PRHLT-UPV at SemEval-2020 Task 8: Study of Multimodal Techniques for Memes Analysis

Gretel Liz De la Peña Sarracén and Paolo Rosso and Anastasia Giachanou
Universitat Politècnica de València, València, Spain
gredela@posgrado.upv.es
prosso@dsic.upv.es
angia9@upv.es

Abstract

This paper describes the system submitted by the PRHLT-UPV team for the task 8 of SemEval-2020: Memotion Analysis. We propose a multimodal model that combines pretrained models of the BERT and VGG architectures. The BERT model is used to process the textual information and VGG the images. The multimodal model is used to classify memes according to the presence of offensive, sarcastic, humorous and motivating content. Also, a sentiment analysis of memes is carried out with the proposed model. In the experiments, the model is compared with other approaches to analyze the relevance of the multimodal model. The results show encouraging performances on the final leaderboard of the competition, reaching good positions in the ranking of systems.

1 Introduction

Nowadays, social networks have become one of the main means of messages spread. Shared information usually consists of different types of content. In this sense, we can find publications of videos, images, texts or audios. Memes are examples of posts that usually combine image and text. Daily the number of memes on popular sites like Twitter and Facebook increases considerably. This makes automatic memes analysis a very important task at present. Often, memes contain subjective information, and can have ironic or sarcastic content. Irony and sarcasm are forms of figurative speech in which authors write the opposite of what they mean (Hernández Farıas and Rosso, 2016). Irony is an umbrella term whereas sarcasm is considered by many researchers a more direct and sometimes somewhat aggressive form of irony. Therefore, the analysis of memes requires an extensive study that combines several tasks. On the one hand, it is necessary to design multimodal systems that can handle both textual and visual information. On the other hand, different studies should be carried out to obtain all the semantic information that the texts may contain. In this way, the study can involve sentiment analysis, irony or sarcasm detection, and so on. Offensive language detection is an important task among them. It has gained great interest with the growth of the spread of offenses on Internet and its implication in society. The Memotion Analysis shared task has been organized in order to bring the attention towards Internet memes processing. The three subtasks defined in the task are Sentiment Classification (subtask A), Humour Classification (subtask B) and Scales of Semantic Classes (subtask C). The main goal of the subtask A is to classify a meme as positive, negative or neutral according to sentiment content. The subtask B aims to identify the type of humour expressed among the categories: sarcasm, humour, offense and motivation. It should be noted that many researchers do not consider sarcasm as a humour category since irony/sarcasm partially overlap with humour (Reyes et al., 2012). Finally, the subtask C focuses on quantifying the level to which a particular category is expressed.

We propose a multimodal model that combines the analysis of textual and visual information. The pretrained BERT base model is used for the text processing, and a pretrained VGG model for images. Features obtained from both text and image analysis are combined to feed a simple classifier that obtains the final categories. We perform an ablation study in course of the proposal design to analyze the

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 908–915
Barcelona, Spain (Online), December 12, 2020.

importance of the multimodal model. Furthermore, we analyze different models rather than BERT and VGG, including traditional machine learning models such as Support Vector Machines and Logistic Regression.

The rest of the paper is organized as follows. Next section briefly describes the principal approaches used in multimodal models. The details of the proposed methodology for each subtask and the dataset are described in Section 3. Experimental results are presented in Section 4, and finally the study is concluded in Section 5.

2 Related Work

Recently, different proposals for some tasks have focused on multimodal analysis. In general, strategies can be divided into two approaches (Corchs et al., 2019), regardless of whether they are based on deep learning or not. On the one hand, some models use a feature level fusion approach, where each input source, for instance text or image, is processed to extract a set of features. Then, the features sets are put together for the final decision (Atrey et al., 2007; Wang et al., 2018). On the other hand, other models carry out a complete processing of each input source and the fusion is made at the decision level (Atrey et al., 2010; Poria et al., 2017). Our proposal is based on the first category.

