Amazon PARS at Memotion 2.0 2022: Multi-modal Multi-task Learning for Memotion 2.0 Challenge

Gwang Gook Lee and Mingwei Shen

Amazon.com, 410 Terry Ave N., Seattle, 98109, United States

Abstract

Over the years, memes became very popular as social media services growing rapidly. Understanding meme images as humans do is very complicated because of its multi-modal nature (texts on images). In this paper, we describe our approach for classifying sentiment and emotion of memes for Memotion 2.0 challenge. Assuming correlation between three sub-tasks, we implemented and compared four different multi-task network heads having different level of interactions. Experiments showed that multi-task classification network could perform better than individual networks for single tasks. We won 6th, 4th and 1st place for task A, B and C respectively.

Keywords

Emotion classification, multi-modal, multi-task

1. Introduction

Memes are images (usually have short texts on them) used to deliver ideas and jokes. They are transmitted and replicated rapidly through social media. With the growth of social media, memes became a popular culture online over the decade. However, understanding the contents of memes is not an easy task because of its multimodal nature (text+image). Also, oftentimes it requires deeper understanding in culture (for example, many memes are produced based on movies or TV shows).

This paper presents our solution for Memotion 2.0 challenge [1][2]. The challenge consists of three sub-tasks to classify (a) sentiment, (b) emotion and (c) emotion intensity of memes. We used VisualBERT [3] to process text and image modalities of memes. Considering correlations among three sub-tasks, four multi-task classification heads are tested on top of VisualBERT architecture. Experiments showed that multi-task networks could perform better than single-task networks dedicated for each task.

2. Challenge Dataset

Memotion 2.0 challenge data consist of train and validation sets which have 7,000 and 1,500 samples respectively. Each sample includes a meme image (have texts on them) as well as recognized meme texts.

The challenge consists of three sub-tasks:

- Task A: sentiment analysis (positive, neutral or negative)
- Task B: emotion classification (humorous, sarcastic, offensive and motivational). A meme can have more than one label.
- Task C: intensity of emotion classes quantifying their extent. Intensities vary upon emotions as in Table 1.

Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, Feb 28, Virtual Location
EMAIL: gglee@amazon.com (G. G. Lee); mingweis@amazon.com (M. Shen)
ORCID: 0000-0002-8776-6871 (G. G. Lee)


Table 1
Intensities for emotion classes

Emotion	Intensity
Humorous	not_funny, funny, very_funny, hilarious
Sarcastic	not_sarcastic, little_sarcastic, very_sarcastic, extremely_sarcastic
Offensive	not_offensive, slight, very_offensive, hateful_offensive
Motivational	not_motivational, motivational

Figure 1: An example of Memotion 2.0 data. This meme has *neutral* sentiment and *funny*/ *little*_sarcastic emotions.

3. Proposed Method

We chose VisusalBERT as our baseline because it showed the best performance among Hateful Memes Challenge [4] baselines which is in a similar domain with Memotion 2.0 Challenge. ViLBERT [5] also showed comparable performance to VisualBERT at the same challenge. However, compared to ViLBERT, VisualBERT employs only single transformer for both text and image modalities hence requires smaller memory and training time.

For text features, *bert-base-uncased toke*nizer is used and maximum sequence length is fixed to 128. For VisualBERT, images features can be extracted from grid maps or regions from object detection results. It has been shown that grid-based features can perform on par with region-based features [6][7] while keeping the entire pipeline simpler (enabling end-to-end pretraining). However, we chose region-based image features for two reasons. First, most meme images have large empty background regions as Fig. 1. For such images, region-based features, which are prepared only for detected objects, would help the model to focus more on image contents compared to grid-based features where all image regions are treated equally. For the next, the challenge dataset has small amount of data for training (7,000 samples). Object detectors are trained on large datasets and frozen while extracting features which makes it more reliable to small size data compared to grid configurations.

