Explaining Emotional Attitude Through the Task of Image-captioning

Oleg Bisikalo [1] , Volodymyr Kovenko [1] Ilona Bogach [1] and Olha Chorna [2]

[1] Vinnytsia National Technical University, Khmelnytsky highway 95, Vinnytsya, 21021, Ukraine
[2] Kremenchuk Mykhailo Ostrohradskyi National University, Pershotravneva Street, 20, Kremenchuk, 39600, Ukraine

Abstract

Deep learning algorithms trained on huge datasets containing visual and textual information, have shown to learn useful features for other downstream tasks. This implies that such models understand the data on different levels of hierarchies. In this paper we study the ability of SOTA (state-of-the-art) models for both texts and images to understand the emotional attitude caused by a situation. For this purpose we gathered a small size dataset based on IMDB-WIKI one and annotated it specifically for the task. In order to investigate the ability of pretrained models to understand the data, the KNN clustering procedure over representations of text and images is utilized in parallel. It's shown that although used models are not capable of understanding the task at hand, a transfer learning procedure based on them helps to improve results on such tasks as image-captioning and sentiment analysis. We then frame our problem as the task of image captioning and experiment with different architectures and approaches to training. Finally, we show that adding additional biometric features such as probabilities of emotions and gender probabilities improves the results and leads to better understanding of emotional attitude.

Keywords

Deep learning algorithms; Emotional attitude; SOTA models; Image-captioning; NLP; Transfer-learning

1. Introduction

Recent development of hardware and access to big datasets allowed researchers to train sophisticated deep learning based algorithms, which suppressed many other approaches. The deep learning revolution affected many fields, whereas the most interesting results were obtained in the field of NLP (natural language processing) [1] and CV (computer vision) [2]. It was shown that SOTA models trained on big datasets (ImageNet [3], Google News) tend to learn useful features that can be used for other downstream tasks [4]. Building on that idea we study how well such models understand the emotional attitude and its cause implicitly or explicitly introduced by visual and textual data. Understanding the emotional attitude and explaining it is a pretty hard task even for a human, as the solution requires the exact understanding of cause and consequence that are affected by the environment and biometric features. For the purpose of experiments a new small-size dataset containing image-text pairs called "EmoAtCap" is collected. The overall contribution of our work is summarized below:

1. A small-size dataset "EmoAtCap" which is based on IMDB-WIKI one, that can be used for image-captioning and sentiment analysis. It is publicly available [5] for facilitating future research in this domain.

COLINS-2022: 6th International Conference on Computational Linguistics and Intelligent Systems, May 12–13, 2022, Gliwice, Poland.

EMAIL: obisikalo@gmail.com (O.Bisikalo); urumipainblackreaper@gmail.com (V.Kovenko); ilona.bogach@gmail.com (I.Bogach), diolgan@gmail.com (O.Chorna)
ORCID: 0000-0002-7607-1943 (O.Bisikalo); 0000-0003-3825-1115 (V.Kovenko); 0000-0001-9398-8529 (I.Bogach); 0000-0002-7113-1572 (O.Chorna);

2. A set of experiments on the tasks of image-captioning and sentiment analysis, based on features extracted from highlighted models. It’s also shown that adding biometric features as gender and emotions distribution improves the performance of image-captioning models.

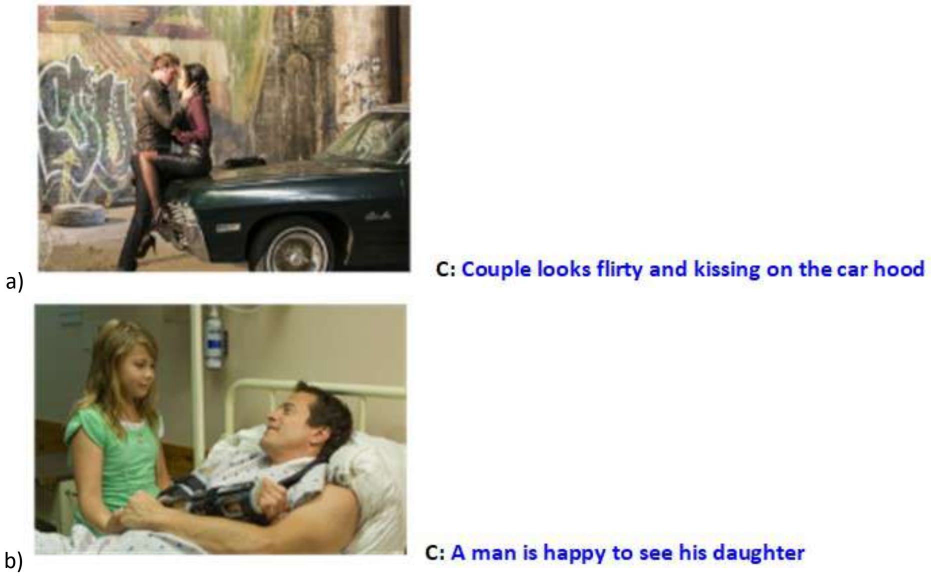
The training procedure was conducted using tensorflow [6] and pytorch [7].