3 Methodology and Dataset

The dataset provided for the task consists of a set of 7000 memes (Sharma et al., 2020). Some information is supplied for each meme with the image, such as the url and the text extracted from the image. For the latter one, the text extracted by an Optical Character Recognition (OCR) system and the corrected text are given. Each meme contains labels for the five categories analyzed in the task. The categories and their labels are listed below.

- **Sentiment Analysis:** *very positive, positive, neutral, negative, very negative*
- **Offense:** *not offensive, slight, hateful offensive, very offensive*
- **Humour:** *not funny, funny, very funny, hilarious*
- **Sarcasm:** *not sarcastic, general, twisted meaning, very twisted*
- **Motivation:** *not motivational, motivational*[1]

The labels are ordered by level from left to right for each category. Then, the *not name_category* tag is level zero in each case (Offense, Humour, Sarcasm, Motivation). There are other level for the Motivation category and 3 levels from 1 to 3 for the rest of categories.

The sentiment analysis category corresponds to the subtask A, where a meme must be classified as positive, neutral or negative, considering very positive as positive, and very negative as negative. The rest of the categories are used in the subtasks B and C. In the subtask B, the category present in each meme must be identified. The memes can contain none, one or more categories. In the subtask C, each category is analyzed independently to identify its level according to the labels in the dataset.

3.1 Dataset Details

Table 1 summarizes the distribution of labels per category, where a large imbalance is observed in each category. An aspect that draws attention when analyzing the data is the relationship between the Offense category and the labels for the Sentiment Analysis. The 41% of negative memes (negative and very negative) are offensive (levels 1 to 3 of the category), whereas in case of positives (positive and very positive) this percentage is 46%. Thus, among positive memes there is a higher percentage of offensive memes than among negative ones. However, we would normally expect a greater relationship between the offensive and the negative content. On the other hand, when analyzing the overlapping between pairs of categories, we observe that the highest percentage of overlapping is between Humour and Offense categories, followed by the pair Sarcasm and Offense.

[1] We define a motivational meme as a meme that develops a positive attitude in the receiver.

Labels	Categories				
	SA	Off	Hum	Sar	Mot
0	1034	2715	1651	1545	4530
1	3131	2596	2457	3512	2470
2	2204	221	2241	1549	—
3	480	1468	651	394	—
4	151	—	—	—	—

Table 1: Category statistics: Sentiment Analysis (SA), Offense (Off), Humour (Hum), Sarcasm (Sar) and Motivation (Mot). The labels are represented by values from 0 to 4. In SA the labels are very positive (0), positive (1), neutral (2), negative (3) and very negative (4). In Mot the labels are not motivational (0) and motivational (1). The labels for the rest of the categories are the levels in each case in the range of 0 to 3.

Often the labeling of the memes for each category is not very clear given that the annotation of these categories can be very subjective. This can affect the training of machine learning models and, therefore, the results of automatic classification obtained with those models.

3.2 Methods

Our proposal for the subtasks B and C consists of a model that extracts features from the memes for the categories: offense, sarcasm, humour and motivation at the same time. The model is composed of one model per category. The categories are fused to obtain an overall classification as shown in the Figure 1. The outputs of the models per category correspond to the subtask C, while the output of the general model corresponds to the subtask B.