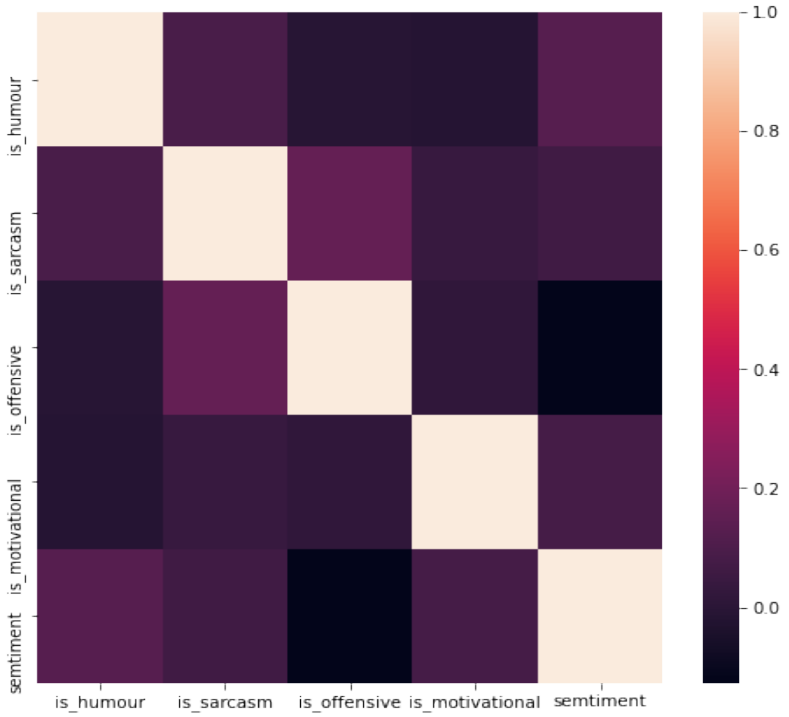
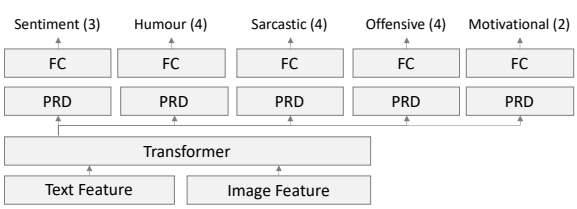


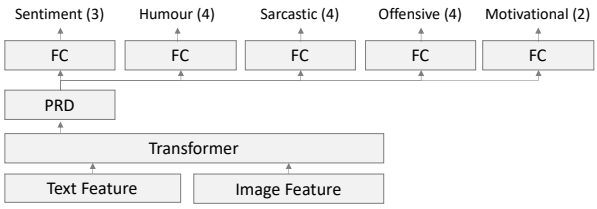
Figure 2: Correlation among labels: sentiment (task A) has positive correlation with *is_humour* (task B) but negative relation with *is_offensive* (task B)

We chose a multi-task model rather than dedicated models for different tasks hypothesizing sub-tasks are related. For example, memes with funny emotion would have higher probability of having positive sentiment compared to offensive memes. Oftentimes, funny memes are sarcastic at the same time. Fig. 2 shows correlation among labels in task A and task B. We also could easily expect Task B and C are highly correlated as Task C is actually a fine-grained version of Task B.

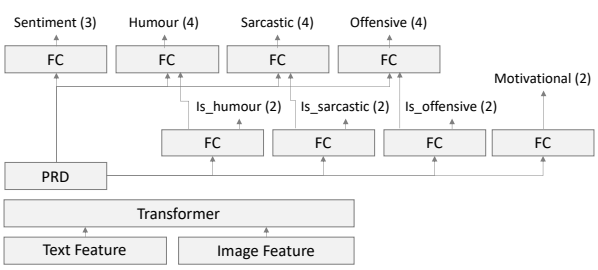
We designed four multi-task classification heads as illustrated in Fig 3. PRD and FC stand for BERT prediction head and fully connected layer respectively. Numbers in parenthesis describes number of channels in each output. For example, heads for sentiment classification have three out-puts: negative, neutral and positive.



