Multimodal Classifier for Emotion Recognition with Convolutional Neural Networks

Montana St. Pierre
Department of Computer Science
Columbia University
New York, USA
mrs2296@columbia.edu

Abstract

Emotion recognition is a burgeoning task within the fields of computer vision and natural language processing that can help elucidate how the incomprehensibly vast corpus of online visual and textual user-generated content elicits emotional arousal. However, the creation of largescale annotated emotion datasets is hindered by the inherent ambiguity in both the definition of the task and the conception of emotion, and research thus far has made do with difficult to reproduce weakly labeled datasets or small annotated but imbalanced datasets. Both typed limit the capacity of machine learning algorithms and instill models with high domain specificity. In this paper, we utilize the latest advancements in deep learning to implement visual and textual convolutional neural networks (CNN) to recognize eight emotional classes. Furthermore, we address the dataset quality and quantity issues by incorporating the often colocated visual and textual features—yet never simultaneously used—in a multimodal fusion CNN. Experimental results show a substantial improvement in single label multiclass emotion classification accuracy.

1 Introduction

Emotion is an innate component of human nature; while the lexical representation and conceptualization of emotion must be learned, the sensation of feeling emotion is inextricable from humanity. In conjunction with cognition, it centrally influences executive function, social behavior and overall health. As such, this topic has long been a focal research subject in the fields of neuroscience and psychology. A multitude of mechanisms and models have been proposed in the quest for understanding emotion; however, meta-analysis uncovers traditional observational studies to be as inconsistent and subjective as emotion itself (Weidman et al., 2017). Although the contribution of these recent smaller-scale studies is contested, new large scale multidisciplinary research with computer scientists has harnessed big data—in the form of self-generated multimodal content produced by social media users—to overcome the pitfalls of previous work.

Technology mediates the human experience of emotion similar to that resulting from human-human interactions. Reeves and Nass (1996) theorized the anthropomorphic treatment of electronic media and they observed a psychological response to messages conveyed via virtual audio, text and video consistent with conventional human communication. As the availability and latency of digital media platforms have become on-demand and instantaneous, resulting largely from the ubiquity of smartphones and social media, there exists the potential for constant emotional arousal arising from their usage. Early multidisciplinary research by Bollen et al. (2011) analyzed the sentiment of microblog posts (tweets) on Twitter and they found a significant correlation between mood and emotion on a public scale with world events. The controversial work by Kramer et al. (2014) manipulated the average sentiment of content visible by unknowing Facebook users and uncovered an emotional contagion effect. Both studies exemplify the utility of research and emotional effect at scale, and the importance of multidisciplinary research informed by ethics and existing psychological theories.

Inspired by the rapidly emerging field of visual emotion recognition and advancements in deep learning, we endeavor to develop a multimodal emotion response classifier incorporating visual and textual features. Using the strongly labeled Flickr dataset developed by You et al. (2016a), we ground the solution in developed emotional response models rather than borrowing from those

intended for affective computing—the interpretation task of human affect. In summary, we contribute the following.

1. Using state-of-the-art CNNs trained for ImageNet classification, we employ transfer learning to train new models specialized to the domain of emotion recognition.

2. Similarly embracing CNNs, we nontraditionally develop a CNN model instead of a recurrent neural network (RNN) to classify emotions on textual features.

3. We combine the independently trained models above using late fusion resulting in a novel deep learning approach to multimodal textual and visual features emotion recognition.

4. We analyze the two unimodal models through comparison to state-of-the-art textual and visual emotion recognition models and the ultimate multimodal model, and we further observe the similarity of classification performance by the two modes.