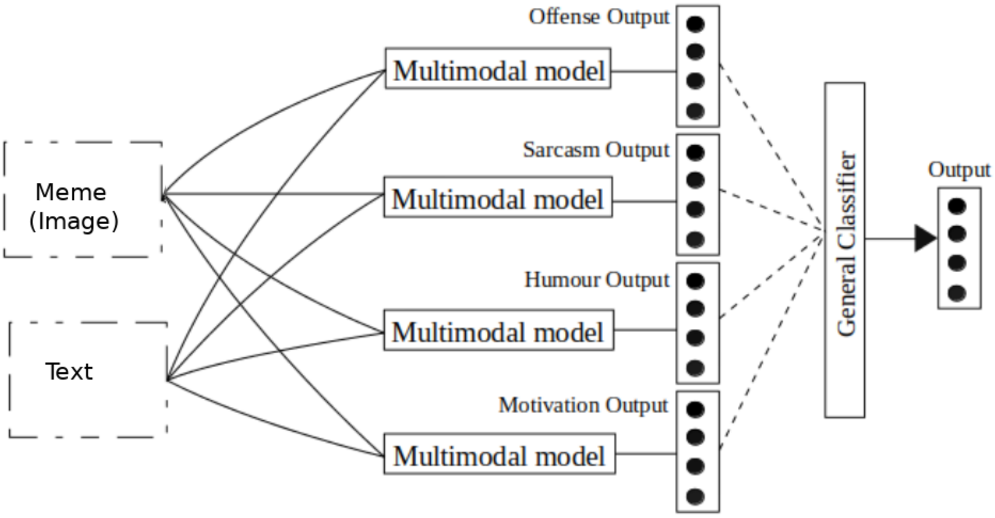

Figure 1: General model for the categories: offense, sarcasm, humour and motivation

The models used in each of the categories for the subtask C are multimodal, so that they combine features extracted from the text and the image. They have the same architecture, which is detailed in the next section. In sentiment analysis, the architecture is expanded to take into account other textual features. In this end, 3 lexicons are used as external resources for the text polarity analysis.

The global classifier for the subtask B is based on a very simple idea, which identifies the presence of a category in the meme if the corresponding label, assigned before in the subtask C, is positive. That is, the assigned label is different from the level 0 (*not name_category*). For a better understanding it must be remembered that each category has different possible levels. In each case, the first level (level 0) corresponds to the absence of the category in a meme. The other levels define to a greater or lesser degree the presence of the category. What the classifier does is interpret a category to be present in the meme if the detected level for that category by the corresponding model is greater than zero.

3.3 Multimodal Model

The multimodal model is composed of both the model that analyzes the image information and the model that analyzes the text, as it is depicted in Figure 2.

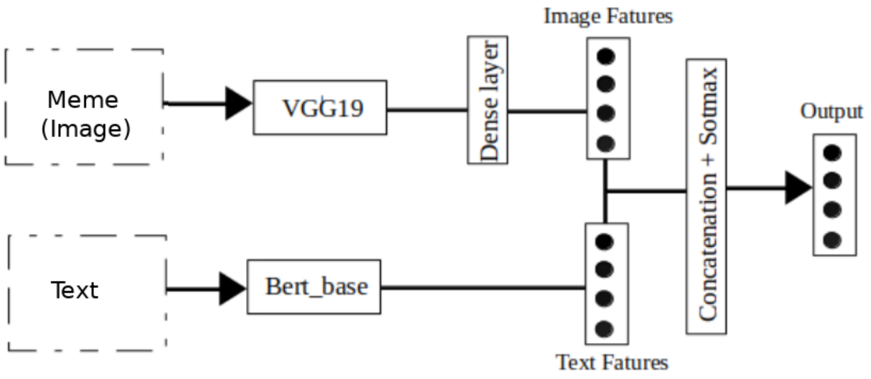

Figure 2: Multimodal model

The image features vector ϕ_{Im} is concatenated with the textual features vector ϕ_{Tx} to obtain a general features vector. This vector is a high level representation of a meme and is used in the final softmax layer to obtain the output as equation 1.

$$\phi = softmax(W^s \cdot [\phi_{Im}, \phi_{Tx}] + b^s) \tag{1}$$

where $W^s \in \mathbb{R}^{N_c*(V+T)}$ and $b^s \in \mathbb{R}^{N_c}$ are the parameters for the softmax layer. N_c is the size of the model output, corresponding to the number of possible labels into the category analyzed for the model. V and T are the dimensions of the visual and textual features vectors respectively. Finally, cross entropy is used as the loss function, defined as equation 2, where y_i is the true classification of the meme.