2. Data collection

The actual data needed to include both images and their captions. As the main intent was to capture the emotional attitude, the images would have to contain people and explicit or implicit information about the cause of their emotional state. The captions should have contained an exhaustive unbiased description of the situation. Based on highlighted requirements, the first idea was to make a dataset from the subset of existing image-captioning datasets.

Image-captioning is the process of generating textual description of an image. The task implies that the relevant dataset consists of image-text pairs. One of the most popular datasets for the discussed task is COCO [8], which consists of 330K images. We used only the subset of dataset related to image-captioning, mainly the 2014 train split, which consisted of 29766 images along with 5 captions per each image. As it would be pretty hard and cumbersome to filter out images manually, a YoloV3 [9] object-detection algorithm trained on the discussed dataset was used. Only images that contained objects of class “person” were left. As a result, the COCO dataset was shrunk to 3731 images. However, filtered images and captions only contained the actual plot of the image without any emotional attitude. The other analyzed dataset was a VizWiz [10] one. VizWiz is the first goal-oriented VQA (visual question answering) dataset arising from a natural VQA setting, which consists of over 31,000 visual questions originating from blind people. Needed data subset was found by filtering the captions using people related words. As the final data was of a poor quality, this variant was declined. The last image-text data we experimented with was SentiCap [11] one. SentiCap consists of 2360 images containing sentiments. After filtering the dataset in the same way as it was done for VizWiz one, we arrived with only 830 samples, which was not enough for our task.

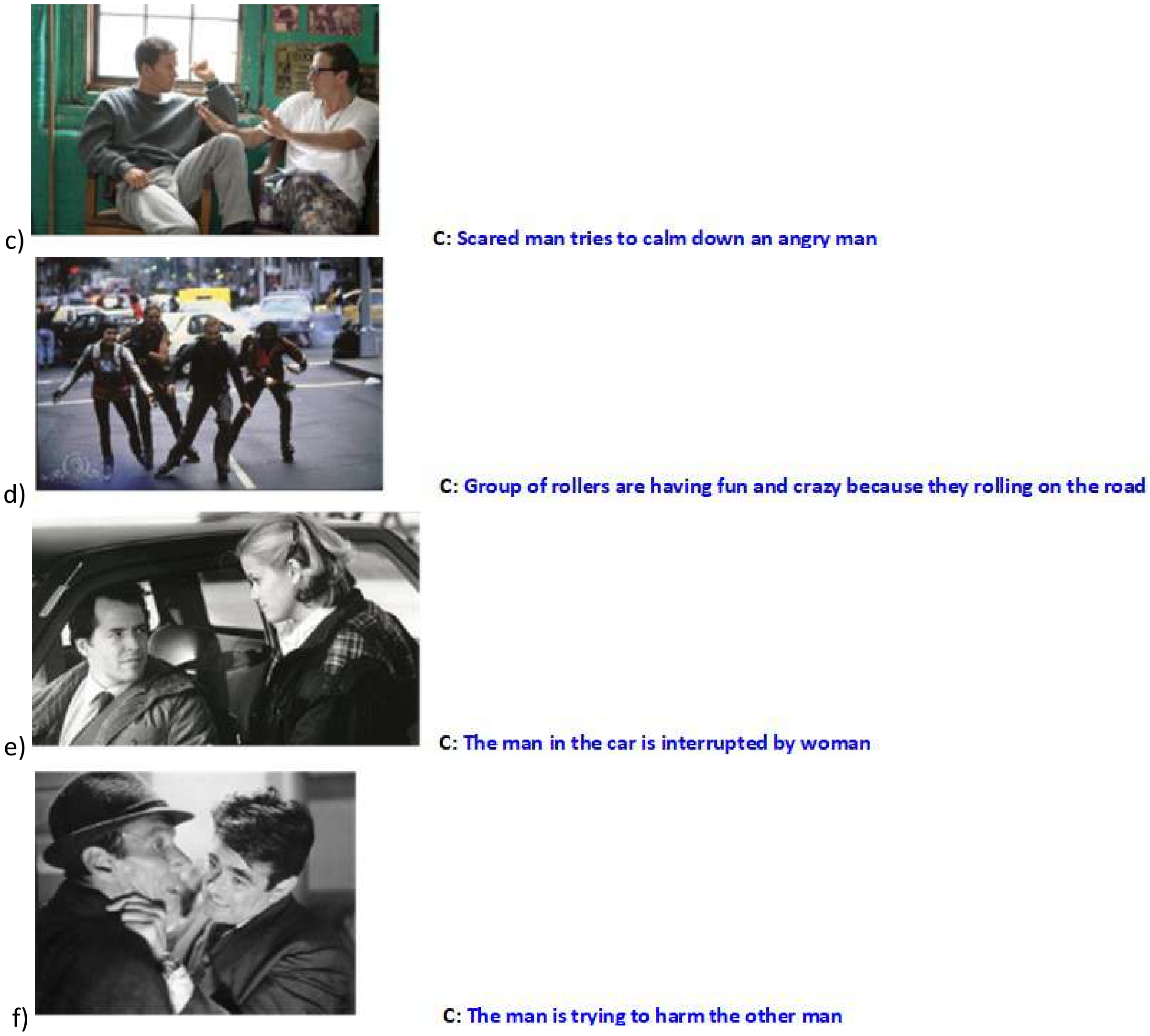
The other variant was to gather a dataset from the very beginning and annotate it. The images were taken from the IMDB-WIKI [12] dataset for age and gender detection. Each image was annotated with a description of the emotional attitude of the person or people on it. As a result we arrived with the dataset of 3840 image-text pairs, where each image was resized to 224x224 pixels (Fig. 1).



