Can a Pretrained Language Model Make Sense with Pretrained Neural Extractors? An Application to Multimodal Classification

Bhagyashree Gaikwad, Bhargav Kurma, Manasi Patwardhan, Shirish Karande and Niranjan Pedanekar

TCS Research

Abstract

Identifying the topics in online images such as memes and advertisements, and the sentiments evoked by them is a challenging task. It requires consideration of both visual and textual content in the images, and what they mean together. Both these modalities of information contain a cognitive component (what is going on in the image) and an affective component (what might be the emotional effect of the image). The main idea of this work is to extract such cognitive and affective concepts from visual and textual modalities using distinct pre-trained neural extractors and applying a language model (BERT) to reason over it for multimodal classification tasks. We benchmark the performance of our approach on the topic and sentiment classification tasks of the CVPR 2018 advertisement dataset and achieve state-of-the-art (SOTA) on topic classification with improvement in terms of mAP and F1-O score over current SOTA for sentiment classification. We also use it on the hate detection task of the Facebook hateful memes dataset and report a decent performance. We demonstrate that BERT can utilize textual inputs from different neural extractors in different formats, to obtain a performance improvement.

Keywords

Multi-modal Classification, Hateful Memes, Topic and Sentiment Detection, BERT

1. Introduction

Expressing hate to a large audience has gotten easier with the freedom of reach offered by social media platforms like Twitter, Facebook, and Instagram. Hence it is being increasingly important to understand, monitor, and filter such hate content. Memes are one such medium of expressing hate that incorporates both image and text data. An understanding of both modalities is required to detect the hate in them. A similar understanding is required to detect the topics and the sentiments inspired by advertisement (ad) images. This understanding can be used in ad placement based on its match with a given context, sensitive ad filtering, etc. The viewer's perception of such images is a function of its cognitive and affective components from both modalities.

De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, co-located with AAAI 2022. 2022 Vancouver, Canada

bhagyashree.gaikwad@tcs.com (B. Gaikwad); bhargav.kurma@tcs.com (B. Kurma); manasi.patwardhan@tcs.com (M. Patwardhan); shirish.karande@tcs.com (S. Karande); n.pedanekar@tcs.com (N. Pedanekar)

(a) (b)

Figure 1: Example of (a) Hateful meme image with scene-text 'on my way to run over w*men and minorities' and (b) Advertisement image with ground truth Sentiment label: Alert, Topic label: Safety, Alcohol abuse

In this work, we exploit pretrained neural networks to extract such cognitive and affective concepts relevant to the task in an advertisement or meme image and use a language model (BERT) [1] to reason over these concepts to perform the tasks of the topic and sentiment detection. We demonstrate our approach on Facebook Hateful Memes [2] and Ad understanding [3] datasets. These datasets involve images with embedded text (henceforth termed as scene-text). To address the memes classification and ads topic or sentiment classification tasks defined on these datasets consideration of information from both the textual and visual modalities is required. The memes dataset contains the 'benign confounders' [2] which makes it hard for models to rely on unimodal information. Figure 1a shows an example of a hateful meme. In this example, visual information about 'the dog driving a car' alone would not suffice to infer the hatefulness in the meme. A model also has to consider the scene-text 'on my way to run over w*men and minorities' along with the image. Similarly, to infer the topic and sentiment labels of the advertisement image in figure 1b, a model should consider both the scene-text 'What's the price of a bottle of Champagne? Don't drink and drive' and visual concepts such as 'car', 'broken alcohol bottle' symbolizing a 'car crash' for inferring the sentiment label 'Alert'. The same goes for the inference of topic labels 'Safety' and 'alcohol abuse'.

With this approach, we achieve SOTA results on the ads topic classification task and an improvement in mean average precision (mAP) and F1-O scores over the current SOTA on the ads sentiment classification task. On the Hate Memes dataset, our approach achieves similar performance to other approaches that use multi-modal transformer-based architectures. Our approach is flexible as one can add more neural extractors to improve performance.