$$L = -\Sigma_i \; y_i * log(\phi_i) \tag{2}$$

3.3.1 Model for Image Processing

A model based on the VGG19 architecture is used in the visual information analysis of a meme. The input is the corresponding image from the meme and the output is a features vector. VGG19 is a deep neural network with 16 convolutional layers and 3 fully connected layers, that is 19 layers deep (Simonyan and Zisserman, 2014). Basically, the model consists of the sequence of the VGG19 layers and a dense layer, so that the last layer of VGG19 feeds the dense layer with the ReLU activation function. The features vector is obtained with the last dense layer. Then, the visual features vector ϕ_{Im} is obtain by the equation 3, where ϕ_{VGG} is the output of the last layer from VGG19. $W^{Im} \in \mathbb{R}^{V*|\phi_{VGG}|}$ and $b^{Im} \in \mathbb{R}^V$ are parameters to learn.

$$\phi_{Im} = ReLU(W^{Im} \cdot \phi_{VGG} + b^{Im}) \tag{3}$$

3.3.2 Model for Text Processing

The text from the meme is analyzed with the pretrained BERT base model. Then, the features vector is obtained as the mean of the vectors of all tokens in the last layer of BERT (layer 12). BERT is a model with state-of-the-art results on many tasks. It is composed of several transformer encoders stacked together, where the multi-head self-attention mechanism is used (Devlin et al., 2018). The textual features vector ϕ_{Tx} is obtained by the equation 4, where $\{\phi_{BERT}\}_s$ is the output for all the tokens in the last layer of BERT.

$$\phi_{Tx} = mean(\{\phi_{BERT}\}_s) \tag{4}$$

3.4 Sentiment Analysis Model

Sentiment analysis is performed with the same architecture of the proposed multimodal model. In addition, 3 lexicons are used to analyze the polarity of the texts.

We use the NRC emotion lexicon (Mohammad and Turney, 2013), the SenticNet sentiment lexicon (Biagioni, 2016) and a lexicon based on WordNet (Miller, 1995). A two-dimensional vector is obtained from each lexicon, with scores for the positive and negative text polarities. Then, the features vector of the lexicons is the concatenation of the vectors obtained with each of them. Hence, the textual features vector is a combination of the BERT output and the features from the tools instead of only the first features.

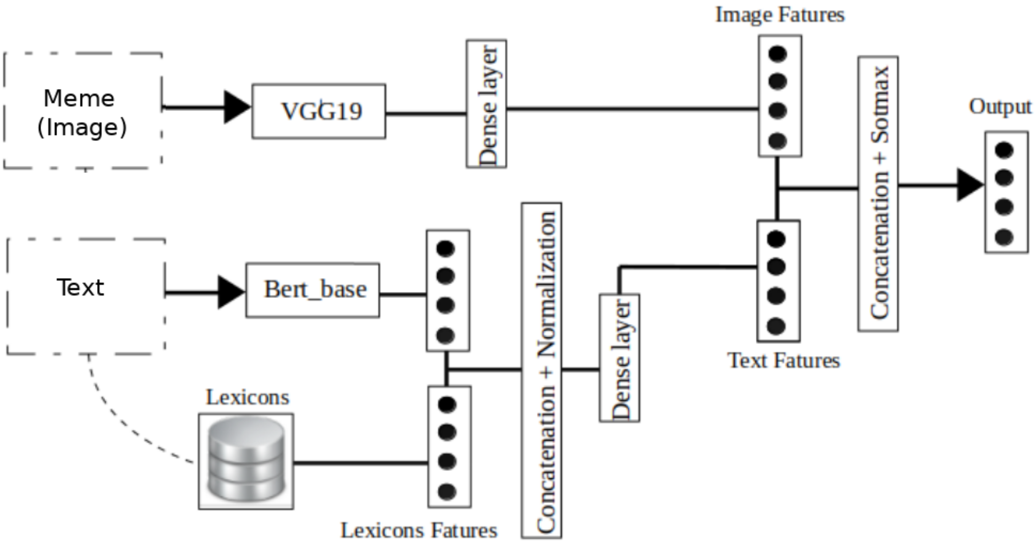

Figure 3: Multimodal model for sentiment analysis

As Figure 3 shows, the features vector obtained with the lexicons is concatenated with the vector from BERT, and a normalization layer is applied to the result. Then, the vector feeds a dense layer to generate the new textual features vector.

