MemoSYS at SemEval-2020 Task 8: Multimodal Emotion Analysis in Memes

Irina Bejan
University A. I. Cuza Iasi / Iasi, Romania
irina.bejan@info.uaic.ro

Abstract

Internet memes are one of the most viral types of content in social media and are equally used in promoting hate speech. Towards a more broad understanding of memes, this paper describes the MemoSys system submitted in Task 8 of SemEval 2020, which aims to classify the sentiment of Internet memes and provide a minimum description of the type of humor it depicts (sarcastic, humorous, offensive, motivational) and its semantic scale. The solution presented covers four deep model architectures which are based on a joint fusion between the VGG16 pre-trained model for extracting visual information and the canonical BERT model or TF-IDF for text understanding. The system placed 5th of 36 participating systems in the task A, offering promising prospects to the use of transfer learning to approach Internet memes understanding.

1 Introduction

Among all social networks, Internet memes are one of the most widespread types of content and have a huge virality factor. These memes are usually thought of as countless derivatives of template images, that become ingrained in our culture and very easily convey a message. While their most common purpose is to serve in a humorous or ironic way, they can unintentionally exhibit racism or promote hate speech that goes undetected because we lack a form of meme understanding at the moment. The attributes that illustrate the anatomy of a meme are the textual information, which consists of a couple of highly informal words, along with one (or more) images based on either TV shows or ”meme characters”. These characters are symbolic entities that set expectations on what kind of expression it conveys (Kolawole, 2015), an example being ”Bad Luck Brian”, which usually presents an unlucky, yet funny situation.

The trouble in understanding memes is caused by their humor - they cannot be taken individually, as standalone jokes. The best analog to internet memes is “joke cycles”, which are large clusters of mutually related texts (Attardo, 2001), sometimes relying on implicit references that can impose trouble in understanding even in human cognition. Secondly, the incredible amount of variations that can emerge from a meme, such as gross occlusion by textual elements, mutation of the meme characters, or even collages, makes it very challenging to discover features that can contribute to the classification.

An approach to tackle memes understanding has been proposed in the Memotion Analysis Task (Sharma et al., 2020), which consists of the following sub-tasks:

Task A - Sentiment Classification Given an Internet meme, the first task is to classify it as a positive, negative or neutral meme.

Task B - Humor Classification Given an Internet meme, the system has to identify the type of humor expressed. The categories are sarcastic, humorous, offensive and motivation meme. A meme can have more than one category.

Task C - Scales of Semantic Classes The third task is to quantify the extent to which a particular effect is being expressed on a scale from 0 to 3 for humour, offense and sarcasm.

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 1172–1178
Barcelona, Spain (Online), December 12, 2020.

To address this challenge we propose a set of four multimodal architectures that are based on fusioning and fine-tuning two pre-trained deep neural networks, in order to adequately combine the visual and the textual elements a meme provides.

2 Related Work

Memes Understanding: Internet memes have not been widely explored from the perspective of sentiment classification. Kolawole (2015) approached the classification task on a small personal dataset and used a linear SVM on handcrafted features, relying only on the visual information and not captions. More consistent efforts have been deployed into generating memes by representing the meme image and the catchphrase in the same word space based on a deep neural network (Kido Shimomoto et al., 2019), by using embeddings resulted from a pre-trained Inception-v3 network and feeding them to an attention-based layer to generate the caption (V and Tolunay, 2018) or based on a rule-based classifier (Gonçalo Oliveira et al., 2016). The quality was judged through human assessment and the results were better than random choices, but still far behind human-produced memes, not being able to produce the humor value.

Multimodal analysis: Great advancement has been made in the ability to detect the emotion and the sentiment depicted in multimodal content. While some focus was directed to the feature selection, mapping the input into the same latent embedding space (Katsurai and Satoh, 2016) or representing it as a cross-media bag-of-words (Wang et al., 2014), increasing attention is being invested into the usage of deep neural networks. Deep neural networks have outperformed traditional methods in understanding textual information (Wang et al., 2019) and image classification (Krizhevsky et al., 2012), showing successful results in multiple multimodal problems such as reasoning, question answering, machine translation or visual and caption generation (Mogadala et al., 2019). A successful attempt was combining features extracted from two CNN-based models for text and images using fully connected layers (Cai and Xia, 2015) . Recently, multiple alternative methods for multimodal fusioning have been analyzed (Liu et al., 2018)

Visual and language sentiment analysis Visual sentiment prediction has been extensively approached using domain transferred deep networks. Using deep convolutional networks previously trained on ImageNet for object recognition, such as AlexNet and VGG (Simonyan and Zisserman, 2014), to extract features from Twitter, Tumblr and Flickr datasets, in combination with SVMs and logistic regression classifiers yielded a set of great results (Xu et al., 2014), (Campos et al., 2016), (You et al., 2015). The task of text-based sentiment classification to identify sarcasm or humor currently has state of the art performance using transformer architectures (Potamias et al., 2019), such as BERT (Devlin et al., 2018), which offered promising results and outperformed most architectures for natural language inference (Munikar et al., 2019).