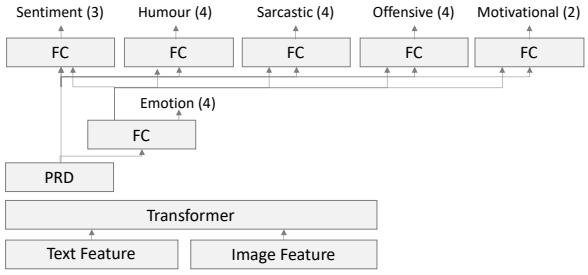
(a) Multi-task classification head 1



(b) Multi-task classification head 2



(c) Multi-task classification head 3



(d) Multi-task classification head 4

Figure 3: Multi-task classification heads

Classification head 1 has separate PRD for each output to learn task specific predictions. In contrast, head 2 shares PRD for all tasks to strengthen benefits of multi-task learning. Head 3 classifies emotions from Task B as binary classes first (*is_humour*, *is_sarcastic* and *is_offensive*). The outputs of emotion predictions (Task B) are then con-catenated with multimodal feature embeddings (from PRD) to make predictions on emotion intensities (Task C) expecting existence of emotions would help to classify how strong the emotions are. This is analogous to the two-stage architectures in object detection [8]. The first stage (region proposal networks, RPN) only produces output for the existence of an object and the following lay-er classifies in which category the object belongs. Head 4 is similar with head 3 but a multi-label classifier is utilized rather than four binary classifiers in head 3.

Multi-task loss is defined as weighted sum of losses for each task. Cross entropy loss is used for multi-class task (Task A and C) and binary cross entropy with logits is utilized for multi-label task (Task B). Weights for Task A, B and C are chosen as 0.4, 0.3 and 0.3. Predictions for Task B is generated from emotion intensity output when they are not explicitly predicted. For example, *slightly_funny*, *funny* or *hilarious* predictions are all considered as *humourous*.

4. Experiments

4.1 Pre-trained Weights

We used MMF [10] a vision and language multimodal research from Facebook AI Research for implementation. MMF provides pretrained weights for various models for different types of tasks. It is well known that the similarity between source domain and target do-main for fine-tuning affects to the model performance. Also, it has been shown that when target domain is very different from source domain, it gives better performance to train the model directly on the target domain rather than fine-tuning from a pre-trained model [8].

We tested several different pretrained models as summarized in Table 2. COCO, VQA2 and Conceptual Captions (CC) are the most common datasets for pre-training multi-modal models. We also trained the model directly on the train data without pre-training. For the last, two models finetuned on hateful memes dataset with pretraining (on COCO) and without pretraining are tested because of the resemblance of dataset. To compare pre-trained weights, two classification networks for Task A and Task C are trained and averages of their F1 are measured on the validation set. Surprisingly, direct training gave a good performance, even better than some pre-trained weights. However, fine-tuning on the weights directly trained on hateful memes dataset gives much lower performance than all other models. Hence, we could expect that directly trained model would be lacking in generalization. Among the models finetuned on pre-trained weights, the weights pre-trained on COCO and then finetuned on hateful memes gave the best performance and used for the following experiments.

Table 2
Comparison of different pre-trained weights

Name	F1
Direct	0.5390
Pretrained.coco	0.5175
Pretrained.vqa2	0.5424
Pretrained.cc	0.5405
Finetuned.hateful_memes.from_coco	**0.5488**
Finetuned.hateful_memes.from_coco	0.4763

4.2 Multi-task Classification Head

Table 3 compares performance of four multi-class classification heads on the validation set. Weighted F1 is used as the evaluation metric in this table. We also implemented classification networks for individual tasks. For single task classification, simply an FC layer is place after PRD layer that produces output for given tasks.