a) C: Couple looks flirty and kissing on the car hood

b) C: A man is happy to see his daughter

Figure 1 (a, b): Dataset examples with corresponding captions



C: Scared man tries to calm down an angry man

C: Group of rollers are having fun and crazy because they rolling on the road

C: The man in the car is interrupted by woman

C: The man is trying to harm the other man

Figure 1 (c - f): Dataset examples with corresponding captions

In order to categorize the dataset, sentiments related to captions were added using Vader [13], which is a rule based model for sentiment analysis. Then the sentiments were checked by humans one more time to produce more meaningful ones. As the result of the analysis, the data appeared to be imbalanced in terms of the new category (Fig. 2).

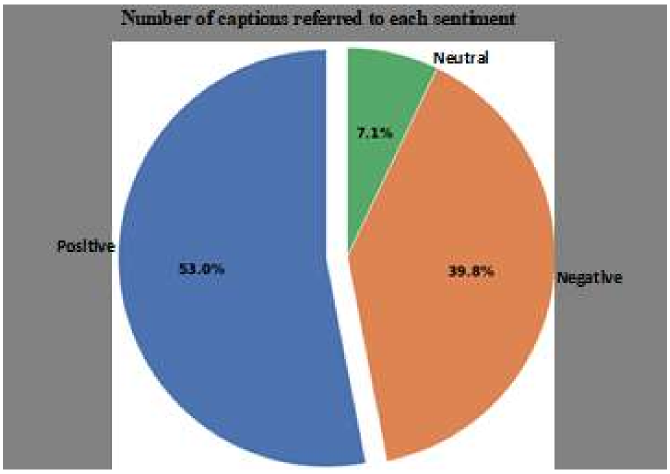


Figure 2: Distribution of caption sentiments in the dataset

New sentiment category was used for analysis of clustering and for solving the task of sentiment analysis given the captions.

3. Pretrained models overview

In order to analyze the ability of pretrained models to understand such difficult information as emotional attitude, recent SOTA models trained on big datasets of textual and visual information were chosen.

3.1. ResNet

ResNet, introduced by Kaiming He et.al, is a deep convolutional architecture, which suppressed previous results on Imagenet benchmark and showed to be pretty successful for object detection by obtaining a 28% relative improvement on the COCO object detection dataset. Main advantage of such architecture is the addition of residual connections that help to fight the problem of vanishing gradients, which is typical for deep neural networks. This advantage gave a possibility to train a very deep network, each layer of which learned different useful features. In our work ResNet152V2 pretrained on the Imagenet dataset was used. We also experimented with ResNet50 trained on FER [14] dataset.

3.2. EfficientNet

EfficientNet, introduced by Tan et.al, is a deep convolutional neural network architecture and scaling method that uniformly scales all dimensions of depth/width/resolution using a compound coefficient. It achieves state-of-the-art 84.3% top-1 accuracy on ImageNet and transfers well to other tasks, reaching state-of-the-art accuracy on CIFAR-100 [15] (91.7%), Flowers [16] (98.8%), and 3 other transfer learning datasets. In our work EfficientNet trained on age-gender IMDB-WIKI dataset was used.

3.3. Word2Vec

Word2Vec, introduced by Mikolov et.al [17] is a neural network based approach to learning word embeddings. The approach gives a possibility to use two methods of learning: CBOW and skip-gramm. During the CBOW approach, the model is asked to predict the current word given the context, whereas skip-gram one tries to predict words within a certain range before and after the current word. As a result of such training, the model learns meaningful word vectors that are often used for transfer learning. Word2Vec embeddings pretrained on Google News with the vectors' dimensionality of 300 were used in the paper.

The exact setup of experiments and description of layers using which the data representation was derived along with experimental results is discussed further in the paper

4. Experiments

4.1 Image-captioning

Image understanding is the process of interpreting regions/objects to figure out what's happening in the image. This may include figuring out what the objects are, their spatial relationship to each other, etc [18]. This statement implies that one of the definitions of scene understanding is a capability of describing its context. Thus, we theorize that a model which can describe the emotional attitude based on image is capable of understanding it. The task of describing the image is known as image-captioning, and gained a huge popularity with the development of deep neural networks [19]. Though there are many different approaches to the task [20], we exploit only the encoder-decoder architecture, where encoder's goal is to encode the representation of the image into the feature vector and decoder's one is to generate the captions based on this information. The theoretical foundations of constructing text messages / captions by modeling combinations of significant words are considered in [21]. For the role of encoder, a convolutional neural network is often exploited, whereas for the role of decoder - recurrent one. In our work the research is done due to different encoder-decoder architectures used to solve the task of image-captioning.