2. Related work

There have been active efforts in recent years to collect and analyze memes datasets [2, 8, 9, 10, 11]. These works address binary (hateful or not) or multi-class (humor, sarcasm, offensive, motivational) classification tasks using concatenated scene-text and visual representations from pretrained models. The Facebook hateful memes dataset has emerged as a prominent dataset on multimodal hate detection. The dataset was released as a part of a challenge that took place in 2 phases. In phase 1, results were reported on a 'seen' test set which was made available

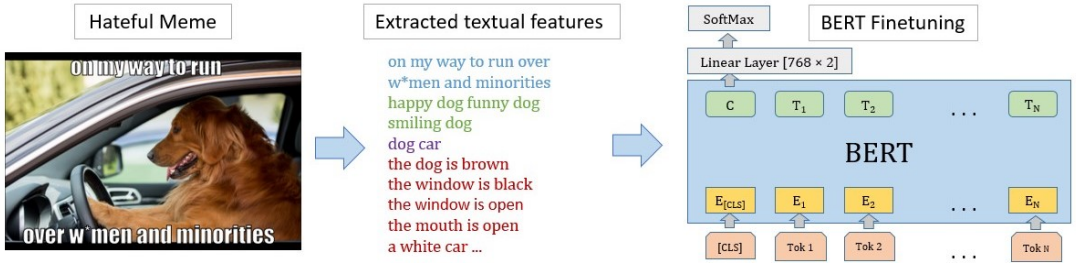

Figure 2: Proposed method of Memes classification based on image textual features. Each feature is shown in a different highlight viz. scene-text, adjective-noun pairs, MRCNN objects, and densecaps. Concatenated combinations of these features in the form of a single string are provided as an input to BERT.

Table 1
Neural Modules (CT: Cognitive Textual, CV: Cognitive Visual, AF: Affective Visual, AT: Affective Textual; VSO: Visual Sentiment Ontology [4]) *https://cloud.google.com/vision/docs/ocr, **https://deepai.org/machine-learning-model/densecap, §https://github.com/matterport/Mask_RCNN, †https://cloud.google.com/natural-language/docs/basics#sentiment_analysis, ′https://github.com/imatge-upc/affective-2017-musa2

Image Feature	Format	Neural Module	Input
Scene-TeXt (STX)	Text	Google Vision API*	CT
Densecaps [5] (DC)	Sentences	Densecap API**	CV
MRCNN Objects [6] (MR)	Label	MRCNN trained on MS COCO§	CV
Scene-text Valance Arousal (TVA)	Numeric	Google NLP API†	AT
Adjective-Noun Pairs [7] (ANP)	Phrases	ANP model trained on VSO′	AV

to the participants, and in phase 2 results were reported on an 'unseen' test set that wasn't made available [12]. A key point to note is that all of the top 5 winners [13, 14, 15, 16, 17] of phase 2 of this challenge utilized ensembles of one or more multimodal transformer based model architectures such as UNITER[18], VisualBERT[19], ERNIE-ViL[20], VL-BERT[21], etc. In contrast, we utilize a pretrained language model BERT for this multimodal task. The top-ranked model achieving 0.845 AUROC score on this dataset [13] utilizes Google Vision Web Entity Detection and FairFace classifier to capture the image's context and race, gender labels respectively that are relevant to the task. They also note that it is unrealistic to expect the pretrained transformer to learn to identify information such as race with just a few thousand samples. This resonates with our ideology of extracting relevant information from data using pretrained neural extractors and reasoning over the extracted information. Muennighoff [14] uses masked pretraining to adapt the models to the hateful memes dataset. Velioglu and Rose [15] show how the pretraining of the multimodal transformer model plays a crucial role in achieving good performance. VisualBERT[19], which they utilize for the task was originally pretrained on COCO image captions dataset[22]. However, they pretrain VisualBERT on Conceptual Captions[23] dataset and fine-tune it on the training dataset augmented with the Memotion dataset[24] for better results.

Understanding image advertisements also requires consideration of both visual and textual

Table 2
Hateful Memes results on 'seen' validation and test splits. (*Models of Kiela et al. [2], rest are the results of our experiments; UP: Unimodal Pre-training; MP: Multimodal Pre-training; GSTX: Groundtruth Scene-TeXt; ESTX: Extracted Scene-TeXt; +: Concatenation; NP: No Pre-training; Underlined & bold - best scores, underlined - second best.)

