Emotional classification on Reddit dataset using embedding fusion approach

Akshay uppal
Master's Student at UIC
akshayuppal3@gmail.com

Prof. Natalie Parde
Assistant Professor
Department of computer science, UIC
nparade@uic.edu

Abstract

Emotions form a biological necessity to express ourselves to other in a non-verbal fashion. The world around us is multimodal, consisting of sense of sight, smell, hear, feel and taste. Similarly, it makes sense to study emotion in a multimodal fashion. We have seen quite advancements in the area of multimodal learning. In this paper, we explore the joint fusion of image and language modality which seems to be a promising area in the area of emotional classification and affective computing. This paper also proposes a variant of joint fusion to fuse the representation of different text to improve the existing baselines of joint fusion method.

Figure 1: Plutchik's wheel of emotions(Plutchik, 1980)

1 Introduction

Emotions have been studied quite extensively in the domain of art, psychology, cognitive neuroscience and also in the multidisciplinary field of affective computing. Emotions evolved as a biological necessity to communicate our thoughts to others in a non-verbal fashion. Emotions are a universal phenomenon and each of the humans expresses the same facial expression corresponding to various emotions that we know, irrespective of race, color or gender(Porter and Ten Brinke, 2008). We express emotions in our everyday lives which help in decision making, learning, communication, and situation awareness (Poria et al., 2017). From cognitive neuroscience, perspective emotions are thought of as judgments, about the extent that the current situation meets our goals and perceptions, indicating a change in the internal state of our body like heart rate, hormones level, breathing rate(Thagard and Aubie, 2008). Over the last decade, AI has made significant evolution in order to make the machine capable of recognizing, interpreting and also expressing similar human emotions. These can be attributed to the domain of affective computing. In order to study emotions, people categorized into discrete categories, from the time of Romans and Greeks to the present time of the AI revolution. In this paper, we categorize emotions into discrete categories based on the idea of Plutchik's 1980's wheel of emotions which is based on evolutionary principles of Darwin (Plutchik, 1980).

The rest of the paper is organized as follows, section 2 discusses related work in the field of emotional classification; section 3 discusses the different theories in the unimodal perspective; section 4 further explains the fusion of these modalities; section 5 we discuss experimentation settings, section 6 explores our proposed approach and finally we conclude in section 7 with potential future work.

2 Related work

Multimodal classification can be classified mainly under two categories depending upon how we combine the information together. In early fusion,

emotion	#posts
joy	1119
fear	697
anger	613
disgust	810

emotion	#posts
joy	384
fear	384
anger	384
disgust	384

Table 1: Distribution of posts for each of the emotion category in the Reddit Dataset (main dataset) (Duong et al., 2017) and the corresponding dataset after filtering invalid URLs (filtered dataset) and creating a balances dataset.

we combine the feature vectors of different modalities together to form a resultant multimodal vector which is then passed to the classifier for classification. The idea of feature level fusion is that concatenation of feature vectors at an early stage helps to exploit the correlation and interaction of low-level features of each modality. (Poria et al., 2017)(Baltrušaitis et al., 2019), which is also referred to as multimodal representation learning. This helps the model to learn from both the modalities and leads to improvement as compared to unimodal counterparts where we have only single modality. In contrast to early fusion, *late fusion* refers to decision-based fusion, features of each modality is classified independently and results are based on a fusion of the decision vectors (Poria et al., 2017). Fusion mechanism employed by late fusion includes averaging, voting schemes, etc. Late fusion is based on the assumption that the underlying data from different modality is independent of each other. It ignores low-level interaction between the multiple modalities, which makes late fusion a better predictor of multi-modal learning tasks. (Baltrušaitis et al., 2019) Our work is mostly inspired by Duong and et al. paper on Multimodal emotional classification. They employed a novel and simple variant of early fusion and is still capable of handling missing data in any modality. They apply their method in the context of emotional classification and also created two social networking datasets (Flick and Reddit), which were the first comprehensive large scale dataset containing both the visual and textual data in the context emotional domain. The content posted by users on a social network is a way to express their emotions and feelings and share their stories with others. Users are free to post text and images as a means of self-expression, which makes it an ideal place to understand human behavior and in our case to study emotional classification.

