Urszula Walińska at SemEval-2020 Task 8: Fusion of text and image features using LSTM and VGG16 for Memotion Analysis

Urszula Walińska and **Jedrzej Potoniec**
Faculty of Computing and Telecommunications
Poznan University of Technology
Poznan, Poland
urszula.walinska96@gmail.com,jedrzej.potoniec@cs.put.poznan.pl

Abstract

In the paper, we describe the Urszula Walińska's entry to the SemEval-2020 Task 8: *Memotion Analysis*. The sentiment analysis of memes task, is motivated by a pervasive problem of offensive content spread in social media up to the present time. In fact, memes are an important medium of expressing opinion and emotions, therefore they can be hateful at many times. In order to identify emotions expressed by memes we construct a tool based on neural networks and deep learning methods. It takes an advantage of a multi-modal nature of the task and performs fusion of image and text features extracted by models dedicated to this task. Our solution achieved 0.346 macro F1-score in Task A – Sentiment Classification, which brought us to the 7th place in the official rank of the competition.

1 Introduction

Nowadays, spreading of abusive content in social media is becoming an increasingly serious problem. Due to the anonymity provided by the Internet, users posting offensive content feel that they can ignore the law with impunity. A remedy to this problem is automatic detection and elimination of such content, and a tremendous progress was made in the past years in the area of sentiment analysis of text (Mäntylä et al., 2018). However, little research was done when it comes to image sentiment analysis, which is a grave omission, as people frequently express emotions in a multi-modal way, using both text and image. A particular example of such multi-modal and potentially harmful content are Internet memes.

In this paper we tackle a problem of Memotion Analysis, which is the 8-th task of SemEval-2020 competition. Dataset with memes in English, provided by Sharma et al. (2020) defines the following sub-tasks:

Task A – Sentiment Classification Given an Internet meme, classify it as a positive, negative or neutral meme.

Task B – Humor Classification Given an Internet meme, identify expressed humor type: *sarcastic*, *humorous*, *offensive*, *motivation*. A meme can express more than one humor type at the same time.

Task C – Scales of Semantic Classes Quantify the extent to which humor types are being expressed by the given meme, using the following 4-point scale scale: *not* (0), *slightly* (1), *mildly* (2) and *very* (3).

We participated only in the Task A and the consecutive analysis is based only on this task. The dataset provided for this task contains 6946 unique, complete examples. Each example consists of the following features: an image name, an image URL, text extracted using OCR from the image, corrected text and overall sentiment. The sentiment is expressed on the following 5-point scale: very positive, positive, neutral, negative, negative, very negative, but in the Task A defining whether given example is positive, neutral or negative is sufficient, i.e., very positive and positive were considered as a single class positive, while negative and very negative as a single class negative.

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 1215–1220
Barcelona, Spain (Online), December 12, 2020.

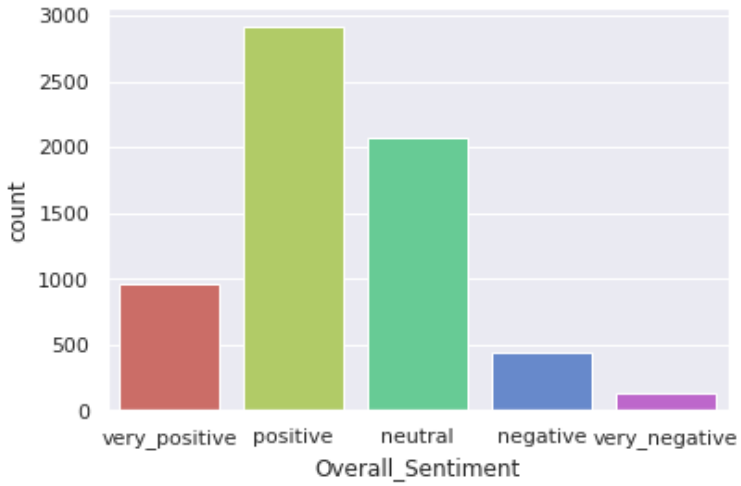

Figure 1: Class imbalance in Task A. Negative and very negative memes represent the minority, but are the most important for the use case of automatic moderation.

Figure 1 presents the distribution of the classes across the whole dataset and we observe that they are heavily imbalanced. Even if we collapse categories into a single negative, a single positive and a neutral category, the problem still remains.

While developing our solution we split provided dataset into training, validation and test set in the proportion 8:1:1. Organizers of the competition provided us also with the official test set for the final evaluation.

2 Related work

Sentiment analysis of text is an established field within the computer science, employing numerous approaches such as dictionaries (Hu and Liu, 2004), ontologies (Dragoni et al., 2018), statistical approaches (Turney and Littman, 2003), machine learning (Pang et al., 2002) or a combination thereof. Recently, deep learning became a very popular tool, e.g., (Howard and Ruder, 2018).