2 Related Work

2.1 Multimodal Sentiment Analysis

Sentiment analysis can be viewed as an easier subproblem of the emotion recognition problem for it only performs binary classification into positive or negative sentiment. As with many problems in machine learning, the visual sentiment analysis problem is more difficult and less studied than the textual counterpart due to the costs and increased complexity of higher dimensionality. Borth et al. (2013) undertook one of the earliest works in visual sentiment analysis by creating an ontology of adjective-noun pairs which were reduced to a sentiment score using traditional machine learning algorithms. This ontology comprises images queried from Flickr on Plutchik's (1980) twenty-four emotions—representing the technique of weak labeling which often used to build datasets involving sentiments and emotions—from which combinations of adjective-noun pairs are discovered. A novel multimodal sentiment analysis model was developed by You et al. (2016b) showing a gain in classification accuracy through the combination of visual and textual features. One of the simpler fused neural network architectures from their work is later seen in the most relevant experiments performed by Hu and Flaxman (2018) on yet another different weakly labeled dataset scraped from Tumblr. Their model slightly outperforms that of the sub-architecture on solely textual features as confirmed with earlier work. Despite describing the problem as sentiment analysis, the results are confusingly framed in terms of how to better understand the structure of emotions. Yet, it remained unclear whether the model directly predicted emotion or sentiment. Regardless, this paper succumbs to the common weakness shared by the above papers in that the use of weakly labeled data—tags or content embedded by the author—predict the sentiment (or emotion) prescribed by the author.

3 Multimodal Emotion Models

In anticipation of fusing together a visual and textual model for the multimodal emotion classifier, we opt for the selection of individually efficient and lightweight architectures. This choice is motivated by the potential for computation time to balloon when fused or overtraining to occur from too many parameters. Moreover, the ultimate exploration of whether a fused model can improve emotion classification performance makes little sense if one of the component models were complex and highly accurate. The multimodal fusion model should result in better performance for less cost than either method alone.

3.1 Visual Model

We craft a simple deep CNN by utilizing a state of the art model pretrained for the ImageNet Large Scale Visual Recognition Challenge (Russakovsky et al., 2015). As exemplified by Yosinski et al. (2014), transfer learning adapts the high-level features learned by a CNN across domains while providing better than random initialization for the learning of low-level features in the new domain. Upon this principle and considerations of computational performance, we choose DenseNets to serve as the base model.

DenseNet Connectivity Overview

Huang et al. (2016) proposed the Dense Neural Network (DenseNet), an extremely deep neural network, to circumvent the impositions placed on network depth by limited connectivity, poor feature reuse and vanishing gradients. A traditional feed-forward network accepting input vector $\mathbf{x}_0$

and consisting of L layers, each applying a non-linear transformation function H_l, is described by the recurrence relation:

$$\mathbf{x}_l = H_l(\mathbf{x}_{l-1}),$$

such that the input for each layer only connects to the output of the previous layer. DenseNets directly increase network connectivity by progressively concatenating features:

$$\mathbf{x}_l = H_l([\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_{l-1}]),$$

thus allowing each layer a connection with all preceding layer outputs. In a transfer learning application, we purport that the direct propagation and preservation of the high generality features to deeper layers will reduce the processing of fine-tuning to all but the very lowest layers.

Image CNN Architecture

We select DenseNet-121, the least complex DenseNet model for ImageNet classification, as the foundation of our emotion classifier on images. The model consumes images of size 224 x 224 x 3 through an initial convolutional layer followed by 119 composite convolutional layers arranged into 4 blocks with an intermediary reduction in channels. Here, composite convolutional layers comprise batch normalization (BN) (Ioffe and Szegedy, 2015), rectified linear unit activation (ReLU) (Nair and Hinton, 2010) and convolution in series. The final fully connected layer with 1000 neurons is pruned in order to enable transfer to the emotion task with only 8 classes. Therefore, we take the pooled output that precedes the fully connected layer and augments it by appending a dropout (Srivastava et al., 2014) layer and a new fully-connected layer with 8 neurons and softmax activation to predict the emotion class. Dropout randomly drops units from the network during training to combat overfitting, and we find it beneficial to improve the validation accuracy (generalization power) concomitant with the reduction in the dimensionality of the classification targets and the small training dataset.

3.2 Textual Model

Recurrent Neural Networks (RNN)

Long short-term memory (LSTM) recurrent networks, proposed by Hochreiter and Schmidhuber (1997), have long been a standard approach for learning temporal in addition to spatial features on sequences of text. More recently, Cho et al. (2014) refactored the LSTM architecture into the computationally and conceptually simpler gated recurrent neural network (GRNN). While these architectures provide state of the art performance across a diverse range of natural language processing tasks, they become prohibitively more difficult to train as sequence lengths increase. Sequence lengths in practice are limited to hundreds of tokens after which the depth, in a one-to-one relationship with sequence length, incurs vanishing gradients and long processing times. Particular domains fit this usage pattern well, especially when a limit on sequence length is prescribed (e.g. Tweets have a 140 character limit), but a more general solution to the multimodal text accompanying an image should elide domain specificity and perform equally on the long-form text.