As it was stated by Kovenko et.al [22], by solving the problem of data reconstruction, autoencoders tend to learn low-level features, which are useful for transfer learning. Based on this idea we train the deep convolutional autoencoder on our dataset and use latent code produced by the encoder part for encoding images in image-captioning task. Also the experiments include the output of 4th block of ResNet, along with the logits of ResNet as the encoders. In order to compare this transfer learning approaches, we also experiment with custom not pretrained convolutional encoder.

The decoder part is represented by the embedding layer and LSTM (Long-short-term-memory) [23] network. LSTM is capable of learning long-time dependencies, which is especially useful when working with sequential data. As the embedding layer, for all the experiments, Word2Vec was used. For all the approaches, layer normalization [24] after LSTM was used. As it was stated by Xu et.al [25], the attention mechanism applied to image-captioning tasks can greatly improve results. Nezami et.al [26] showed that usage of additional features of emotions helps to improve results on the image-captioning datasets that include emotional aspects. Based on these ideas, we experimented with using attention and conditioning LSTM on additional features. Different from Nezami's approach, gender features were also used and the emotional ones were encoded as probability distribution. Specifically, YoloV3 is used to extract face regions from the images and EfficientNet trained on Age-Gender dataset along with ResNet trained on FER one are used to predict gender and emotions.

Gender features are produced using predicted probabilities for each face presented on the image (formula 1).

$$S_g = \frac{S_g}{\sum_g^G S_g}, \quad Sg = \sum_{i=1}^{N} 1_{P_i}^{g}, \quad P_i = argmax(pred_i) \qquad (1)$$

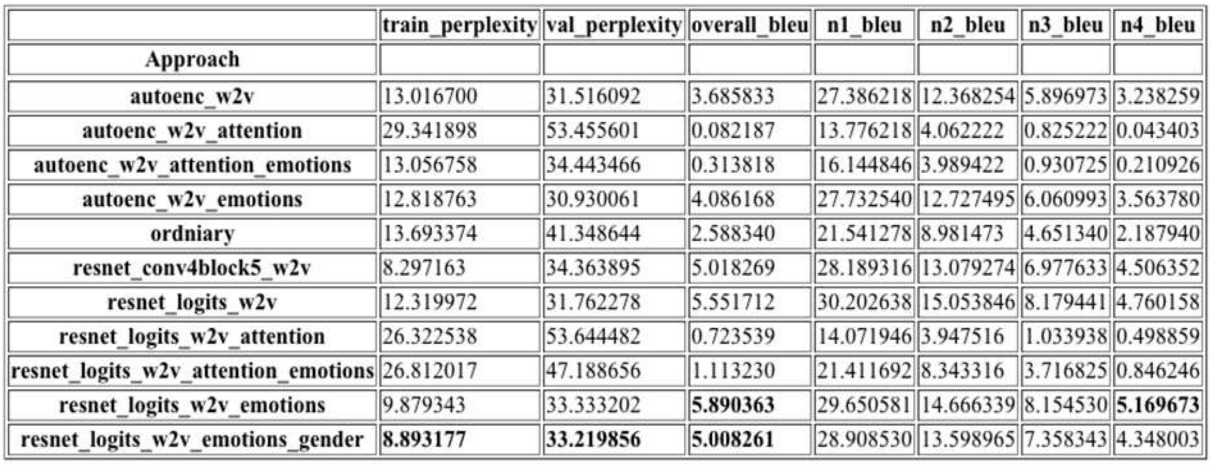
where G - number of unique genders, g - gender, S^{RxG} - normalized vector with gender probabilities, N - number of faces presented on the image, $1_{P_i}^{g}$ - identifier of Pibeing equal to specific g, P_i - result of an argmax operation over prediction probability vector for specific face i.

Emotional features are produced as normalized probability distribution of the sum of probability vectors for each face presented on the image (formula 2).

$$E = \sum_{i}^{N} pred_i \ , \quad E \ = \frac{E}{\sum_j^M E_j} \qquad (2)$$

where E^{RxM} - vector of averaged emotion probabilities, N - number of faces presented on the image, pred$_i$ - prediction probability vector for specific face i, M - number of unique genders.