3 Theory

We use the Reddit dataset as provided in the paper(Duong et al., 2017). It is a dataset that has been crawled for the online social networking website and has been organized into four categories, mentioned in Table 1 **??**. In this section, we explain the unimodal models for language and image modality.

3.1 language modality

For the language modality, we exploit the Bi-LSTM network which is the state of the art in the language domain (Huang et al., 2015) as it preserves the long term dependency of words in the text as opposed to BOW (bag of words) model which assumes independence among the words. Although the version mentioned in ((Huang et al., 2015) is for entity extraction and uses CRF(Conditional random field) it can be extended for simple text classification as well. We remove the top CRF layer and uses a variant bidirectional LSTM (referred as Bi-LSTM further) for simple classification of emotional categories.

3.2 image modality

In the image domain, we experiment with CNN models which have been proven quite successful in object classification domain ((Duong et al., 2017)). CNN is able to capture features of the images which leads to accurate classification. For our problem, we can use CNN to extract those features vectors for the join fusion model. Experimentally we take a pre-trained CNN on Imagenet (Deng et al., 2009) data and remove the last fully connected layer and keeping all of the remaining layers intact (a process similar to mention in paper (Duong et al., 2017)). We experiment with various CNN architectures namely VGG-16, Resnet50, and Inception Net, details of which are mentioned in the experiment section.

model	F1-score
embedding-dense - (no LSTM)	0.70
lstm-dense	0.72
bi-lstm-flatten layer	0.74
2 Bi-(flatten layer)	0.75
2 Bi- (mean layer)	0.68
2 Bi -(sum layer)	0.70

Table 2: various configurations of Bi-LSTM model architecture

embeddings	precision	recall	F1-score
glove	0.72	0.71	0.71
fastText	0.72	0.71	0.71

Table 3: various configurations of Bi-LSTM model architecture

model	F1-score
inception-net (2048-D)	0.40
VGG-16 (4096-D)	0.23
Resnet50 (Resnet50)	0.61

Table 4: Final summary of various models for the image classification task on Reddit dataset (Duong et al., 2017)

4 Fusion of image and language modality

The paper (Duong et al., 2017) mentions four different fusion techniques including late fusion, early fusion, joint fusion, and common space fusion. Main techniques mentioned in this paper is the joint fusion approach we concatenate the image and text vectors. In the joint fusion approach, we still consider both of the modalities in separate planes and keep the dimension of each of the modality of the same size.

4.1 discrete emotion categories

The study of emotional categories dates back to the ancient Greeks and Romans. An early example is by Cicero where he organized the emotions into four basic categories: metus *fear*, aegritudo *pain*, libido *lust* and leatitia *pleasure* (Poria et al., 2017). The most popular one in the recent times are Ekman Categories which organize the emotions into six basic categories: happiness, sadness, fear, anger disgust and surprise, which has led to a new field of affective computing (Poria et al., 2017). As human emotions are complex instead of categorical, there have been studies to represent emotions in multidimensional space. Most of the dimensional approach define the emotions into two categories, including Russell's circumplex model (Russell, 1980), which employs 2 dimensions of arousal and valence to plot 150 affective labels (Poria et al., 2017) or the Whissel's (Whissel, 1989) 2D emotional emotions whose dimensions include evaluation and activation. In this paper, we refer to the Plutchik's 1980's wheel of emotions which categorizes emotions on evolutionary principles of Darwin (Plutchik, 1980). The wheel of emotions consists of 8 basic emotions and 8 advanced emotions Fig 1, vertical dimension represents the intensity and radial dimension represent the degree of similarity among the emotions. The dataset referred in the paper is crawled from the social networking website Reddit by (Duong et al., 2017). Reddit is a discussion website where discussions are organized by topics (called subreddits). Each of the discussion threads in Reddit is labeled by a number of users, based on a reputation system, which enforces each subreddit to belong to a relevant category. Duong et al. focus on 4 subreddits (happy, creepy, gore, rage) which has the text and associated images in their discussion threads. Duong et al. map these subreddits categories to emotions of joy, fear, anger and disgust in the Plutchik's wheel of emotions, Fig **??**. For each subreddit, they crawled 1000 submissions with the highest number of up-votes and discarded the ones that didn't contain both image and text. In the experimentation stage, some of the URLs were invalid so we further filtered those posts. Finally to create a balances set we