3 Dataset

The dataset proposed (Sharma et al., 2020) for the task consisted of 7000 memes, including their caption extracted and manually corrected. These memes include a great level of diversity, do not exhibit a similar structure, and do not represent a subset of meme templates or characters. Additionally to having no limitation regarding the content of the memes, there is also extra noise added by being only a part of the actual image, as in a screenshot of the phone screen or from social media platforms.

4 Preprocessing

The images that were chosen for the dataset present a big variety of quality and scales. We resized all images to address this issue to the size of 320x320, which seemed to preserve most of the features of the sub-images of the memes. Given the multiple font styles, sizes and placement of the caption, usually occluding a big part of the image, we aimed to ease the noise in the visual information by using the pre-trained EAST text detector (Zhou et al., 2017) to replace the textual elements with white boxes.

This allowed the neural network to learn to ignore such areas and to easier distinguish the features that contribute to the classification.
We have noticed that data augmentation, such as mirroring and random cropping of the images, helps the fine-tuning of the model. To train the models further, we split the dataset 70%-30% into training and validation subsets.

5 Models

For tasks B and C we trained four models for each attribute (humor, motivation, sarcasm, offense). Transforming the task B from multi-label classification to multiple binary classifications led to better results in all experiments, showing that our models were not able to learn and use the potential inter-dependencies between the attributes in a meaningful manner. To address the highly imbalanced data in all three sub-tasks and avoid biasing the prediction towards the majority classes, we applied a weighting factor in the loss computation that is inversely proportional to the number of samples of that class. We used Adam (Kingma and Ba, 2014) optimization for training, a learning rate starting with 0.0001 and a mini-batch size of 32. The stopping condition was observing whether any improvement has been encountered in the last three epochs. For implementation the Tensorflow 2.0 libray was used with the built-in Keras API.

5.1 Baseline

As a baseline, we opted for a unimodal approach that ignores the extracted caption. We extracted features by fine-tuning the pre-trained VGG16-ImageNet and fed it to a soft-max classifier layer for Task A and C with a categorical cross-entropy loss function, and to a sigmoid classifier for Task B with the binary cross-entropy loss function. Running only the VGG16 model inference for object recognition on memes mostly yielded classes like ”website”, ”internet website” or ”bookcase”, revealing the complexity of the input that stops the model from leveraging previous learning on memes, as memes might be made of collages or screenshots that are very different in structure from the photos used for training ImageNet. are more complex Internet memes with a lot of sub-images or noise.

5.2 TF-IDF + VGG16 Fusion

Given the good results TF-IDF has on sentiment analysis, we combined the features extracted by the TF-IDF from the meme captions after removing punctuation, POS tagging and lemmatization with high-level features learned by the convolutional layers of the VGG model through two locally connected layers with 256 hidden units each of the rectified linear type. This allows the model to learn adaptive weights for the input features, being able to measure the importance of each side, and generate a better-fused representation. Like in the baseline, we added on top of the locally-connected fusion a soft-max classifier layer for Task A and C and a sigmoid classifier on top for Task B. To avoid over-fitting, we apply Dropout after the fusion layers, as well as after feature extraction.

5.3 BERT + VGG16 Fusion using Softmax

BERT (Devlin et al., 2018) is a well-established model with state of the art results in sentiment classification on datasets such as Yelp, SST or IMDb (Sun et al., 2019). We hypothesized that we can leverage the language understanding of BERT on meme captions, therefore we fine-tuned the BERT base uncased model, which has 12 transformer layers, 12 self-attention heads and with a hidden size 768, along with the VGG16 and fusioned the extracted features using two locally-connected layers of 512 and 256 hidden units respectively of the rectified linear type. We fed it further to soft-max classifier for task A and C based on categorical cross-entropy loss function and to a sigmoid classifier task B based on the binary cross-entropy loss function to produce a distribution over the prediction classes. After the fusion of the inputs, we applied Dropout using a factor of 0.4. This model was used for submission and offered the best results in Task A (ranking 5), but we believe it could have ranked much better during the competition after more training in Task B (ranking 26) and C (ranking 23).