Temporal (1D) Convolution

Pure CNNs have performed well on textual classification, where for example, Kim (2014) deployed a network convolving a k-dimensional sequence $\mathbf{x} \in \mathbb{R}^{nk}$ of length n embedded words with a kernel $\mathbf{w} \in \mathbb{R}^{hk}$ of width h. Unfortunately, for the successful production of word embeddings through dimensionality reduction, a vocabulary must have a degree of normalization. This precludes learning features that cannot be captured by a limited vocabulary, most often equating to colloquialisms, non-English words, stylistic capitalization and typographical errors. Zhang and LeCun (2015); Zhang et al. (2015) found that alternatively, convolving at the character level exploited signal-like qualities of the text, behaving akin to image convolution, and in turn solved the dimensionality crisis and narrow contextual range of model applicability. Although they find performance improved on large datasets, we predict character level CNNs to excel on the noisy, user-generated textual content.

Character CNN Architecture

Following from the Small Frame model described in Zhang and LeCun (2015); Zhang et al. (2015) with 6 consecutive temporal convolutions, we propose a similar model with the main addition of parallel convolutional blocks.

As depicted in Table 1, the convolutional block is formed by a sequence of convolutions with a given kernel and dilation size, followed by convolution with an equally sized kernel, dilation of

Depth	Kernel	Stride	Dilation
128	7	1	D
128	7	2	1
128	5	1	D
128	5	2	1
128	3	1	D
128	3	2	1

Table 1: Block of convolutional layers where D determines the dilation factor.

Layer	Size
Embedding	$N * 64$
Conv($D = 1$) **:** Conv($D = 2$)	$\frac{N}{8} * (128 + 128)$
Concatenation	$\frac{N}{8} * 256$
Global Average Pooling	256
Dense	8

Table 2: Macroscopic architecture with description of layers and output sizes where N is sequence length. Note the parallel convolutional blocks from 1 with dilations of 1 and 2.

1 and stride of 2. This pattern is repeated three times for kernel sizes of 7, 5 and 3 in order to ensure a large receptive at the input. We remove the fixed pooling operation, instead favoring the halving of the sequence dimension by means of strided convolution thus introducing a learnable temporal downsampling operation. Additionally, we adjust the dilation rate to implement atrous convolution from Chen et al. (2016) to further increase the receptive field without increasing the number of parameters. After each convolutional layer in the block, we apply layer normalization (LN), redeveloped by Ba et al. (2016) as an alternative regularization to BN where the minibatch size varies, and subsequently from He et al. (2015), the parameterized rectified linear unit (PReLU) activation.

The overall architecture, as described in Table 2, first embeds each character from the input sequence into a vector of length 64. Two convolution blocks as described above, of dilation rates 1 and 2, transform the embedded sequence in parallel with 128 features each. The results of these blocks are concatenated along the feature axis, averaged with global pooling along the sequence axis and finally connected densely to eight softmax activated neurons to predict the emotion class.

3.3 Fused Model

The fusion of the individual modality models is similar in practice to instantiating a transfer model from pretrained networks. We could have inserted additional convolutional layers to combine the model outputs or fused the output of one modality into the middle layers of the other, but we decided to use simple late fusion. The output layers of the base Image CNN and Character CNN models are replaced by dense layers with eight units, layer normalization and PReLU activation. These two results are summed post-activation and subsequently passed to a final dense layer with eight units and softmax activation.

4 Experiments

The Tensorflow 2.0[1] deep learning framework is used to develop and evaluate our three model architectures (Abadi et al., 2015). Motivated by newly introduced tight coupling with Keras (Chollet et al., 2015), we adopt the second iteration of this well-established framework for the developmental efficiency facilitated by the functional paradigm and seamless integration of pretrained models.

