performed further experiments using only attention-based fusion with image features as queries(Q). We used MPNet, RoBERTa, and DeBERTa with ViT, which outperformed all unimodal and multimodal baselines. For Subtask B, we experimented with only concatenation and attention methods and found that attention (image as a query) performed better than concatenation.

6 Conclusion

This paper describes our approach for SemEval 2022 Task 5: MAMI - Multimedia Automatic Misogyny Identification. We propose a dot product attention-based mechanism for learning multimodal representation from independent text/image features. Our work also describes a comprehensive set of experiments using unimodal/multimodal models using different pre-trained models. Our system performed well, attaining 3^{rd} rank on Subtask A and 10^{th} in Subtask B. Our experiments highlight the non-commutative nature of dot product attention, with the choice of the query being a critical design decision. Our results showed that textual features dominate over image features in multimodal understanding. In the future, we would like to explore more on how attentions learn multimodal features and further compare the role of individual modalities in multimodal tasks.

References

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *Natural Language Processing and Information Systems*, pages 57–64, Cham. Springer International Publishing.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2022. BEit: BERT pre-training of image transform-

ers. In *International Conference on Learning Representations*.

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. DataStories at SemEval-2017 task 4: Deep LSTM with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, Vancouver, Canada. Association for Computational Linguistics.

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 *Computer Vision – ECCV 2020*, pages 104–120, Cham. Springer International Publishing.

François Chollet et al. 2015. Keras. `https://keras.io`.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

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.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018a. Overview of the evalita 2018 task on automatic misogyny identification (AMI). In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*, volume 2263 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2020. Ami @ evalita2020: Automatic misogyny identification. In *EVALITA*.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018b. Overview of the task on automatic misogyny identification at ibereval 2018. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018*, volume 2150 of *CEUR Workshop Proceedings*, pages 214–228. CEUR-WS.org.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. {DEBERTA}: {DECODING}-{enhanced} {bert} {with} {disentangled} {attention}. In *International Conference on Learning Representations*.

Sarah Hewitt, T. Tiropanis, and C. Bokhove. 2016. The problem of identifying misogynist language on twitter (and other online social spaces). In *Proceedings of the 8th ACM Conference on Web Science*, WebSci '16, page 333–335, New York, NY, USA. Association for Computing Machinery.

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. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. 2020. Big transfer (bit): General visual representation learning. In *Computer Vision – ECCV 2020*, pages 491–507, Cham. Springer International Publishing.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2020. Evaluating aggression identification in social media. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 1–5, Marseille, France. European Language Resources Association (ELRA).

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. Misogyny detection in twitter: a multilingual and cross-domain study. *Information Processing & Management*, 57(6):102360.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. SemEval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 59–69, Online. Association for Computational Linguistics.

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

Raj Ratn Pranesh and Ambesh Shekhar. 2020. Memesem:a multi-modal framework for sentimental analysis of meme via transfer learning.

XiPeng Qiu, TianXiang Sun, YiGe Xu, YunFan Shao, Ning Dai, and XuanJing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.

Niloofar Safi Samghabadi, Parth Patwa, Srinivas PYKL, Prerana Mukherjee, Amitava Das, and Thamar Solorio. 2020. Aggression and misogyny detection using BERT: A multi-task approach. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 126–131, Marseille, France. European Language Resources Association (ELRA).

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020a. SemEval-2020 task 8: Memotion analysis- the visuolingual metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 759–773, Barcelona (online). International Committee for Computational Linguistics.

Mayukh Sharma, Ilanthenral Kandasamy, and Vasantha Kandasamy. 2021a. Deep learning for predicting neutralities in offensive language identification dataset. *Expert Systems with Applications*, 185:115458.

Mayukh Sharma, Ilanthenral Kandasamy, and W.b. Vasantha. 2020b. Memebusters at SemEval-2020 task 8: Feature fusion model for sentiment analysis on memes using transfer learning. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1163–1171, Barcelona (online). International Committee for Computational Linguistics.

Mayukh Sharma, Ilanthenral Kandasamy, and W.b. Vasantha. 2021b. YoungSheldon at SemEval-2021 task 7: Fine-tuning is all you need. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1146–1152, Online. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. In *NeurIPS 2020*. ACM.

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.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. 2021. Training data-efficient image transformers and distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10347–10357. PMLR.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. 2020. Self-training with noisy student

improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.