The area of image sentiment analysis is much younger, and rooted mostly in machine learning and deep learning techniques, e.g., (You et al., 2015), however, images alone are rarely considered and instead the problem of multi-modal classification is considered (Soleymani et al., 2017).

A notable example is work by Sabat et al. (2019), concerned with using visual and textual information to automatically detect hate speech in Internet memes. The authors built a dataset of 5,020 memes to train and evaluate a model capable of identifying hateful memes. To extract text from images, they performed OCR first and later used a pre-trained BERT model (Devlin et al., 2019) to extract relevant features from the text. For image feature extraction, a pretrained VGG-16 network (Simonyan and Zisserman, 2015a) was used. The features for both modalities were then combined using a multi-layer perceptron (MLP). Moreover, their experiments indicated how the visual modality can be much more informative for hate speech detection than the linguistic one in memes.

3 Methodology

3.1 General concept

The general concept of a solution to multi-modal sentiment classification is presented in the Figure 2 and basically follows an architecture proposed in (Sabat et al., 2019). First, we perform text preprocessing. Then, we separately extract text and image features using appropriate models and methods. Later, we freeze the features and concatenate them to obtain a unified representation of a given example over its both modalities. Finally, we use a dense feed-forward neural network to perform final classification over the unified representation.

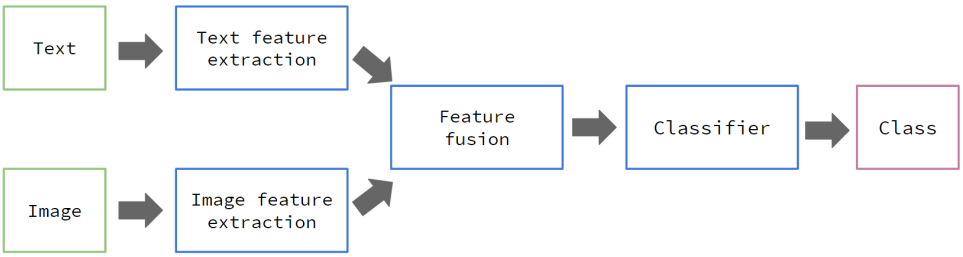

Figure 2: General concept of a solution

3.2 Text preprocessing

As our solution uses pre-trained word embeddings down the pipeline, we decided to perform preprocessing in order to ensure that the embeddings are available for as many words as possible. We started with basic operations such as removing white spaces, accented and special characters, numbers, punctuation and stop words. Moreover, we expanded all contractions to full forms, performed lemmatization and corrected misspellings as language used in memes is often sloppy. As some of the memes contain hash tags, which can convey an important message when it comes to sentiment analysis, we decided to split them, so that our model is able to understand them, e.g., *#10YearChallenge* $\rightarrow$ *10 Year Challenge*.

3.3 Text feature extraction model

To be able to solve any NLP problem, we need to find an appropriate data representation first. In this case, we decided to use pre-trained GloVe embeddings, which basically are vectors of numbers used in the subsequent method of classification. GloVe (Pennington et al., 2014) stands for Global Vectors for Word Representation and is an unsupervised learning algorithm for obtaining vector representations of words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations display interesting linear substructures of the word vector space. We decided to use word vectors pre-trained on tweets (2B tweets, 27B tokens, 1.2M vocabulary, 100d vectors).

In order to obtain an informative single feature vector for the whole text contained by an example, we created and trained a neural network model. Embeddings were included as the first layer in our model and were freezed during the training.

The second layer of the model was a Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997), a kind of recurrent neural network (RNN) able to process entire sequences of data. It remembers previous states and uses them to predict the next ones. One of their most prominent use cases is NLP. We used a layer of 100 LSTM units, each unit consisting of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell.

RNNs tend to overfit very often and in order to alleviate this effect a recurrent dropout was used with the probability of 0.2 (Semeniuta et al., 2016).

Subsequently, we added also a single dense layer, which contained 3 units with softmax activation. We employed Adam optimizer (Kingma and Ba, 2015) with the default parameter settings to train the obtained neural network with categorical cross-entropy loss. We used also early-stopping with F1-macro score on the validation set. As the classes are heavily imbalanced, we mitigate this problem by levelling instances' importance. Each example is assigned a weight inversely proportional to the total number of examples in the example's class. This way, we assign higher weight to examples of class with small number of instances, and therefore indicate its higher importance. Moreover, to evaluate our prediction we used F1-score (macro) which is robust when it comes to class imbalance problem.

Table 1: Evaluation of individual models of single modalities and feature fusion

Method	**Macro F1**
text only	0.32
image only	0.34
fusion	*0.39*

3.4 Image feature extraction model

To perform feature extraction from images we used a VGG-16 model (Simonyan and Zisserman, 2015b) pre-trained on the ImageNet dataset (Deng et al., 2009), available in *Keras*[1] package *applications*. VGG-16 is a deep convolutional neural network (CNN) for image classification, of state-of-the-art performance on ImageNet dataset in the year of publication. As the classes of the original ImageNet dataset do not match our case, we removed the dense part of the network, obtaining a fully convolutional network (FCN).