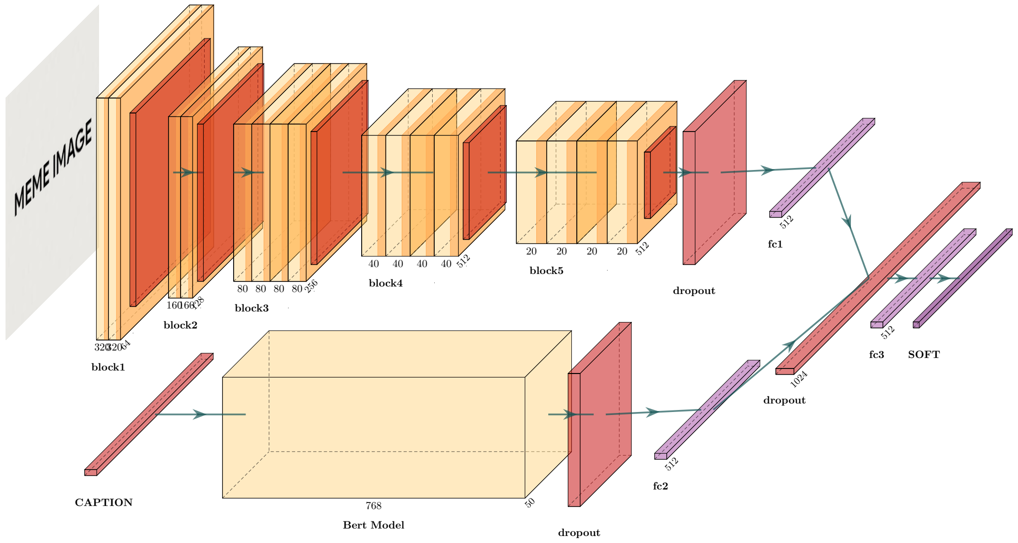

Figure 1: BERT + VGG Fusion using Softmax Architecture (Task A and C)

5.4 BERT + VGG16 Fusion using SVM

This model replaces the soft-max/sigmoid layer and the cross-entropy loss in the previous approach with a linear support vector machine and minimizes the squared hinge loss, method that has been proved to give better accuracy on CIFRAR-10 and MNIST dataset (Tang, 2013). For multiclass classification that cannot be done directly using the linear SVM, we use the One vs rest approach to obtain the final labels.

6 Analysis of Results

The results presented in Table 1 are based on the validation dataset resulted in the initial split, because not all models' predictions have been officially submitted and a fair comparison would be difficult to make otherwise. For reference to the official test data, our submission scores, the best overall results in the leader board and the baseline proposed by the task organizers can be found in Table 2.

Model	**Task A**		**Task B**		**Task C**	
	F1-macro	Accuracy	F1-macro	Accuracy	F1-macro	Accuracy
baseline	0.3268	0.4473	0.6920	0.5778	0.2314	0.3751
TF-IDF + VGG	0.3283	0.4668	0.6525	0.5400	**0.2864**	0.3063
BERT + VGG using SVM	0.2671	0.3938	0.7064	**0.6977**	0.2261	**0.3764**
BERT + VGG using Softmax	**0.3513**	**0.4988**	**0.7283**	0.6142	0.2833	0.2937

Table 1: My results on my test dataset (split from training set)

Model	**Task A**	**Task B**	**Task C**
My submission	0.3475	0.4519	0.2796
Top scores	**0.3546**	**0.5183**	**0.3224**

Table 2: My submission compared to top scores on leaderboard on hidden test data

7 Interpretability

It is nontrivial to understand the predictions made by these models, as memes understanding requires a high level of abstraction, cultural knowledge, and subjectivity that are hard to interpret even in human reasoning. Based on prediction results, we observed that less complex memes based on templates, such as ”Success Kid” meme, tend to be classified correctly. We considered this is probably linked to image understanding and we used the Grad-Cam approach (Selvaraju et al., 2019) to gather some visual explanations of the high level features the VGG16 network learns after training. It was surprising to see that many of the highlighted regions did not reveal any important information. However, on a certain subset, it seems that it highlights the face of the characters present), showing the strong tie meme understanding has with the character it is based on.

8 Challenges

As the dataset contains a very diverse collection of memes, it was difficult for all participants to score way above chance (33% for Task A, 50% for Task B, 25% for Task C). Although deep neural networks are powerful, it is of great importance to have enough training data to generalize learning. Extending the current dataset requires a certain level of consistency in the labeling process. People may have, for example, a different understanding of the humor depicted in the same meme and an even inaccurate rating of the sarcasm, the humor of offense. Since the images are weakly labeled, it may lead to poor generalizability of the network.