We execute the experiment within a Windows 10 Python 3.6.8 environment on a laptop equipped with one hexacore Intel i7-8750H processor, 16 GB of DDR4-2666 memory, one NVIDIA GeForce GTX 1070 with Max-Q and 1 TB of solid-state storage on one Samsung SSD 960 Pro. GPU acceleration makes use of NVIDIA driver 417.35, CUDA 10.0 and cuDNN 7.5.1.

4.1 Dataset

We experiment on the largest annotated multiclass single label emotion dataset featuring visual and textual content created by You et al. (2016a) from images hosted on the Flickr social photo sharing community. It was crafted by querying Flickr [2] on the eight emotion classes proposed by Mikels et al. (2005) (i.e. anger, amusement, awe, contentment, disgust, excitement, fear and sadness) and filtering duplicate and multiple label photos. After this weakly labeled construction, an Amazon Mechanical Turk task asked workers to answer whether they felt the weakly labeled emotion upon viewing. A majority of five votes constituted label verification, and the resultant set contained

[1] In alpha at the time of this experiment.

[2] Additonally, Instagram was queried in the construction of the dataset, but the number of annotated results is insignificant compared to that from Flickr. Thus, we ignore these auxiliary images to more uniformly conduct this experiment.

21,110 strongly labeled Flickr images across the eight emotions.

Emotion	Train	Val	Test	Total	Weight
Contentment	3218	690	690	4598	23.59%
Amusement	2883	618	618	4119	21.14%
Awe	1895	406	406	2707	13.89%
Excitement	1749	374	375	2498	12.82%
Sadness	1662	356	357	2375	12.19%
Disgust	982	210	211	1403	7.20%
Anger	703	151	150	1004	5.15%
Fear	549	118	117	784	4.02%
Total	13641	2923	2924	19488	100.00%

Table 3: Label distribution of the train, validation and test datasets. The use of a stratified split on the overall dataset produces these subsets maintaining distributions equal to the whole.

Since the initial publication of the Flickr dataset, a faction of the images have been removed or are no longer accessible. The raw dataset provides image URLs and labels agreement/disagreement votes, and by querying the Flickr API on image identifiers extracted from the URLs, we find 19,488 images and metadata available at the time of writing. We save the descriptions, tags and titles for feature use alongside the images, and randomly split the data into stratified (i.e. preserving the overall label distribution in the subsets) training (70%), validating (15%) and testing (15%) subsets as described in Table 3.

Preprocessing

Images are scaled[3] to 256 pixels along the smallest dimension, and then center cropped to the 224 x 224 input size of the CNN. The text undergoes equally little preprocessing. Line breaks are replaced with spaces as necessary to store the text for each datum on one line, and hypertext is removed entirely as it functions outside the domain of the content. Unlike previous assemblages of textual data in emotion classification tasks, such as those from Twitter by Abdul-Mageed and Ungar (2017) and Saravia et al. (2018), URLs and non-English words are not removed. The character level CNN is able to learn the emotional concepts from the temporal relationship of characters in sequence without the a priori knowledge required with a high dimensional word level deep learning model.

[3]Images are retrieved at the Flickr default medium 640 x 480 resolution, but are smaller in cases where the original resolutions are less than medium. Thus, some images may require upsampling.

4.2 Learning Details

Augmentations are applied to the training data to reduce the likelihood of a biased fit of the most represented classes in the imbalanced dataset. Training images are jittered to randomly crop rather than the deterministic center crop, and then have a 50% probability to be flipped lengthwise. Training text sequences are limited to a length of 4096 to prevent excessive overpadding of the frequent short sequences to the length of the longest sequence in the dataset.

We train all models on the categorical cross-entropy objective function. The objective function for the Image CNN is minimized by traditional stochastic gradient descent (SGD) (i.e. no momentum or decay) with 0.1 set as the initial learning rate. However, it is minimized for the Character CNN and Fused CNN by an adaptive version of SGD, the Adam optimizer (Kingma and Ba, 2014), with initial learning rates of 0.001 and the default parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1 \times 10^{-8}$ of the Tensorflow implementation. We prescribe the learning rates to be automatically halved after three consecutive epochs without an improvement in the validation loss. Data is minibatched with size 32 across the board and randomly shuffled every epoch. Specific to models involving sequential textual features, the Character CNN and Fused CNN, minibatches are formed from the binning of sequences by length into buckets equally spaced by 128 from 0 - 2048 and 512 from 2048 - 4096. Sequences are then zero padded as needed to the upper bound of their respective buckets as mini batching necessitates all sequences to be equal in length, and this improves training efficiency compared to padding all sequences to the maximum length of 4096.