4 Experiments and Results

To measure the performance of the approaches we use the metrics proposed in the Memotion Analysis task. For the subtask A: macro F1, and for the subtasks B and C: macro F1 for each of the subtasks, and then the average.

4.1 Our Baselines

We used three simple models as baselines that only take into account the text of the meme. Each model is based on one of the following classifiers: Random Forest (**RF**), Logistic Regression (**LR**) and Support Vector Machines (**SVM**). The parameters were selected by optimization with the GridSearchCV[2] tool from the sklearn library.

Additionally, we use a model based on Support Vector Machines whose input is the concatenation of the text representation with the image representation (**SVM-image**). The image representation is basically the vector of pixels, whose dimensionality was reduced with the Principal Component Analysis.

[2] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

4.2 Implementation Details

The dimension of each dense layer was fixed to 100. For all the models, we use the same batch size of 50 instances in the training with 20 epochs.

The number of layers of BERT trained for fine tuning was 5 and all the layers of VGG19 were trained. For the baselines, the representation based on TF-IDF word ngrams was used for the texts.

4.3 Results

Different models were evaluated in the experiments. We varied the model used for image analysis while keeping the model for text processing and vice versa. **DenseNet** (Huang et al., 2017) and **NasNet** (Zoph et al., 2018) architectures were evaluated for the images. Furthermore, a Convolutional Neural Network (**CNN**) with a convolutional layer of 32 filters of 3x3 and a maxpooling layer of 2x2 was used. The model evaluated for the text was a Bidirectional LSTM network (**BiLSTM**). In this last case, the number of units is 64 and the FastText words embeddings were used for text representation.

Table 2 shows the results obtained for each of the models evaluated. In general, the results are very low for all models and the insignificant differences among the scores makes the comparison of the considered mechanisms hard. This may indicate that the models fail to learn correctly from the data. As discussed in Section 3.1, some labels are unclear and this might have influenced the model training. However, we can see a small improvement with BERT instead of BiLSTM for text analysis, and with VGG instead of DenseNet and NasNet for images.

Model	Subtask A	Subtask B	Subtask C	Average
Proposed Model	0.4285	0.6487	0.2847	0.4300
Baselines				
SVM	0.3355	0.6414	0.2838	0.4202
LR	0.2858	0.5857	0.1856	0.3524
RF	0.2482	0.5924	0.2250	0.3552
SVM-image	0.2478	0.5638	0.1507	0.3808
Varying model for image processing				
DenseNet	0.3244	0.6228	0.2721	0.4064
NasNet	0.3211	0.6380	0.2803	0.4132
CNN	0.3279	0.6386	0.2854	0.4173
Varying model for text processing				
BiLSTM	0.2823	0.4257	0.1416	0.2832

Table 2: Macro F1 in all tasks including average

Even the baselines, which only take into account the text, obtain very similar results. It should be noted that in the baseline where visual information is taken into account, the results are worse. This does not have to mean that the idea of taking only text is better, but rather may be due to the strategy used that does not allow to extract good visual features.

In addition, Table 3 shows the results obtained with different techniques to deal with data imbalance for the subtask A. The random undersampling technique is used in the proposal, so that some samples from the majority classes are removed. Another technique evaluated was random oversampling which involves supplementing the training data with multiple copies of some of the minority classes.

Furthermore, class weighting was evaluated as another alternative. For this, it is used a dictionary that maps the indices of each class to a weight value. Thus, a greater weight is assigned to the classes with less representation. It was used for weighting the loss function during training. This can be useful for the model to pay more attention to samples from a class less represented. The weight vector (class_weight) has been obtained with the equation 5, where Prior(classes) is the prior probability distribution over the classes in the dataset.

Technique	Subtask A
Undersampling (used)	**0.4285**
Oversampling	0.4093
Class weights	0.4116
No sampling	0.3279

Table 3: Macro F1 results of different techniques that deal with data imbalance

$$class\_weight = \frac{1}{log(1 + Prior(classes))} \quad (5)$$

The results are worse when the data imbalance problem is not addressed, as almost all samples are classified in the majority class. Undersampling is the technique with the best performance, hence we used it in the submission, although the results are not very far from those obtained with the other techniques.