Model	Features	Validation		Test	
		Accuracy	AUROC	Accuracy	AUROC
Image Grid*	Object Features (O)	52.73	58.79	52.00	52.63
Image Region*	O	52.66	57.98	52.13	55.92
ResNet152	Image Features	57.00	59.52	53.90	56.20
Text BERT*	Ground-truth Scene-text (GSTX)	58.26	64.65	59.20	65.08
Visual BERT (UP)*	O+GSTX	65.01	74.14	66.67	74.42
Visual BERT (MP)*	O+GSTX	**65.93**	**74.14**	**69.47**	**75.44**
BERT	GSTX	58.40	65.06	59.20	66.45
BERT	Extracted Scene-text (ESTX)	56.40	63.89	58.10	65.31
BERT	Densecaps (DC)	52.60	54.13	51.70	54.38
BERT	MRCNN Objects (MR)	50.20	54.64	49.80	52.87
BERT	Adjective-Noun Pairs (ANPs)	50.00	51.72	50.00	51.99
BERT	ANP + DC	53.00	54.50	51.20	52.99
BERT	TVA + GSTX	56.60	63.36	58.40	66.43
BERT	GSTX + MR	61.00	69.51	60.70	68.67
BERT	GSTX + DC	62.60	70.72	63.70	71.08
BERT	GSTX + ANP + MR	61.80	70.80	66.00	72.14
BERT NP	GSTX + ANP + MR	54.40	58.62	54.10	60.49

modalities. Zhang et al. [25] address the topic and sentiment classification tasks on the ads dataset [3] in a multimodal-multitask setting. They use image and object features for the visual modality, BiLSTM features for the text modality and combine them with inter and intra attention. On the other hand, our approach takes advantage of BERT that can perform multiple levels of attention between and within modalities. Pilli et al. [26] addresses ads sentiment prediction task by exploiting the semantic relationship among sentiment labels and under-sampling noisy labels resulting in an improvement in mAP and F1-O score over [25]. On the CVPR ads challenge task, [27] uses Densecap captions [5] to convert visual input to text sentences to feed as an input to BERT. Our approach uses several neural modules to extract cognitive and affective concepts which can be in various formats as listed in table 1. Our results demonstrate that BERT can utilize the information presented in these different formats and achieve decent performance.

3. Datasets' Statistics and Metrics

The Facebook Hateful Memes dataset [2] contains 10000 memes annotated as hateful and not-hateful with equal distribution. 40% of the memes are constructed using a "benign confounder" text or image. These memes are hateful unimodally but not hateful multimodally. This makes it difficult for models to rely on just unimodal information. This data has 85%-5%-10% train-val-test splits. We report both area under the receiver operating characteristic curve (AUROC) [28] which is considered as the main metric in this dataset and also the accuracy score. The ad

Table 3
Ads topic and sentiment classification results. (ST: Single Task; MT: Multi-Task; NA: Not Applicable; OL: Original Labels; USL: Under-Sampled Labels; other abbreviations from table 2)

Model	Features	Topic Classification			Sentiment Classification		
		mAP	F1-C	F1-O	mAP	F1-C	F1-O
Zhang et al. [25] ST	Image, Object Features + Scene-text	0.290	0.214	0.482	0.284	0.192	0.439
Zhang et al. [25] MT	Image, Object Features + Scene-text	0.382	0.371	0.585	0.292	**0.216**	0.453
Pilli et al. [26] OL	Image Features	NA	NA	NA	0.327	-	0.466
Pilli et al. [26] USL	Image Features	NA	NA	NA	0.332	-	0.470
ResNet152	Image Features	0.3091	0.2324	0.4810	0.3459	0.1520	0.4126
BERT	Extracted Scene-text (ESTX)	0.5501	0.5363	0.7167	0.3577	0.1650	0.4429
BERT	Densecaps (DC)	0.2448	0.1502	0.4052	0.3321	0.1280	0.3948
BERT	MRCNN Objects (MR)	0.1618	0.0082	0.2486	0.3080	0.0897	0.3433
BERT	Adjective-Noun Pairs (ANPs)	0.1492	0.0534	0.2156	0.3117	0.0945	0.3494
BERT	ANP + DC	0.2354	0.1434	0.3846	0.3355	0.1366	0.3989
BERT	Text Valence-Arousal + ESTX	0.5558	0.5405	0.7178	0.3640	0.1780	0.4529
BERT	ESTX + ANP	0.5626	0.5355	0.7255	0.3702	0.1790	0.4598
BERT	ESTX + DC	**0.5694**	**0.5365**	**0.7298**	0.3790	0.1929	**0.4728**
BERT	ESTX + ANP + DC	0.5613	0.5222	0.7251	**0.3794**	0.1977	0.4663
BERT NP	ESTX + ANP + DC	-	-	-	0.3534	0.1544	0.4304
BERT NP	ESTX + DC	0.4524	0.3323	0.6351	-	-	-