In training the Image CNN from the pretrained DenseNet base model, we freeze the weights in all but the layers of last dense block in order to fine-tune the domain-specific low-level features to the emotion classification task. We train the Image CNN for 25 epochs.

For the Character CNN, we must choose the combination of textual features from the descriptions, tags and titles to train on. Titles are a dense feature, most photographs having a title, of on average short sequence length; descriptions are a

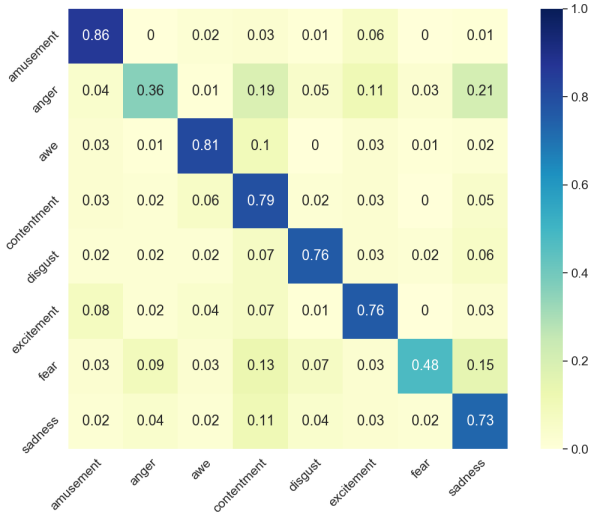

Figure 1: Image CNN confusion matrix.

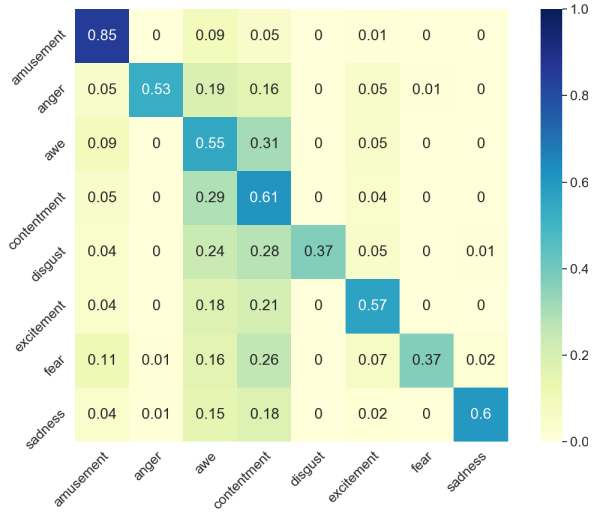

Figure 2: Tags CNN confusion matrix.

more sparse feature, but tend to have much greater lengths on average; while the tags feature are of intermediate sparsity and average length. We train two networks for 50 epochs each, one on the tags feature alone, which we henceforth refer to as the Tags CNN and another on the concatenation of all three features.

The Fused CNN is trained for 50 epochs using the trained Image CNN and Tags CNN with all weights frozen.

4.3 Results

We evaluate the classification performance of the three models on the test subset of the Flickr emotion dataset by comparing accuracy, F_1 score, precision and recall. In terms of the latter three metrics, it is readily apparent in Table **??** that the

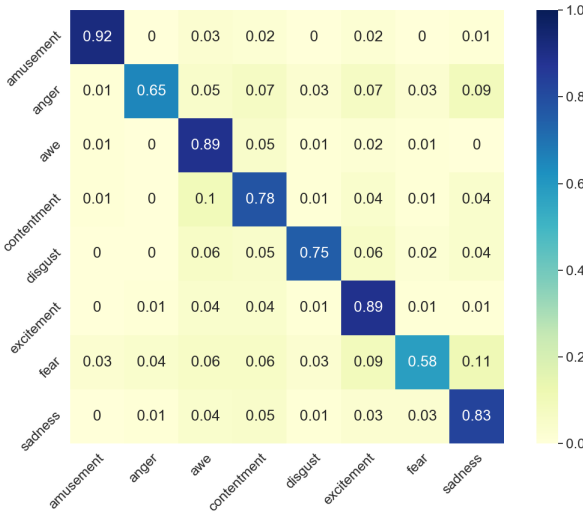

Figure 3: Fused CNN confusion matrix.