The data was splitted in the same way as for sentiment analysis. The approaches were validated based on the test set performance using beam search technique with the beam size of 5. BLEU score along with perplexity were used as the main metrics. For all the experiments RMSprop optimizer was used, with the initial learning rate of 0.0001. In order not to overfit, the loss reduction technique was used. If there was no improvement in validation perplexity for two epochs, the loss was reduced by the factor 10. All the models were trained with a batch size of 64 for 30 epochs (Fig. 3).



	train_perplexity	val_perplexity	overall_bleu	n1_bleu	n2_bleu	n3_bleu	n4_bleu
Approach							
autoenc_w2v	13.016700	31.516092	3.685833	27.386218	12.368254	5.896973	3.238259
autoenc_w2v_attention	29.341898	53.455601	0.082187	13.776218	4.062222	0.825222	0.043403
autoenc_w2v_attention_emotions	13.056758	34.443466	0.313818	16.144846	3.989422	0.930725	0.210926
autoenc_w2v_emotions	12.818763	30.930061	4.086168	27.732540	12.727495	6.060993	3.563780
ordniary	13.693374	41.348644	2.588340	21.541278	8.981473	4.651340	2.187940
resnet_conv4block5_w2v	8.297163	34.363895	5.018269	28.189316	13.079274	6.977633	4.506352
resnet_logits_w2v	12.319972	31.762278	5.551712	30.202638	15.053846	8.179441	4.760158
resnet_logits_w2v_attention	26.322538	53.644482	0.723539	14.071946	3.947516	1.033938	0.498859
resnet_logits_w2v_attention_emotions	26.812017	47.188656	1.113230	21.411692	8.343316	3.716825	0.846246
resnet_logits_w2v_emotions	9.879343	33.333202	**5.890363**	29.650581	14.666339	8.154530	**5.169673**
resnet_logits_w2v_emotions_gender	**8.893177**	**33.219856**	**5.008261**	28.908530	13.598965	7.358343	4.348003

Figure 3. Comparison of image-captioning models and approaches. For train and validation perplexity, the values are shown for the last epoch of training

Analyzing the results it's obvious that the transfer-learning procedure gives better results than training from scratch (ordinary) w.r.t BLEU on a test set. It's also clear that ResNet representation tends to give better results than autoencoder's one, possibly because of a deeper architecture and better learned features. Attention didn't work well for all the approaches, probably because of the low number of samples in the dataset and small amount of epochs. So far an approach that utilized logits output of ResNet for encoder part of the network along with Word2Vec embeddings and additional features of emotions (resnet_logits_w2v_emotions) gave the best results on test data w.r.t to averaged BLEU score. The other appomodelrach, which is also worth paying attention to is the one, which incorporates both emotions and gender features. Despite resnet_logits_w2v_emotions_gender didn't achieve the best performance on test BLEU, it reached the best balanced performance on all the data splits, and thus was chosen as the best one. The architecture of the overall prediction pipeline is shown in Fig. 4.

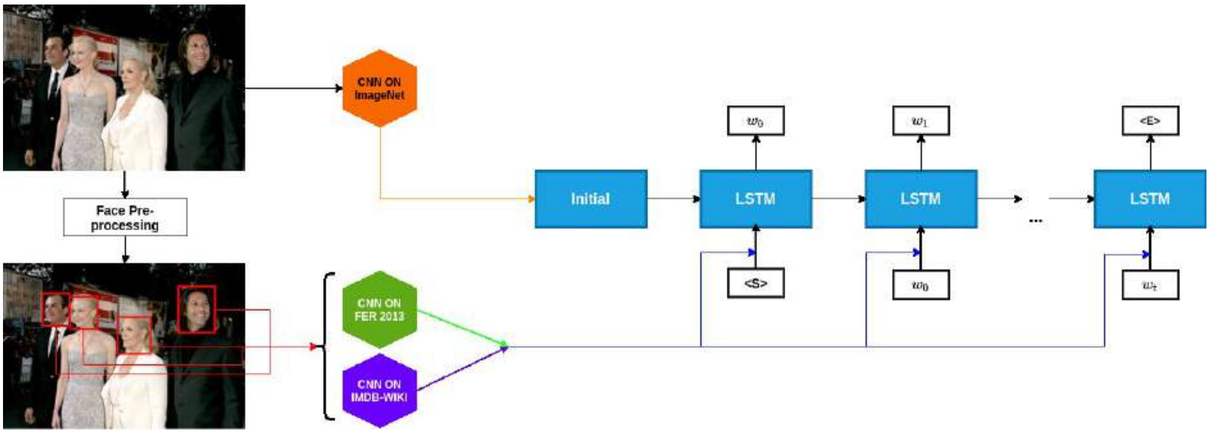
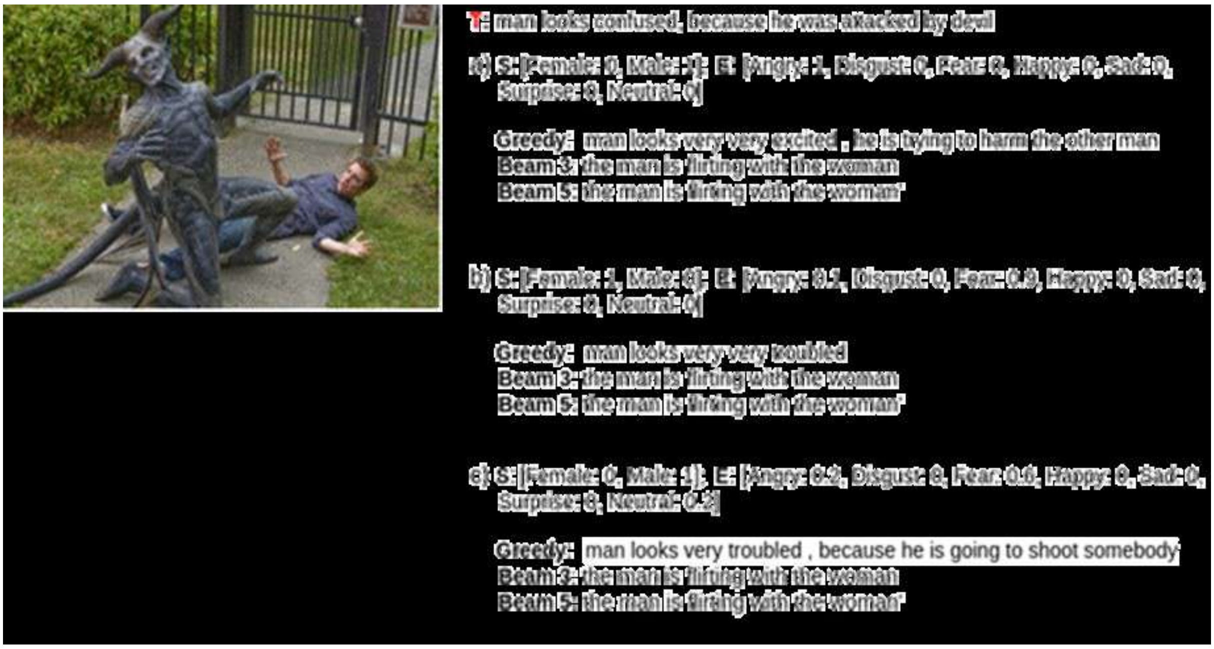