Multi-task head 1 that shares the least information among tasks showed the lowest performance. Multi-task head 3 achieved the best performance. This could be interpreted that predicting emotions as individual binary classification problems would be easier than solving it as a multi-label problem as head 4.

Models are trained for 100 epochs with AdamW optimizer and a learning rate of 5e-5. To avoid overfit due to small data size, dropout (with probability 0.2) and early stopping (after 4,000 epochs) are both employed. For test set, we obtained F1 scores of 0.5025, 0.7609 and 0.5564 for Task A, B and C, placing 6_{th}, 4_{th} and 1_{st} on each task.

Table 3
Comparison of classification heads

Name	Task A	Task B	Task C	Average
Multi-task 1	0.4934	0.6862	0.5691	0.5691
Multi-task 2	0.5160	0.7049	0.5545	0.5918
Multi-task 3	0.5090	**0.7101**	**0.5746**	**0.5979**
Multi-task 4	**0.5347**	0.6581	0.5174	0.5701
Single task A	0.5241	-	-	
Single task B	-	0.5523	-	
Single task C	-		0.5606	

5. Conclusions

In this paper, we presented our approach to the Memotion 2.0 Challenge. We tackled the problem as a multi-task classification considering correlation among individual tasks. Experiments showed that multi-task models could outperform single-task models.

Though we achieved competitive performance, there is much room for improvement. One direction could be end-to-end training of the multi-modal model. In this work, images features are extracted from an object detector. As the object detector is not updated during the training, image features might not be fully aligned with the downstream task. Exploring end-to-end training on grid image features (Huang el. al. 2020) would be interesting which we leave as a future work.

6. References

[1] S. Mishra, S. Suryavardan, A. Bhaskar, P. Chopra, A. Reganti, P. Patwa, A. Das, T. Chakraborty , A. Sheth, A. Ekbal and C. Ahuja, FACTIFY: A Multi-Modal Fact Verification Dataset, in: Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection , CEUR, 2022

[2] P. Patwa, S. Mishra, S. Suryavardan, A. Bhaskar, P. Chopra, A. Reganti, A. Das, T. Chakraborty, A. Sheth, A. Ekbal and C. Ahuja, Benchmarking Multi-Modal Entailment for Fact Verification, in: Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR, 2022

[3] L. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang. VisualBERT: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557, 2019.

[4] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia and D. Testuggine, The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. In: Proceedings of Neural Information Processing Systems, 2020.

[5] J. Lu, D. Batra, D. Parikh, and S. Lee, ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks, in: Proceedings of the 33rd International Conference on Neural Information Processing Systems, 2019, No. 2, pp. 13–23.

[6] H. Jiang, I. Misra, M. Rohrbach, E. Learned-Miller and X. Chen, In Defense of Grid Features for Visual Question Answering, in: Proceedings of Computer Vision and Pattern Recognition (CVPR 2020), 2020

[7] Z. Huang, Z. Zhaoyang, L. Bei, F. Dongmei, and F. Jianlong, Pixel-BERT: Aligning Image Pixels with Text by Deep Multi-Modal Transformers. 2020. arXive preprint arXiv:2004.00849

[8] S. Ren. K. He, G. Ross and J. Sun, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, in Proceedings of Advances in Neural Information Processing Systems 28, 2015.

[9] A. Singh, V. Goswami, V. Natarajan, Y. Jiang, X. Chen, M. Shah, M. Rohrbach, D. Batra, and D. Parikh, MMF: A multimodal framework for vision and language research, 2020. URL: https://github.com/facebookresearch/mmf.

[10] A. Singh, V. Goswami, and D. Parikh, Are we pretraining it right? Digging deeper into visio-linguistic pretraining. 2020, arXiv preprint arXiv:2004.08744.