Fused CNN model outperforms both the Image CNN and Tags CNN. With the exception of the awe class, the Fused CNN produces significant improvements in F_1 score. This may be explained by the disparity between the unimodular models being the largest for awe at 41%. Otherwise, we observe the hypothesized synergistic effect with an increase in weighted average F_1 score from 76% in the second most performant Image CNN to 83% in the multimodal Fused CNN.

Furthermore, we calculate the confusion matrix for each of the three models to examine their accuracies. Figure 1 shows that the Image CNN model is highly susceptible to the imbalanced dataset, with the two least represented emotions anger and fear having around half of the accuracy of the other six emotions. Conversely, for all classes, we note the tendency of misclassification as contentment which is the most represented emotion in the dataset. This effect is much more pronounced in the Tags CNN model, as displayed in Figure 2, with almost all misclassifications being attributed to well-represented awe or contentment. Finally, from Figure 3 representing the Fused CNN approach, we see the misclassifications are nearly distributed evenly across all emotions. Therefore, the visual modality would benefit to learn from a more balanced dataset while the textual model on intermediately sparse data could have an increased capacity for learning and more regularization. Notwithstanding, the multimodal Fused CNN attacks these different weaknesses by incorporating the learned features from both.

We proffer a final benchmark against two other

	Image CNN			Tags CNN			Fused CNN		
Emotion	Precison	Recall	F_1 Score	Precison	Recall	F_1 Score	Precison	Recall	F_1 Score
Amusement	0.85	0.87	0.86	0.80	0.85	0.83	0.97	0.92	**0.94**
Anger	0.50	0.35	0.41	0.95	0.53	0.68	0.84	0.65	**0.74**
Awe	0.81	0.81	**0.81**	0.32	0.55	0.40	0.72	0.89	0.79
Contentment	0.76	0.79	0.77	0.50	0.61	0.55	0.85	0.78	**0.81**
Disgust	0.74	0.76	0.75	0.96	0.37	0.53	0.88	0.75	**0.81**
Excitement	0.75	0.77	0.76	0.70	0.57	0.63	0.78	0.89	**0.83**
Fear	0.64	0.49	0.55	0.90	0.37	0.52	0.65	0.58	**0.61**
Sadness	0.69	0.73	0.71	0.97	0.60	0.74	0.80	0.83	**0.82**
Micro Avg	0.76	0.76	0.76	0.61	0.61	0.61	0.83	0.83	**0.83**
Macro Avg	0.72	0.70	0.70	0.76	0.56	0.61	0.81	0.79	**0.80**
Weighted Avg	0.76	0.76	0.76	0.70	0.61	0.63	0.84	0.83	**0.83**

Table 4: Percision, recall and F_1 score results with the Image CNN, Tags CNN and Fused CNN on the Flickr dataset with eight emotion classes. The micro, macro and weighted averages are provided below each metric. For the three models, the best F_1 score is bolded for each category.

Method	Accuracy
You et al. (2016a)	58.30%
Yang et al. (2017)	67.48%
Image CNN	75.96%
Tags CNN	61.50%
Fused CNN	82.82%

Table 5: A comparison of the accuracy of our three models to the original CNN trained alongside the creation of the dataset and the state-of-the-art ensemble CNN.

deep learning solutions on this Flickr dataset, including the latest state-of-the-art CNN ensemble by Yang et al. (2017). Both approaches only consider visual features and are grounded in transfer learning like our Image CNN. However, our Image CNN already improves significantly over the state-of-the-art which we attribute to the suitability of DenseNets to transfer features to a new domain. At 82.82% accuracy, our multimodal Fused CNN highlights the additional gain to be had from textual data that has previously gone wasted.