understanding dataset [3, 29] contains 64,340 images annotated with 38 topic labels. Out of these, 30,275 are annotated with 30 sentiment labels. Each image can be tagged with multiple topic and sentiment labels making it a multi-label classification task. For our experiments, we consider these 30,275 images that have both topic and sentiment labels to benchmark against [25] and [26]. We consider a train-val-test split of 72%-18%-10%. Following the conventional settings [25, 30], we report the results in terms of micro-F1 or average overall F1 score (F1-O), macro-F1 or average class wise F1 score (F1-C), and mean average precision (mAP) metrics.

4. Approach

The tasks in hand may require both cognitive (Objects, their attributes, and relations) and affective (image valance arousal scores, object attributes such as happy, angry, calm, etc.) type of semantic information. We use a set of neural modules to extract this information we term as image features (table 1), from visual and textual modalities. These neural modules provide noisy outputs on our datasets as they are used in a zero-shot setting. We concatenate combinations of these features to form a single string and provide it as an input to BERT for a sentence classification task. For the task of binary classification of the memes dataset, we pass the [CLS] pin output to a fully connected layer with two output neurons followed by a softmax function. For all the experiments, we use a batch size of 16, learning rate of 2e-5, AdamW [31] optimizer, cross-entropy loss, and train the models for 5 epochs. In the case of classification on ads dataset, we pass the [CLS] pin output to a fully connected layer followed by a sigmoid function. We consider the prediction to be positive if the output is greater than 0.5.

Our approach is modular and efficient as one can use off-the-shelf neural extractors to extract information relevant to the task from both visual and textual modalities. A powerful language model is then used to reason over this information extracted in the form of text.

5. Results and Discussion

We report the results for both topic and sentiment classification tasks on the ads dataset in table 3. On the hateful memes dataset, we report the results on the 'seen' test and validation sets (table 2) for the results to be comparable with [2]. With our approach we achieve SOTA on ads topic classification and improvement in terms of mAP and F1-O score for sentiment classification. On the hateful memes dataset, we report comparable performance with approaches that use multimodal transformer based architectures.

5.1. Contribution of Individual Features

We perform experiments with individual outputs of neural modules (features) for each task to identify the contribution of each feature. For the memes dataset, we perform experiments with the scene-text provided in the dataset (GSTX) to benchmark against [2], however, we can see that the model trained on the extracted scene-text performs nearly as good as the model trained on the ground-truth scene-text. We observe that for all the tasks on both datasets, in terms of individual feature contribution, scene-text contributes the highest followed by densecaps, and then, MRCNN objects and ANPs.

5.2. Feature Sequence Construction

To avoid important information getting lost due to the truncation beyond the maximum token length (512) allowed by BERT architecture, we decide the feature order to be fed to BERT as follows. Scene-text being the most contributing feature, it is the first to appear in order. We consider the top 5 ANP predictions, which are 10 words. On average we get 4 MRCNN objects per sample, which are 4 words. Length of concatenated densecaps may go beyond the allowed max sequence length. To avoid truncation of ANPs and/or MRCNN Objects, we append them after the scene-text and before densecaps with descending order of their confidence. We concatenate text valence-arousal scores at the beginning so that the positions of these values remain the same across examples. This allows BERT to utilize the positional encodings to treat this as a distinct signal. We follow the same order for both datasets.