5 Ablation Analysis

Table 4 shows a comparison among the results obtained with the proposed model and variants in which some of its components are removed. Basically, in each evaluated model, one of the models that processes text or images was eliminated. As before, it is difficult to carry out a good analysis since the results suggest that the models do not learn well. However, it should be noted that when a simple model is used instead of the multimodal model, the results are even worse. They are even worse than the SVM baseline, with which practically the same results are obtained as with the proposed model.

Model	Subtask A	Subtask B	Subtask C	Average
Multimodal Model	**0.4285**	**0.6487**	**0.2847**	**0.4300**
Text Model	0.2569	0.4880	0.1678	0.3042
Image Model	0.2563	0.5311	0.2043	0.3306

Table 4: Macro F1 results

5.1 Results in the test Set

Table 5 shows a summarization of the results obtained in the test set. The number of participants was 35, 31 and 28 for the subtasks A, B and C respectively, where our system reached the positions 19, 4 and 6. In the the subtask A, the baseline of the organizers was positioned at the end of the ranking. When analyzing the general ranking, it can be seen that even the best results are around to the value 0.5, which suggests once again that the models do not learn correctly.

Model	Subtask A		Subtask B		Subtask C	
	Position	Macro F1	Position	Macro F1	Position	Macro F1
Best system	1	0.3546	1	0.5183	1	0.3224
Our proposal	19	0.3355	4	0.5093	6	0.3143
Last system	35	0.2477	32	0.4002	29	0.1267
Baseline	36	0.2176	10	0.5002	19	0.3008

Table 5: Summary of the results in the test set

6 Conclusion

In this work, we studied the problem of Internet memes analysis as part of a shared task of SemEval 2020. We proposed a multimodal approach for each subtask that combines both textual and visual modalities. The architecture contains the BERT model for text processing and VGG19 for images. We obtained a good ranking position regarding the second subtask. However, our experimental results do not allow a thorough analysis of the proposed approach, obtaining less than 0.65 of macro F1 in the subtask B and a lower performance for subtasks A and C.

Acknowledgements

The research work was partially funded by the Spanish MICINN under the project MISMIS-FAKEnHATE on Misinformation and Miscommunication in social media: FAKE news and HATE speech (PGC2018-096212-B-C31). Anastasia Giachanou is supported by the SNSF Early Postdoc Mobility grant P2TIP2_181441 under the project Early Fake News Detection on Social Media, Switzerland.

References

Pradeep K Atrey, Mohan S Kankanhalli, and John B Oommen. 2007. Goal-oriented optimal subset selection of correlated multimedia streams. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 3(1):2–es.

Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. 2010. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6):345–379.

Raoul Biagioni. 2016. *The SenticNet Sentiment Lexicon: Exploring Semantic Richness in Multi-Word Concepts*, volume 4 of *SpringerBriefs in Cognitive Computation*. Springer International Publishing.

Silvia Corchs, Elisabetta Fersini, and Francesca Gasparini. 2019. Ensemble learning on visual and textual data for social image emotion classification. *International Journal of Machine Learning and Cybernetics*, 10(8):2057–2070.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Irazú Hernández Farıas and Paolo Rosso. 2016. Irony, sarcasm, and sentiment analysis. chapter 7. *Sentiment Analysis in Social Networks, F.A. Pozzi, E. Fersini, E. Messina, and B. Liu (Eds.), Elsevier Science and Technology*, pages 113–128.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708.

George A. Miller. 1995. Wordnet: A lexical database for english. *Communications of the ACM*, 38:39–41.

Saif M Mohammad and Peter D Turney. 2013. Nrc emotion lexicon. *National Research Council, Canada*.

Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125.

Antonio Reyes, Paolo Rosso, and Davide Buscaldi. 2012. From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering*, 74:1–12.

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. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM Sigkdd International Conference on Knowledge Discovery & Data Mining*, pages 849–857.

Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. 2018. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8697–8710.