Learning to Feel: Training a CNN to recognize emotion

Paul Carroll, Connor Smith, Frank Zheng, William Jarrold, Mana Lewis*
Stanford University
{paulc3, csmith95, fzheng}@stanford.edu
william.jarrold@gmail.com
mana@chezmana.com

Abstract

In the last decade, the development of convolutional neural nets (CNNs) has made it possible for computers to accurately identify the objects in particular photos. Now that object recognition is working well, it seems natural to ask the question of whether we can use CNNs to recognize higher-level concepts in photos. In this paper, we apply a modified CNN based on Inception-Resnet to the task of classifying the emotion evoked by particular images. We begin by training our network to classify photos as having positive or negative valence. Next, we set our network a harder challenge, training it to classify photos as belonging to one of eight emotional states. We use a dataset of 13K photos collected from social networks. For our valence classification problem, we achieve an accuracy of 91%. For our eight-fold problem, we achieve a top-one classification accuracy of 69%.

1. Introduction

In the past few years, with the advent of deep convolutional neural networks, it seems like computers have finally "solved" the object classification problem. In the ImageNet Large-Scale Visual Recognition Challenge[3], algorithms are tasked with identifying which of 1000 possible classes of object is present in a particular image. Inception-Resnet, a powerful CNN which we decided to adapt to our classification challenge, has a top-5 classification error of just 3.08% on the ImageNet test set[8].

This suggests a question: is it possible to adapt convolutional neural networks to more abstract classification problems? When a human looks at an image, he or she is not merely identifying the objects in that photo. The viewer might also have an emotional reaction to that photo. Would it be possible for some kind of deep network to "learn" the expected emotional reaction?

This task is more abstract, and correspondingly more difficult, than object recognition. One immediate challenge is that not all humans will agree on which emotions are evoked by which photos. Unlike in the case of object recognition, in which there is an objective source of truth for which objects are or are not contained in a photo, there is no way to "prove" that a particular photo evokes excitement or fear or sadness. The only question which we can ask is whether a consensus of humans seems to agree on how this photo makes them feel.

It was also not clear when we started this research whether transfer learning from ImageNet would be effective when applied to the task of emotion detection. Transfer learning from ImageNet has successfully been applied to many datasets and classification challenges[1]. Our classification problem, however, is sufficiently different from image classification that it was not clear whether ImageNet would discover "useful" features. It is well understood in the computer vision community that the early layers of a CNN encode edge detection. Edge detection is useful if the goal of a neural net is to detect an object class, since every object in a particular class is likely to have a similar shape. But does every photo with negative valence have similar edges? Probably not. Even if we consider the higher-level features which ImageNet learns, such as some internal representation of a cat or a boat, it is easy to imagine a cat or a boat in a photo with negative or positive valence, though perhaps not with even odds. Part of our research question, then, was to determine if transfer learning from ImageNet would be effective on this task.

We decided to adapt Inception-Resnet to two classification challenges. In the first, we would train it to identify photos as having either positive or negative valence. This simple binary classification challenge would be a good "sanity check" for whether emotion detection is possible at all with transfer learning from ImageNet. Next, we would train the classifier to determine which of eight emotional classes each photo belonged to. This would be a harder

[0]William and Mana helpfully provided data and links to academic papers. Neither is enrolled in CS231N.

test for the classifier, since some of the emotional classes were quite similar to each other: for example, amusement and excitement. We discovered that transfer learning from ImageNet is very effective in the binary classification case, particularly when we augment our training set's data. For the eight-fold classification problem, the classifer's accuracy dropped, but it was still able to identify the correct emotion about two-thirds of the time.

2. Related Work

It has long been understood that, in order for machines to be truly intelligent, they will need to be emotionally intelligent. Classifying objects and answering questions is all very well, but a machine which is unable to understand human emotion is less useful to humans than a machine which can.

In 2001, Rosalind W. Picard et al. trained an algorithm to recognize one of eight human emotions[6]. The input to this algorithm was physiological data–direct measurements of the human subject's vital signs. By using a Fisher projection, the authors were able to achieve 81% classification accuracy. The methodology of their study was very different from ours, since they were trying to fit physiological data instead of pixel data, but the basic goal of the study was similar: to teach a computer to recognize human emotions.

Later authors suggested approaches for teaching an algorithm to recognize latent emotions in a photo. William Jarrold et al.[2] argued that even though emotional reactions vary among different people, there is nevertheless a logic to people's emotional responses. Emotions are not completely unpredictable or arbitrary. Therefore, it should be possible to train an artificial intelligence to recognize human emotions. Jarrold suggests a methodology for how to evaluate an algorithm's emotional intelligence, but does not provide a procedure for how to build such a model.