Model	Classifier	F1-score
Inception-net (2048-D)	SVM	0.38
	Rf	0.40
	PCA-SVM	0.12
Inception-net (1000-D)	SVM	0.24

Table 5: testing outer average pool layer and dense layer in inception-net

Model	Classifier	F1-score
VGG-16 (4096-D)	SVM	0.09
	Rf	0.29
	PCA-SVM	0.14
	Naive bayes	0.24

Table 6: extracting the image features from the fc2 layer of VGGG-16 architecture

Model	Classifier	F1-score
Restnet50 (1000-D)	SVM	0.53
	Rf	0.57
	PCA-SVM	0.61

Table 7: extracting image feature from the last fc2 layer of Resnet50 architecture

took 384 pots for each of the categories. (referred to as dataset 2).

5 Experimentation

5.1 Baselines with language

For the language modality, we take the title of each of the post and use Bi-LSTM model as explained above. It is different from the one used in the (Duong et al., 2017) paper. As the initial baselines gave better result with the Bi-LSTM we end up using it as the text classification and for further fusion process. We preprocess each of the text (or titles of images): namely tokenize, remove stop words (NLTK library) and reduce each of the words to its base form. Initial baseline was captured using the entire dataset Table **??** without filtering as text data. We use a simple variant Bi-LSTM model (Huang et al., 2015) without the CRF layer, having an embedding layer, followed two bi-directional layers and time distributed layer and configuration layer (architecture illustrated in 4) which we replace and test in different configurations. We train the model on the full Reddit dataset utilizing the title of image serving as the language modality. We tested the Bi-LSTM model in various configurations, Table 2. For initializing the weights of the bi-LSTM model we use glove (Pennington et al., 2014) twitter embeddings which are suited for the social networking data and further provided an analysis with the fasttext embedding (Joulin et al., 2016). Based on the results of the Bi-LSTM model configurations we use the Bi-LSTM flatten layer model as it performs the best among the existing models. We finally tune the model on the filtered dataset, also we test the model using the fasttext embeddings as mentioned in the (Duong et al., 2017) paper and compare with the glove embeddings. Although we didn't see any difference in fastText as compared to the glove it can be due to the small dataset. We use customized fastText embeddings trained on Reddit dataset mentioned in 4.

5.2 Baselines with image

For the image, we test various CNN models and we remove the top layer and use the feature extracted from the top layers to classify the images into 4 categories. We get the baselines with Inception-Net (Ioffe and Szegedy, 2015), Resnet (He et al., 2016) and VGG16 model (Han et al., 2015) architecture. All of the weights of each of the CNN models are pre-trained on imagenet data (Deng et al., 2009) and we don't train the models for extracting the initial baselines. Each of the images is preprocessed using Keras (Chollet et al., 2015) preprocessing library, the size of the images are changed based on the input of each of the respective CNN models. For each of the models, we remove the top fully connected layer to extract the linear feature vector layer. We use the extracted features and pass to different classifiers to classify the images into the four categories. For the simplicity purposed we report the F1-score for each of the CNN models. Inception-net model we experiment with avg pooling layer (2046-D) and Dense layer (1000-D), mentioned in 5. We test each top linear 'fc2' layer in each of the VGG-16 and Resnet50 model in a similar fashion presented in 6 and 7 respectively. Among the models, we experimented Resnet50 model performs the best among them. So we further fine-tune the model using all of the images of Reddit dataset. For the image representation, we use Resnet50 as highlighted in Table **??** which seems to give better

model	F1-score
inception-net (2048-D)	0.40
VGG-16 (4096-D)	0.23
Resnet50 (Resnet50)	0.61

Table 8: finaly summarisation of various models for the image classification task on Reddit dataset (Duong et al., 2017)

pooling type	F1-score	accuracy
concat	0.79	0.78
max	0.77	0.77
mul	0.76	0.75
subtract	0.78	0.77
subtract	0.78	0.77

Table 9: Different modes of fusion joint fusion of image and language vectors

results as compared to other models. We further fine tune the Resnet50 model on our dataset. The model is fine tunes by adding a fully connected 100 Dense layer on top of it followed by a softmax layer to categorize the 4 categories of Reddit dataset.