5 Conclusion

In this paper, we conceive a multimodal textual and visual convolutional neural network as a simple late fusion of an image CNN trained using transfer learning and a textual temporal character-level CNN. Evaluation of the three models shows that the multimodal model exceeds the individual modality models in emotion classification performance and that the Image CNN and Fused CNN both outperform the state-of-the-art deep learning solutions.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Muhammad Abdul-Mageed and Lyle Ungar. 2017. EmoNet: Fine-Grained Emotion Detection with Gated Recurrent Neural Networks. pages 718–728.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer Normalization. *arXiv:1607.06450 [cs, stat]*. ArXiv: 1607.06450.

Johan Bollen, Alberto Pepe, and Huina Mao. 2011. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 450–453, Barcelona, Spain. ArXiv: 0911.1583.

Damian Borth, Tao Chen, Rongrong Ji, and Shih-Fu Chang. 2013. SentiBank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content. In *Proceedings of the 21st ACM international conference on Multimedia - MM '13*, pages 459–460, Barcelona, Spain. ACM Press.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. 2016. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *arXiv:1606.00915 [cs]*. ArXiv: 1606.00915.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. *arXiv:1409.1259 [cs, stat]*. ArXiv: 1409.1259.

François Chollet et al. 2015. Keras. `https://keras.io`.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *arXiv:1502.01852 [cs]*. ArXiv: 1502.01852.

Sepp Hochreiter and Jrgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Anthony Hu and Seth Flaxman. 2018. Multimodal Sentiment Analysis To Explore the Structure of Emotions. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - KDD '18*, pages 350–358. ArXiv: 1805.10205.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2016. Densely Connected Convolutional Networks.

Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv:1502.03167 [cs]*. ArXiv: 1502.03167.

Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*. ArXiv: 1412.6980.

Adam D I Kramer, Jamie Guillory, and Jeffrey Hancock. 2014. Experimental Evidence of Massive-Scale Emotional Contagion Through Social Networks. *Proceedings of the National Academy of Sciences of the United States of America*, 111:8788–8790.

Joseph A. Mikels, Barbara L. Fredrickson, Gregory R. Larkin, Casey M. Lindberg, Sam J. Maglio, and Patricia A. Reuter-lorenz. 2005. Emotional category data on images from the International Affective Picture System. *Behavior research methods*, 37(4):626–630.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 807–814, USA. Omnipress. Event-place: Haifa, Israel.

Robert Plutchik. 1980. Chapter 1 - A GENERAL PSYCHOEVOLUTIONARY THEORY OF EMOTION. In Robert Plutchik and Henry Kellerman, editors, *Theories of Emotion*, pages 3 – 33. Academic Press.

Byron Reeves and Clifford Ivar Nass. 1996. *The media equation: How people treat computers, television, and new media like real people and places.* The media equation: How people treat computers, television, and new media like real people and places. Cambridge University Press, New York, NY, US.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252.

Elvis Saravia, Hsien-Chi Toby Liu, and Yi-Shin Chen. 2018. DeepEmo: Learning and Enriching Pattern-Based Emotion Representations. *arXiv:1804.08847 [cs]*. ArXiv: 1804.08847.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

Aaron C. Weidman, Conor M. Steckler, and Jessica L. Tracy. 2017. The jingle and jangle of emotion assessment: Imprecise measurement, casual scale usage, and conceptual fuzziness in emotion research. *Emotion*, 17(2):267–295.

Jufeng Yang, Dongyu She, and Ming Sun. 2017. Joint Image Emotion Classification and Distribution Learning via Deep Convolutional Neural Network. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 3266–3272, Melbourne, Australia. International Joint Conferences on Artificial Intelligence Organization.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks? In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3320–3328. Curran Associates, Inc.

Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2016a. Building a Large Scale Dataset for Image Emotion Recognition: The Fine Print and the Benchmark. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*,

AAAI'16, pages 308–314. AAAI Press. Event-place: Phoenix, Arizona.

Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2016b. Cross-modality Consistent Regression for Joint Visual-Textual Sentiment Analysis of Social Multimedia. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining - WSDM '16*, pages 13–22, San Francisco, California, USA. ACM Press.

Xiang Zhang and Yann LeCun. 2015. Text Understanding from Scratch. *arXiv:1502.01710 [cs]*. ArXiv: 1502.01710.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level Convolutional Networks for Text Classification. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 649–657. Curran Associates, Inc.