In recent years, efforts have begun to build labeled datasets of human emotion for computer vision. In 2017, Kurdi et al. published the OASIS dataset[4], which contains 900 images ranked by human subjects on two dimensions: valence, which is whether the image evokes a negative or positive response, and arousal, which is the intensity of the emotional response. This dataset would be useful for researchers who were interested in a regression algorithm for valence or arousal, but cannot be used for the eight-fold classification problem. Fortunately, another dataset exists for that purpose, published by You et al[10]. We will discuss this dataset further in the next section. You et al. appear to have been the first to train a CNN to classify image data into one of eight emotional categories. On their strongly labeled data, their classifier achieved a top-one accuracy of 58%. They did not attempt the binary classification problem.

A plethora of examples exist for how transfer learning from ImageNet has solved classification problems [5][9][7]. In particular, Oquab et al.[5] argue that transfer learning from ImageNet is possible even to datasets with quite different statistical properties from ImageNet, and to datasets which are too small to enable conventional training of deep neural nets. This is highly relevant to our classification challenge, since our dataset is relatively small–just 13,000 images, as opposed to ImageNet's 14 million images. If transfer learning from ImageNet were not possible, it would be highly difficult for us to train a deep neural net over 13k images without overfitting.

3. Data

As previously mentioned, our data come from Yun et al[10]. It is worth describing this data in some detail, since the manner in which it was collected and labelled is highly relevant to the classification accuracy which we were able to obtain. They began by querying image search engines for the eight emotions in which they were interested: amusement, anger, awe, contentment, disgust, excitement, fear, and sadness. In this way, they gathered 11,000 images for each emotion. They then used Amazon Mechanical Turk to ask five humans each if they felt the same emotion when looking at each image. For many of these images, zero humans agreed with the proposed label; but for others, a consensus of humans agreed that the label matched the image.

One observation to make about this dataset is that the methodology of collection doesn't quite fit what we would prefer for a classification problem. In the ideal methodology described by Jarrold et al.[2], each human participant is asked: Does this photo make you feel angry, awe, amusement...etc.? The question is framed as a multiple-choice. According to You's methodology, a photo which Instagram classifies as "anger" will only ever be considered for that category. The human subjects are never asked if the photo would be better classified as "disgust" or "fear." This is significant, since, assuming that image search engines haven't solved the problem of machine emotional intelligence, a search engine's decision to label a photo with a particular emotion is probably quite arbitrary. The poor quality of these search results is confirmed by the large number of cases in which no humans agreed with the search engine's label. Even if the human subjects agree that a photo returned by the search engine represents "amusement," they are never asked if that photo is an even better representation of "excitement." This flaw in methodology is something that should be corrected in future data-gathering (see the Conclusions section).

Another question which any user of this data must answer is how many of the human participants need to agree with the weak label before the label is accepted as valid. If three of the five human subjects agree that a photo represents excitement, should that photo be accepted into the training/validation/test sets? What if four of the participants

agree? You et al. decided that three was the appropriate cutoff. We decided to use four, in an attempt to improve on You's classification accuracy. After we filtered the dataset in this way, we were left with 13k images.

After the filtering, we were left with a class imbalance. Our largest category, contentment, had 3083 images, while our smallest category, fear, had just 486 images. For more on how we attempted to address this class imbalance, see the Methods section.

NOTE! If we want more space here, we can insert some images from the dataset.

4. Methods

References

[1] M. Huh, P. Agrawal, and A. A. Efros. What makes imagenet good for transfer learning? *CoRR*, abs/1608.08614, 2016.

[2] W. Jarrold and P. Z. Yeh. The social-emotional turing challenge. *AI Magazine*, 37(1), 2016.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. pages 1097–1105, 2012.

[4] B. Kurdi, S. Lozano, and M. Banaji. Introducing the open affective standardized image set (oasis). 49, 02 2016.

[5] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, pages 1717–1724, Washington, DC, USA, 2014. IEEE Computer Society.

[6] R. W. Picard, E. Vyzas, and J. Healey. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(10):1175–1191, Oct. 2001.

[7] H. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. J. Mollura, and R. M. Summers. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *CoRR*, abs/1602.03409, 2016.

[8] C. Szegedy, S. Ioffe, and V. Vanhoucke. Inception-v4, inception-resnet and the impact of residual connections on learning. *CoRR*, abs/1602.07261, 2016.

[9] M. Xie, N. Jean, M. Burke, D. B. Lobell, and S. Ermon. Transfer learning from deep features for remote sensing and poverty mapping. *CoRR*, abs/1510.00098, 2015.

[10] Q. You, J. Luo, H. Jin, and J. Yang. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. *CoRR*, abs/1605.02677, 2016.