We believe that even more cleaning has to be made aside from removing caption, but it is a big challenge to automatically reduce the noise of the images, splitting them into components or removing the outer border from screenshots. The high-class imbalance also raised difficulties, even with balancing techniques as oversampling and class weights. For example, the dataset contains 1544 memes (out of 7000) labeled as being not sarcastic. This brings issues in classifying the input as sarcastic or not (Task B) because the model does not have a broad range of samples to learn from what the lack of sarcasm looks like, but rather naturally over-fit features of the small selection.

Having in mind the goal of identifying hate speech, it is crucial to address the interpretability of the model and the multimodal approaches add an increasing level of complexity.

9 Future Work

The experiments we worked on have shown that pre-trained deep neural networks, along with a certain level of pre-processing of the raw images give promising results in the classification of memes. A very big leap towards meme understanding would be acquiring a bigger and cleaner dataset to analyze to what extent the neural networks can generalize and learn useful features, through accurate labeling methods that are less sensitive to subjectivity. One way would be to focus on viral memes that already have a broad audience such as social media posts and derive their sentiment from Facebook reactions or Reddit comments. Furthermore, by focusing more on reducing the noise and emphasizing the features that contribute most to a person’s understanding of a meme, it is more likely to improve the accuracy. An example would be the knowledge associated with the most prevalent meme characters already documented by websites like KnowYourMeme.

10 Conclusions

Our results show that fusions of pre-trained models have the potential to overcome some of the challenges in memes understanding. This approach can distinguish in certain cases the humor, sarcasm, and most important, offensiveness, being a tool that we can work upon and use further to act against hate speech on social media networks, to asses the quality of the existing content and to advance the automatic generation of memes.

References

S. Attardo. 2001. *Humorous Texts: A Semantic and Pragmatic Analysis*. Humor research. Mouton de Gruyter.

Guoyong Cai and Binbin Xia. 2015. Convolutional neural networks for multimedia sentiment analysis. In *Proceedings of the 4th CCF Conference on Natural Language Processing and Chinese Computing - Volume 9362*, NLPCC 2015, page 159–167, Berlin, Heidelberg. Springer-Verlag.

Victor Campos, Brendan Jou, and Xavier Giro i Nieto. 2016. From pixels to sentiment: Fine-tuning cnns for visual sentiment prediction.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Hugo Gonçalo Oliveira, Diogo Costa, and Alexandre Pinto. 2016. One does not simply produce funny memes! – explorations on the automatic generation of internet humor. 06.

M. Katsurai and S. Satoh. 2016. Image sentiment analysis using latent correlations among visual, textual, and sentiment views. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2837–2841.

Erica Kido Shimomoto, Lincon Souza, Bernardo Gatto, and Kazuhiro Fukui. 2019. News2meme: An automatic content generator from news based on word subspaces from text and image. 05.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

Kolawole. 2015. Classification of internet memes.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Neural Information Processing Systems*, 25, 01.

Kuan Liu, Yanen Li, Ning Xu, and Prem Natarajan. 2018. Learn to combine modalities in multimodal deep learning.

Aditya Mogadala, Marimuthu Kalimuthu, and Dietrich Klakow. 2019. Trends in integration of vision and language research: A survey of tasks, datasets, and methods.

Manish Munikar, Sushil Shakya, and Aakash Shrestha. 2019. Fine-grained sentiment classification using bert.

Rolandos Alexandros Potamias, Georgios Siolas, and Andreas Georgios Stafylopatis. 2019. A transformer-based approach to irony and sarcasm detection.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2019. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, Oct.

Chhavi Sharma, William Paka, Scott, Deepesh Bhageria, Amitava Das, Soujanya Poria, Tanmoy Chakraborty, and Björn Gambäck. 2020. Task Report: Memotion Analysis 1.0 @SemEval 2020: The Visuo-Lingual Metaphor! In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, Sep. Association for Computational Linguistics.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration.

Yichuan Tang. 2013. Deep learning using linear support vector machines.

Abel L Peirson V and E Meltem Tolunay. 2018. Dank learning: Generating memes using deep neural networks.

Min Wang, Donglin Cao, Lingxiao Li, Shaozi Li, and Rongrong Ji. 2014. Microblog sentiment analysis based on cross-media bag-of-words model. In *Proceedings of International Conference on Internet Multimedia Computing and Service*, ICIMCS '14, page 76–80, New York, NY, USA. Association for Computing Machinery.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems.

Can Xu, Suleyman Cetintas, Kuang-Chih Lee, and Li-Jia Li. 2014. Visual sentiment prediction with deep convolutional neural networks.

Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2015. Robust image sentiment analysis using progressively trained and domain transferred deep networks.

Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. 2017. East: An efficient and accurate scene text detector.