5.3 Fusion of image and language modality

Based on the joint fusion (Duong et al., 2017) we fuse the extracted features of image and language together to classify the images among the four discreet categories.

$$x = pooling(\gamma(i), \psi(i)) \tag{1}$$

In the joint fusion approach, the post vector x is constructed by concatenating the image vector $\gamma(i)$ and text vectors $\psi(s)$. In this approach, we want to consider both the modalities of image and text. Although in this approach we still consider both the image and text to be present in vector spaces. We change the way the image and text modality are fused together while keeping other layers similar to late fusion.

With respect to joint fusion, we remove the top layer of bi-lstm model (the softmax layer) and only include the last Dense layer. As explained in section 5.1. For the image modality, we take the fine-tuned Resnet50 model and remove the top layer. To have a similar dimension as language modality we add a dense layer to it.

To reach to best joint fusion we experimented with various embedding dimensions and we get the best results with 100-D vector size for each of the modality. Furthermore we experimented with various pooling strategies in addition to concatenation (mentioned in 10) which discussed in (Duong et al., 2017). Further note we didn't experiment with the common space fusion strategy discussed by (Duong et al., 2017) but we exhaustively experimented with various joint fusion approach 10 and in the next section we discuss the proposed a *embedding fusion* method to improve the existing joint fusion methodology. We see that in 10 we get an improved score using the joint fusion approach in comparison to the unimodal baselines which is expected as mentioned in (Duong et al., 2017). The score on the image-only as presented in 10 is lower as compared to the one presented in **??**, as the current one we reduce the Dimension to 100 and its top layers are trained on the image dataset. One may note that we get a very low score in image 0.57 as opposed to 0.73 and similarly a bit low score for language (0.71 as opposed to 0.78) as mentioned in the paper but this could be attributed to the fact we filter the dataset as some of the images point to invalid URLs.

6 Proposed approach

We experimented with the joint fusion approach in which we project the unimodal representation into multimodal space [@baltruvsaitis2019multimodal]. Although we didn't experiment with the common space fusion model which is state of the art, proposed by Duong et al, we propose a method to improve the joint fusion approach.

modality	Precision	Recall	F1-score
image-(Resnet50)	0.58	0.57	0.57
language-(bi-LSTM)	0.72	0.71	0.71
joint-fusion	0.72	0.71	0.71
embedding-fusion (pooled)	0.78	0.78	0.78
embedding-fusion (concat)	0.79	0.78	0.78

Table 10: Scores related to unimodal, joint fusion and the proposed embedding fusion approach on the filtred dataset

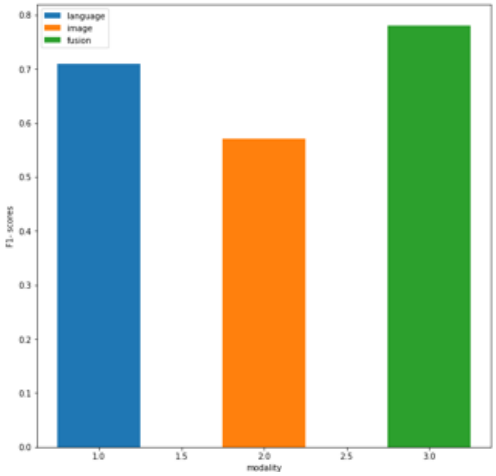


Figure 2: unimodal baselines in comparison with the joint fusion baseline. The Blue bar represent the language baseline, orange image and finally green represent the joint fusion of both modalities. We can see a clear improvement using the fusion approach

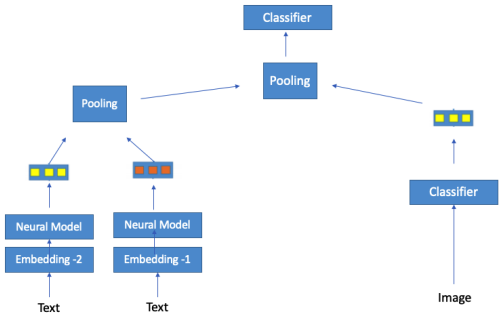


Figure 3: Proposed embedding fusion method as an extension to the early fusion method