Figure 4. Architecture of the pipeline of resnet_logits_w2v_emotions_gender approach

As it can be seen from Fig. 4, the overall pipeline is dependent on the face pre-processing step along with the detection of emotions and gender. Obviously, if the performance of highlighted steps is poor, the final output will be at least biased. The example of such bias is represented in Fig. 5.



T: man looks confused, because he was attacked by devil

a) S: [Female: 0, Male: 1]; E: [Angry: 1, Disgust: 0, Fear: 0, Happy: 0, Sad: 0, Surprise: 0, Neutral: 0]

Greedy: man looks very very excited , he is trying to harm the other man
Beam 3: the man is 'flirting with the woman
Beam 5: the man is flirting with the woman'

b) S: [Female: 1, Male: 0]; E: [Angry: 0.1, Disgust: 0, Fear: 0.9, Happy: 0, Sad: 0, Surprise: 0, Neutral: 0]

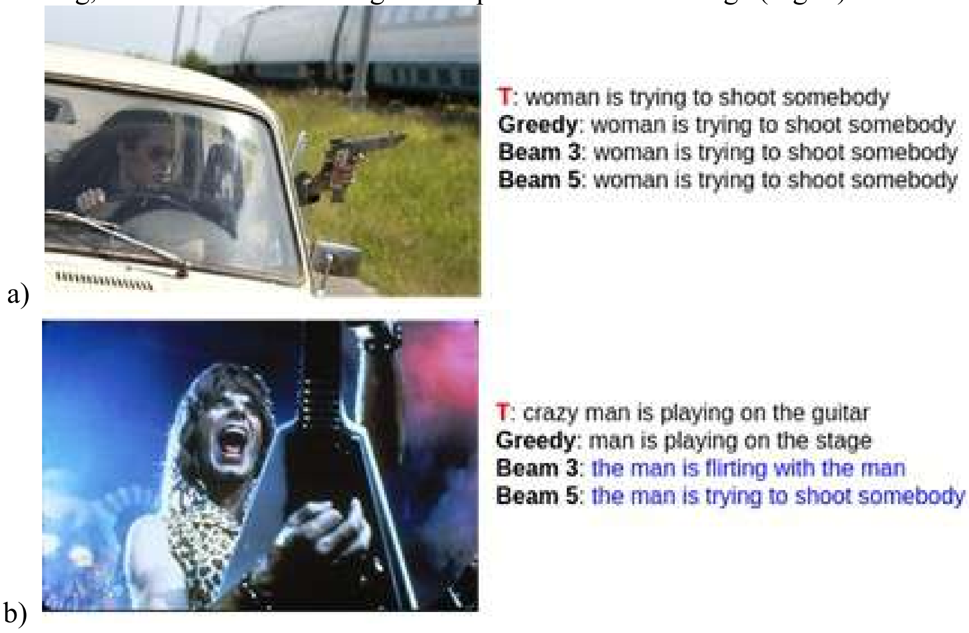
Greedy: man looks very very troubled
Beam 3: the man is 'flirting with the woman
Beam 5: the man is flirting with the woman'

c) S: [Female: 0, Male: 1]; E: [Angry: 0.2, Disgust: 0, Fear: 0.6, Happy: 0, Sad: 0, Surprise: 0, Neutral: 0.2]

Greedy: man looks very troubled , because he is going to shoot somebody
Beam 3: the man is 'flirting with the woman
Beam 5: the man is flirting with the woman'

Figure 5. Example of bias of additional features w.r.t image-captioning process. S - vector of gender features, E - vector of emotions features. T - true caption, greedy - result of greedy decoding, beam - result of beam search decoding. Changing additional features, changes the generation of captions using greedy decoding strategy.