5.3. Contribution of Feature Combinations

We try several combinations of text valence-arousal score, scene-text, ANPs, MRCNN Objects, and densecaps to find out which combination gives the best results on the validation split of each dataset. With scene-text+densecaps as input, we achieve SOTA on the ads dataset for the topic classification task and also on sentiment classification task in terms of F1-O score. After adding ANPs as an additional feature to this combination, we see an improvement in mAP and F1-C scores of sentiment classification. A multi-task setting as in [25] gives us inferior results compared to the single task. The feature combination, scene-text+densecaps also achieves a decent performance on the memes dataset.

5.4. Contribution of Each Modality Features

Within features from just the visual modality (ANPs, densecaps, MRCNN Objects) we find that the combination of ANPs+densecaps gave the best results for both datasets. Within features from just the text modality, text valence-arousal+extracted scene-text gave us the best results for the ads dataset and the ground-truth scene-text gave the best results on the memes dataset. The difference in the performances of models on just the text modality and just the visual modality suggests that the text modality might be contributing more to the tasks compared to the visual modality. For the ads dataset, adding text valence-arousal scores and ANPs individually to scene-text gave an improvement over just scene-text. This shows that we can utilize even numeric or word phrase information from neural modules as input to BERT to get better performance. We fine-tune ResNet152 [32] pretrained on ImageNet to have a benchmark on end-to-end approaches that operate directly on the image features. We observe that this approach performs moderately well on the tasks but the results are always inferior to our approaches that utilize the scene-text. However, when compared with our approaches that utilize just the visual input features (densecaps, ANPs, MRCNN Objects), it performs better indicating some loss of information during feature extraction from the visual input.

5.5. Role of BERT's pre-training

To understand the role of BERT's pre-training, we perform experiments on tasks of ads and memes dataset by using BERT without the pre-trained weights [BERT NP]. For this experiment, we use features that gave us the best results for each task. The significant reduction in scores on both tasks indicates that BERT's pre-training plays a prominent role in obtaining the performance.

Conclusion and Future Work

We demonstrate a powerful approach that utilizes the language model BERT to achieve a good performance on distinct multimodal tasks. Extracting semantic information from the visual input in the form of text by using neural modules and using BERT to reason over this information gives us comparable results with approaches that operate directly on object or image features despite some loss of information. We show that a language model like BERT can utilize inputs in the form of numbers, categorical labels, or phrases to get an improvement. In the future, we would like to validate our approach with more language models and neural modules.

References

[1] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[2] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, D. Testuggine, The hateful memes challenge: Detecting hate speech in multimodal memes, in: Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 2611–2624.

[3] Z. Hussain, M. Zhang, X. Zhang, K. Ye, C. Thomas, Z. Agha, N. Ong, A. Kovashka, Automatic understanding of image and video advertisements, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1705–1715.

[4] D. Borth, R. Ji, T. Chen, T. Breuel, S.-F. Chang, Large-scale visual sentiment ontology and detectors using adjective noun pairs, in: Proceedings of the 21st ACM international conference on Multimedia, 2013, pp. 223–232.

[5] J. Johnson, A. Karpathy, L. Fei-Fei, Densecap: Fully convolutional localization networks for dense captioning, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4565–4574.

[6] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.

[7] D. Fernandez, A. Woodward, V. Campos, X. Giró-i Nieto, B. Jou, S.-F. Chang, More cat than cute? interpretable prediction of adjective-noun pairs, in: Proceedings of the Workshop on Multimodal Understanding of Social, Affective and Subjective Attributes, 2017, pp. 61–69.

[8] Y. Du, M. A. Masood, K. Joseph, Understanding visual memes: An empirical analysis of text superimposed on memes shared on twitter, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 14, 2020, pp. 153–164.

[9] S. Suryawanshi, B. R. Chakravarthi, M. Arcan, P. Buitelaar, Multimodal meme dataset (multioff) for identifying offensive content in image and text, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, 2020, pp. 32–41.