6.1 Embedding Fusion approach

As we have seen a surge of evolution of embeddings like fastText ((Joulin et al., 2016)), glove ((Pennington et al., 2014), bert ((Devlin et al., 2018) ,elmo ((Perone et al., 2018)) etc. we propose a method to incorporate a fusion of the existing embeddings to improve the scores of joint fusion. The idea is to capture the different representation of the same text using a combination of embeddings. The idea is intuitive in case of images where we use different kernels to capture different features of the image. Similarly, we can extend the idea to text with each of different embeddings representing different representation of the text. Although the vector representations would be different of each embeddings but they represent the same words. For instance in our proof of concept approach we fuse the glove and fasttext embeddings together, which can be used in conjugation as they return the word vectors for each word. Glove embeddings are based on shallow networks in which it learn vector representation of the word given the context or context given the word (skip-gram and CBOW). FastText embeddings also uses a shallow network similar to glove but they learn the representation of each of the sub-words as well.(Joulin et al., 2016). For example: they learn the representation of sub-words shal, sha, sh from the word shallow, which help in case of accurately representing rare words. Each of glove and fastText return a static representation of each word based on its context trained on corpus of billion of words. This makes them a great for our fusion experiment. We present our results in (Table 9 10) which comparative study with unimodal baselines and joint fusion approach.

6.2 Analysis

Although the proposed embedding fusion didn't provide better results as compared to the existing joint fusion approach. This could be attributed to various reasons. One reason could be attributed to the fact that as the embedding are associated in different vector space their fusion would not be accurate and we could perform further tests using the common space fusion approach as highlighted by the author (Duong et al., 2017). Other reason could be that pooling of one existing glove embedding in addition to existing embedding of a word

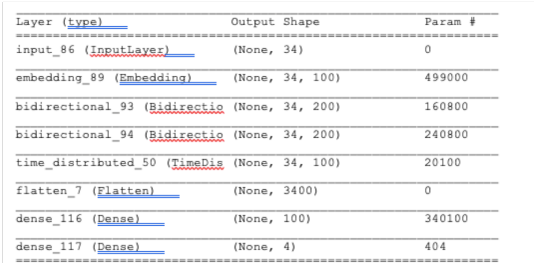
is not contributing any additional information but rather a different representation so the weights of the neural model would only existing embedding and hence assign higher weights (or relevance) to the first part of the existing embedding and lower importance to the second half of the embedding. Thereby, not leading any improvement in the scores with the fusion approach.

7 Conclusion and Future work

Although the proposed models didn't seem to fair any better than compared to existing model but to totally disapprove it we require further test using common space fusion approach (Duong et al., 2017) and with further datasets (eg Flickr) (Duong et al., 2017)).

8 Acknowledgements

I would like to thank Prof. Natalie Parde for their consistent guidance, support and constant feedback which has led to the improvement and evolution of the proposed project.



Layer (type)	Output Shape	Param #
input_86 (InputLayer)	(None, 34)	0
embedding_89 (Embedding)	(None, 34, 100)	499000
bidirectional_93 (Bidirectio	(None, 34, 200)	160800
bidirectional_94 (Bidirectio	(None, 34, 200)	240800
time_distributed_50 (TimeDis	(None, 34, 100)	20100
flatten_7 (Flatten)	(None, 3400)	0
dense_116 (Dense)	(None, 100)	340100
dense_117 (Dense)	(None, 4)	404

Figure 4: bi-lstm architecture

References

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.

François Chollet et al. 2015. Keras. https://github.com/fchollet/keras.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

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*.

Chi Thang Duong, Remi Lebret, and Karl Aberer. 2017. Multimodal classification for analysing social media. *arXiv preprint arXiv:1708.02099*.

Song Han, Huizi Mao, and William J Dally. 2015. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Christian S Perone, Roberto Silveira, and Thomas S Paula. 2018. Evaluation of sentence embeddings in downstream and linguistic probing tasks. *arXiv preprint arXiv:1806.06259*.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier.

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.

Stephen Porter and Leanne Ten Brinke. 2008. Reading between the lies: Identifying concealed and falsified emotions in universal facial expressions. *Psychological science*, 19(5):508–514.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Paul Thagard and Brandon Aubie. 2008. Emotional consciousness: A neural model of how cognitive appraisal and somatic perception interact to produce qualitative experience. *Consciousness and cognition*, 17(3):811–834.

Cynthia M Whissel. 1989. The dictionary of affect in language r. plutchnik and h. kellerman (eds) emotion: Theory, research and experience: vol 4, the measurement of emotions.