During error analysis, it was found that the model suffers from slight overfitting on most frequent words and phrases (like "man is flirting with a woman" presented in Fig. 5), which is a problem caused by a small diversity of the dataset. Despite the fact that the collected data is a noisy one, as each image was annotated by a different expert, which is not very suitable for the task of image-captioning, the model succeeds to give adequate results on average (Fig. 6).



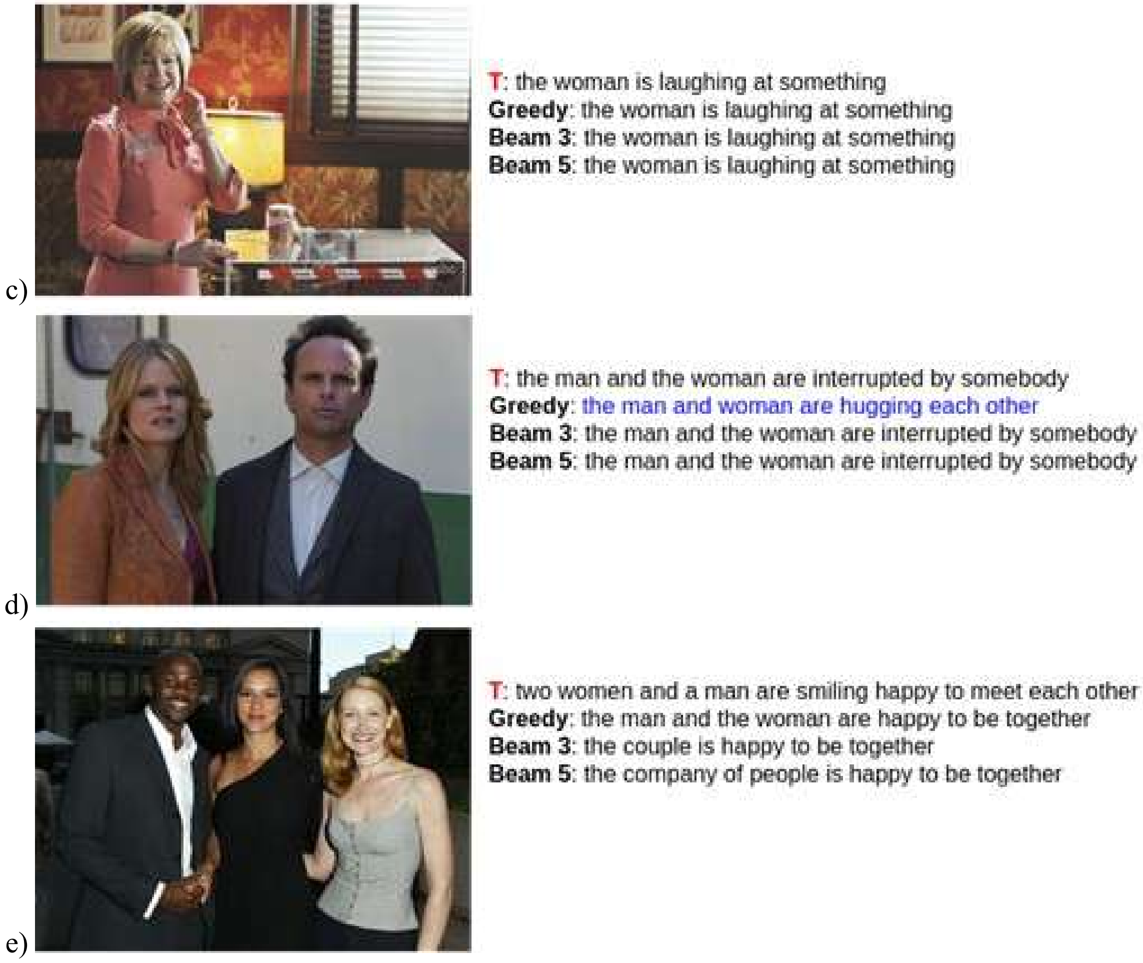
T: woman is trying to shoot somebody
Greedy: woman is trying to shoot somebody
Beam 3: woman is trying to shoot somebody
Beam 5: woman is trying to shoot somebody

a)

T: crazy man is playing on the guitar
Greedy: man is playing on the stage
Beam 3: the man is flirting with the man
Beam 5: the man is trying to shoot somebody

b)

Figure 6 (a, b). Example of generated captions. T - true caption, greedy - result of greedy decoding, beam - result of beam search decoding. Captions which are fully inappropriate are marked with bleu.



T: the woman is laughing at something
Greedy: the woman is laughing at something
Beam 3: the woman is laughing at something
Beam 5: the woman is laughing at something

T: the man and the woman are interrupted by somebody
Greedy: the man and woman are hugging each other
Beam 3: the man and the woman are interrupted by somebody
Beam 5: the man and the woman are interrupted by somebody

T: two women and a man are smiling happy to meet each other
Greedy: the man and the woman are happy to be together
Beam 3: the couple is happy to be together
Beam 5: the company of people is happy to be together

Figure 6 (c – e). Example of generated captions. T - true caption, greedy - result of greedy decoding, beam - result of beam search decoding. Captions which are fully inappropriate are marked with bleu.