In order to obtain image features of a constant shape from the FCN, we resized all the images to the size of 224×224, yielding 25088-dimensional feature vectors (after flattening).

3.5 Feature fusion and classification

Since image feature vector has 25088 dimensions (much more than a text feature vector, which has 100 dimensions), we forced the network to reduce the dimensions into 100, by using dense layer with 100 neurons after image feature input. Later, in order to combine both modalities, we used the concatenate layer of Keras to obtain a single feature vector for classification, consisting of 200 dimensions.

We then attached a dense feed-forward neural network consisting of 2 layers with, respectively, 100 and 3 units. The layers are separated with a dropout layer with the rate of 0.05. All the layers except for the last one used the ReLU activation function, and the last one used the sofmax activation.

We performed the training using the same setting as described in subsection 3.3.

4 Experiment

In order to decide whether using feature fusion has any advantages over using any single modality, we performed the following experiment.

The evaluation is based on the results obtained by each method using a test set according to macro F1-score. The results on the test set are presented in Table 1.

The *text only* approach is the result of the complete network (with the dense part) as described in subsection 3.3. The *image only* approach consisted of features extracted using the method described in subsection 3.4 and an additional dense feed-forward network consisting of 3 dense layers with 512, 100 and 3 neurons, respectively. Layers were separated with dropout with the rate of 0.1. Network was trained under the same regime as described in subsection 3.3.

We observe that feature fusion was the best of the considered methods (F1 score of 0.39) and we hypothesise that this was due to the richer representation obtained by feature fusion, increasing the level of flexibility to fit to the data. Seeing that incorporating visual information increased F1 score by 0.07 (0.32 to 0.39), we hypothesise that the visual information is of utmost importance for the problem of sentiment analysis of memes, and that the knowledge derived from text and from images is, to some extent, complementary. This is also in line with the results of Sabat et al. (2019).

5 Competition results

We took part in Task A of Memotion Analysis competition. Taking into consideration the results of experiment we obtained, our final solution was based on the model with feature fusion of LSTM (textual modality) and VGG-16 (visual modality). Model was trained on training set provided by organizers of Memotion Analysis 2020 and described in the introduction. Our result on the official test set including obtained F1-scores and place in the official rank is presented in Table 2.

[1] `https://keras.io`

Table 2: Our results in Task A

Macro F1	**Micro F1**	**Place**
0.346	0.468	*7th*

6 Conclusions

Taking into account number of cases of public discrimination and humiliation of people on the grounds of race, ethnic origin or religion, we can undoubtedly say that dealing with these pervasive problems in social media is of utmost importance. What is more, we need to focus not only on the hate visible in text posted by people but also on memes which more and more popular way of expressing emotions. Automatic sentiment analysis methods, particularly the method presented in this paper, can help to reduce harmful impact of hate in social media.

As the results show, fusion of features of different modalities using neural networks can be a very powerful tool for multi-modal sentiment analysis. Textual information might enrich visual information and vice versa, therefore the topic of multi-modal classification is very interesting and is undoubtedly worth further analysis also in other tasks.

Acknowledgements

Urszula Walińska executed the research as a part of master thesis project under the supervision of Jedrzej Potoniec. This work was partially funded by project 0311/SBAD/0678.

References

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Mauro Dragoni, Soujanya Poria, and Erik Cambria. 2018. Ontosenticnet: A commonsense ontology for sentiment analysis. *IEEE Intell. Syst.*, 33(3):77–85.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 328–339. Association for Computational Linguistics.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In Won Kim, Ron Kohavi, Johannes Gehrke, and William DuMouchel, editors, *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004*, pages 168–177. ACM.

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

Mika V. Mäntylä, Daniel Graziotin, and Miikka Kuutila. 2018. The evolution of sentiment analysis - A review of research topics, venues, and top cited papers. *Comput. Sci. Rev.*, 27:16–32.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, EMNLP 2002, Philadelphia, PA, USA, July 6-7, 2002*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL.

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

Stanislau Semeniuta, Aliaksei Severyn, and Erhardt Barth. 2016. Recurrent dropout without memory loss. In Nicoletta Calzolari, Yuji Matsumoto, and Rashmi Prasad, editors, *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 1757–1766. ACL.

Chhavi Sharma, Deepesh Bhageria, William Paka, Scott, Srinivas P Y K L, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. SemEval-2020 Task 8: Memotion Analysis-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. 2015a. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Karen Simonyan and Andrew Zisserman. 2015b. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Mohammad Soleymani, David García, Brendan Jou, Björn W. Schuller, Shih-Fu Chang, and Maja Pantic. 2017. A survey of multimodal sentiment analysis. *Image Vis. Comput.*, 65:3–14.

Peter D. Turney and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Trans. Inf. Syst.*, 21(4):315–346.

Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2015. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In Blai Bonet and Sven Koenig, editors, *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 381–388. AAAI Press.