Model documentation and write-up

1. Who are you (mini-bio) and what do you do professionally?

Riza Velioglu:

After receiving my bachelor's degree in Electrical & Electronics Engineering from Istanbul Bilgi University in 2018, I directly started to my master's degree in Intelligent Systems at Bielefeld University. Currently, writing the master's thesis and planning to graduate in February 2021. At the same time, I have been the Teaching Assistant of the courses; Intro to Machine Learning, Intro to Neural Networks, and Intro to Computer Algorithms for the previous 1.5 years. In addition, I am actively looking for PhD opportunities in the areas of Vision and Language Representations, Multi-Modal Deep Learning, Self-Driving Cars (Semantic Segmentation, Lane Detection, Image Classification, Object Detection, etc.)

Jewgeni Rose:

I completed my Master's degree in Computer Science at the University of Brunswick in 2017. After that, I started at the VW Group Innovation Center in Europe on the topic of digital assistants. At the same time, I began as a PhD student at the University of Bonn under the supervision of Prof. Dr. Jens Lehmann. My research area is Question Answering, specifically Context-aware Question Answering, whereby context is external knowledge (e.g. sensory data), chat history, or a user profile.

2. What motivated you to compete in this challenge?

(Riza)

I have a background on Image Processing and Natural Language Processing through university projects and as a personal interest. Therefore, Hateful Memes Challenge was a perfect chance for me to take part because it tries to leverage from joining the two modalities, that is multi-modality, which is perfect for someone working in both areas.

Moreover, I always try to do something that can have a direct impact in the real-world! With this challenge, it was possible to work on Hate Speech which we face in every day of our lives.

3. High level summary of your approach: what did you do and why?

The approach could be summed up as follows:

- Growing training set by finding similar datasets on the web,
- Extracting image features using object detection algorithms (Detectron),
- Fine-tuning a pre-trained V+L model (VisualBERT)
- Hyper-parameter search and applying Majority Voting Technique

Growing training set:

The more samples one has, the better score one will get! Therefore, we added more samples to the training data. We found out that there are 100 memes in *dev_seen* which are NOT in *dev_unseen*! First, we added those 100 memes to the training data.

Secondly, we searched for open-sourced datasets available on the web that is somewhat similar to Hateful Memes dataset. There was none because Hateful Memes was the first dataset on this regard. But we found the one called 'Memotion Dataset' which was planned to be one of the tasks at SemEval 2020, to be specific Task 8 ([Paper], [SemEval]). The dataset is open-sourced and can be downloaded [by this link].

In this dataset all 14k samples were annotated the class labels as *Humorous, Sarcasm, Offensive*, and *Motivational* and quantified the intensity to which a particular effect of a class is expressed. The effects of each of the class is given in the following table:

Humour	**Sarcastic**	**Offensive**	**Motivational**
Not Funny	Not Sarcastic	Not Offensive	Motivational
Funny	General	Slight Offensive	Not Motivational
Very Funny	Twisted Meaning	Very Offensive	
Hilarious	Very Twisted	Hateful Offensive	

For instance, a meme can be annotated as: *"Not Funny, Very Twisted, Hateful Offensive, Not Motivational".* We added all the *"Hateful Offensive"* and *"Very Offensive"* to the training data as hateful memes and added *"Not Offensive"* ones as non-hateful memes. We discarded the *"Slight Offensive"* memes because we thought it might confuse the model. With this training data the model did not achieve a better score. Later we found out that the memes are not labelled correctly (in our opinion). Therefore, we labelled the dataset manually! We searched for the memes that are somewhat similar to the ones in Hateful Memes challenge considering the idea of the challenge. After cherry-picking the "similar" memes, we ended up having 328 memes. Then, we added those 328 memes to the training set.

Extracting image features using object detection algorithms (Detectron):

After collecting the whole dataset, we extracted our own image features using Facebook's Detectron. To be specific we used Mask R-CNN and ResNet 152 as the backbone network architecture.

Fine-tuning a pre-trained V+L model (VisualBERT):

After trying different models and different pre-trained VisualBERT models, we found out that the model which was pre-trained on the Masked Conceptual Captions gave the best score. The pre-trained model is provided by Facebook's MMF and reachable via this link (pretrained key= *visual_bert.pretrained.cc.full*).

Hyper-parameter search and applying Majority Voting Technique:

We then did a hyper-parameter search and ended up having multiple models having different ROC-AUC scores on *"dev_unseen"* dataset. We sorted them by the ROC score and took the first 27 models (27 is chosen arbitrarily). Then, predictions are collected from each of the models and the majority voting technique is applied: the 'class' of a data point is determined by the majority voted class. For instance, if 15 models predicted Class1 for a sample and 12 models predicted Class0, the sample is labelled as Class1 cause it's the majority voted class. The 'proba' which stands for the probability that a sample belongs to a class is then determined as follows:

- if the majority class is 1, then the 'proba' is the **maximum** among all the 27 models
- else if the majority class is 0, then the 'proba' is the **minimum** among all the 27 models

We would argue that this technique works because it brings the "experts of the experts"! Imagine that one model is very good at -in other words it's expert in- detecting hate speech towards Woman but might not be an expert in detecting hate speech towards Religion. Then we might have another expert which has the opposite case. By using the majority voting technique, we bring such experts together and benefit from them as a whole.