It's important to note that the longer training would probably give better results.

5. Conclusion and further work

In this paper we analyzed the ability of deep learning models to understand the emotional attitude driven by the situation. For this purpose, a new dataset with image-text pairs was presented. In result of pretrained SOTA models analysis, it was concluded that some of them can be used in the process of transfer-learning. Through the experiments it was shown that the dataset can be used to solve the problem of sentiment analysis. It was then theorized that the problem of understanding the emotional attitude, can be transferred to the task of image-captioning. Empirical results have shown that addition of emotional and gender features along with transfer-learning based on ResNet network and Word2Vec embeddings improve the overall captioning performance. Our approach gives pleasant results on average, confirming that deep learning models are able to understand emotional attitude if they are trained to. It's important to note that such an approach has many downsides, as it's dependent on the performance of three additional models for face, emotions and gender detection. The other problem that was faced is the noisy nature of the dataset and small variation of phrases in it. In future work it's planned to gather a bigger dataset, label each image with 5 captions and fix current problems.

6. Acknowledgements

We would like to thank Oleksii Abdullaiev, Dmytro Tarasovskyi and Dmytro Maliovanyi for their contribution in terms of the dataset creation.

7. References

[1] Elizabeth D. Liddy. Natural Language Processing. In Encyclopaedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc. https://surface.syr.edu/cgi/viewcontent.cgi?article=1043&context=istpub (accessed 12 December 2021).

[2] IBM. What is Computer Vision? https://www.ibm.com/topics/computer-vision (accessed 12 December 2021).

[3] Deng, Jia. et al., 2009. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255.

[4] Yosinski, Jason, et al. How transferable are features in deep neural networks?. arXiv preprint arXiv:1411.1792, 2014.

[5] Kovenko, Volodymyr; Abdullaiev, Oleksii; Maliovanyi, Dmytro; Tarasovskyi, Dmytro; Bogach, Ilona; Bisikalo, Oleh (2021), "EmoAtCap : Emotional attitude captioning dataset", Mendeley Data, V5, doi: 10.17632/dym6p2pvbt.

[6] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... & Zheng, X. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467.

[7] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 32, 8026-8037.

[8] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014, September). Microsoft coco: Common objects in context. In European conference on computer vision (pp. 740-755). Springer, Cham.

[9] Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767.

[10] Gurari, D., Li, Q., Stangl, A. J., Guo, A., Lin, C., Grauman, K., ... & Bigham, J. P. (2018). Vizwiz grand challenge: Answering visual questions from blind people. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3608-3617).

[11] Mathews, A., Xie, L., & He, X. (2016, March). Senticap: Generating image descriptions with sentiments. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 30, No. 1).

[12] Rothe, R., Timofte, R., & Van Gool, L. (2015). Dex: Deep expectation of apparent age from a single image. In Proceedings of the IEEE international conference on computer vision workshops (pp. 10-15), doi: 10.1109/ICCVW.2015.41.

[13] Hutto, C., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 8, No. 1).

[14] Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., ... & Bengio, Y. Challenges in representation learning: A report on three machine learning contests. Neural Networks, 64:59--63, 2015. doi: 10.1016/j.neunet.2014.09.005.

[15] Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. Tech Report. https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf (accessed 12 December 2021).

[16] Nilsback, M. E., & Zisserman, A. (2008, December). Automated flower classification over a large number of classes. In 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing (pp. 722-729). IEEE.

[17] Barz, B., & Denzler, J. (2019, January). Hierarchy-based image embeddings for semantic image retrieval. In 2019 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 638-647). IEEE.

[18] Bryan S. Morse. Image Understanding. http://www.sci.utah.edu/~gerig/CS6640-F2012/Materials/BMorse-BYU-iu-active-contours.pdf – Title from the screen (accessed 12 December 2021).

[19] Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3156-3164).

[20] [20] Hossain, M. Z., Sohel, F., Shiratuddin, M. F., & Laga, H. (2019). A comprehensive survey of deep learning for image captioning. ACM Computing Surveys (CsUR), 51(6), 1-36.

[21] Bisikalo, O., Bogach, I. & Sholota, V. (2020). The Method of Modelling the Mechanism of Random Access Memory of System for Natural Language Processing. In 2020 IEEE 15th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET) (pp. 472-477). doi: 10.1109/TCSET49122.2020.235477.

[22] Kovenko, V., & Bogach, I. (2020). A Comprehensive Study of Autoencoders' Applications Related to Images. In IT&I Workshops (pp. 43-54).

[23] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

[24] Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. arXiv preprint arXiv:1607.06450.

[25] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., ... & Bengio, Y. (2015, June). Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning (pp. 2048-2057). PMLR.

[26] Nezami, O. M., Dras, M., Anderson, P., & Hamey, L. (2018, September). Face-cap: Image captioning using facial expression analysis. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 226-240). Springer, Cham.