[10] S. D. Das, S. Mandal, Team neuro at semeval-2020 task 8: Multi-modal fine grain emotion classification of memes using multitask learning, arXiv preprint arXiv:2005.10915 (2020).

[11] R. R. Pranesh, A. Shekhar, Memesem: A multi-modal framework for sentimental analysis of meme via transfer learning, in: 4th Lifelong Learning Workshop of ICML 2020, 2020.

[12] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, C. A. Fitzpatrick, P. Bull, G. Lipstein, T. Nelli, R. Zhu, N. Muennighoff, R. Velioglu, J. Rose, P. Lippe, N. Holla, S. Chandra, S. Rajamanickam, G. Antoniou, E. Shutova, H. Yannakoudakis, V. Sandulescu, U. Ozertem, P. Pantel, L. Specia, D. Parikh, The hateful memes challenge: Competition report, in: Proceedings of the NeurIPS 2020 Competition and Demonstration Track, volume 133 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 344–360.

[13] R. Zhu, Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution, arXiv preprint arXiv:2012.08290 (2020).

[14] N. Muennighoff, Vilio: State-of-the-art visio-linguistic models applied to hateful memes, arXiv preprint arXiv:2012.07788 (2020).

[15] R. Velioglu, J. Rose, Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge, arXiv preprint arXiv:2012.12975 (2020).

[16] P. Lippe, N. Holla, S. Chandra, S. Rajamanickam, G. Antoniou, E. Shutova, H. Yannakoudakis, A multimodal framework for the detection of hateful memes, arXiv preprint arXiv:2012.12871 (2020).

[17] V. Sandulescu, Detecting hateful memes using a multimodal deep ensemble, arXiv preprint arXiv:2012.13235 (2020).

[18] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, J. Liu, Uniter: Universal image-text representation learning, in: Computer Vision – ECCV 2020, Springer

International Publishing, Cham, 2020, pp. 104–120.
[19] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, K.-W. Chang, Visualbert: A simple and performant baseline for vision and language, in: Arxiv, 2019.
[20] F. Yu, J. Tang, W. Yin, Y. Sun, H. Tian, H. Wu, H. Wang, Ernie-vil: Knowledge enhanced vision-language representations through scene graphs, Proceedings of the AAAI Conference on Artificial Intelligence 35 (2021) 3208–3216.
[21] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, J. Dai, Vl-bert: Pre-training of generic visual-linguistic representations, in: International Conference on Learning Representations, 2020.
[22] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, C. L. Zitnick, Microsoft coco captions: Data collection and evaluation server, 2015. arXiv:1504.00325.
[23] P. Sharma, N. Ding, S. Goodman, R. Soricut, Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 2556–2565. doi:10.18653/v1/P18-1238.
[24] C. Sharma, D. Bhageria, W. Scott, S. PYKL, A. Das, T. Chakraborty, V. Pulabaigari, B. Gamback, Semeval-2020 task 8: Memotion analysis – the visuo-lingual metaphor!, 2020. arXiv:2008.03781.
[25] H. Zhang, Y. Luo, Q. Ai, Y. Wen, H. Hu, Look, Read and Feel: Benchmarking Ads Understanding with Multimodal Multitask Learning, Association for Computing Machinery, New York, NY, USA, 2020, p. 430–438.
[26] S. Pilli, M. Patwardhan, N. Pedanekar, S. Karande, Predicting sentiments in image advertisements using semantic relations among sentiment labels, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 408–409.
[27] K. Kalra, B. Kurma, S. V. Sreelatha, M. Patwardhan, S. Karande, Understanding advertisements with bert, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 7542–7547.
[28] A. P. Bradley, The use of the area under the roc curve in the evaluation of machine learning algorithms, Pattern recognition 30 (1997) 1145–1159.
[29] K. Yi, J. Wu, C. Gan, A. Torralba, P. Kohli, J. Tenenbaum, Neural-symbolic vqa: Disentangling reasoning from vision and language understanding, in: Advances in neural information processing systems, 2018, pp. 1031–1042.
[30] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, W. Xu, Cnn-rnn: A unified framework for multi-label image classification, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2285–2294.
[31] I. Loshchilov, F. Hutter, Fixing weight decay regularization in adam (2018).
[32] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.