Predicting Human Emotional Response to Images

Ashwin Sreenivas, Jessica Zhao
{ashwinsr, jesszhao}@stanford.edu

Stanford | ENGINEERING
Computer Science

INTRODUCTION

Research in emotion recognition has largely focused on identifying emotions from the human face. There is less literature on identifying human emotional response to images. For example, an image of a sunset may inspire awe, while an image of a graveyard may inspire fear. We are interested in predicting these emotional responses to images. Given an input image, we output 1 of 8 emotion classes: Amusement, Anger, Awe, Contentment, Disgust, Excitement, Fear, or Sadness.

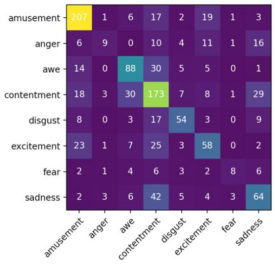
DATA

We use a dataset curated by You, Luo, Jin, and Yang through Amazon Mechanical Turk. The researchers presented an image with an emotion and Mechanical Turkers were asked if they agreed or disagreed. We take the majority label as ground truth (e.g. if 3 agreed and 2 disagreed on an image presented as "Awe", we label the image "Awe"). The dataset contains links to images on Flickr that we scraped. As the original dataset was curated in May 2016, some of the image links have expired. Moreover, we only use images with a positive classification. We also discovered duplicate image classifications; for those, we take the most salient label (least divisive based on number of agrees and disagrees). Thus out of the 89,319 original image links, we use 21,970.

RESULTS

Model	Train Accuracy	Dev Accuracy	Parameters		
			FC Layers	Learning Rate	Dropout
Search for FC layers	**55.6%**	**55.6%**	**8**	**1e-5**	**0.5**
	25.4%	27.7%	8x8	1e-5	0.5
	23.6%	23.0%	16x8	1e-5	0.5
	23.6%	23.0%	32x16x8	1e-5	0.5
Search for Learning Rate	21.6%	23.5%	8	1e-4	0.5
	59.0%	54.8%	8	1e-5	0.5
	56.3%	**54.9%**	**8**	**1e-6**	**0.5**
Search for Dropout	84.1%	57.9%	8	variable*	0.4
	77.5%	**59.8%**	**8**	**variable***	**0.5**
	72.1%	59.1%	8	variable*	0.6

Emotion	Prec.	Recall	F1	Train Count	Dev Count	Test Count
Amusement	73.9%	80.9%	77.2%	4244	251	256
Anger	50.0%	15.8%	24.0%	1088	46	57
Awe	61.1%	61.5%	61.3%	2657	146	143
Contentment	54.1%	64.3%	58.7%	4645	245	269
Disgust	65.1%	57.4%	61.0%	1398	78	94
Excitement	52.7%	48.7%	50.7%	2474	142	119
Fear	57.1%	25.0%	34.8%	874	46	32
Sadness	49.2%	49.6%	49.4%	2393	144	129
Overall	**60.1% test accuracy**			19773	1098	1099



Left to Right: Architecture Search, Emotions Test Statistics, Emotions Confusion Matrix
*combination of LRs in final model

DISCUSSION

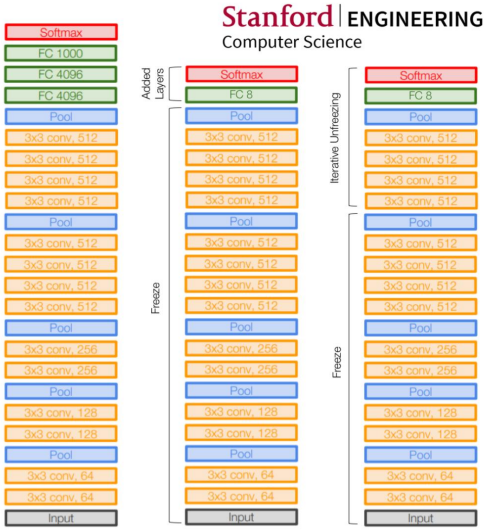
We perform manual error analysis and examine the misclassified images by class to tabulate trends. Globally, we notice there are challenges in extracting connotation from images and that some images can have multiple correct classifications. Most commonly, we notice:

Anger misclassified as Sadness: Out of the 16 misclassified, 11 have a non-unanimous vote and 7 have a strongly-divided vote. 11 are black & white, meaning our network may be associating monochromatism with sadness. Moreover, 9 could have been sadness based on human evaluation. Sadness is challenging to identify as it arrives in many forms not immediately apparent in a facial expression or. landscape A frown can evoke sadness; a pile of rubbish labelled "drugs" can also evoke sadness for the state of the world.

Sadness misclassified as Anger: The 3 misclassified images are all close-ups of faces that contain red, warm temperature tones, meaning our network could be associating red colors with anger.

Contentment misclassified as Sadness: Out of the 29 misclassified, 20 have a non-unanimous vote and 11 have a strongly-divided vote. We notice that 8 images are women looking away from the camera, meaning the network could be associating indirect facial gazes with sadness. 4 are faces of animals or statues, meaning the network is not as skilled as reading non-human faces.

Sadness misclassified as Contentment: Out of the 42 misclassified, 35 have a non-unanimous vote and 23 have a strongly-divided vote. 8 have nature backgrounds but feature a sad person, meaning the network could be associating outdoor landscapes with contentment but overlooking the person. Moreover, we notice 5 graveyard photos that the network misclassifies as contentment, suggesting that it has not learned the connotation of what a gravestone means to humans. We also notice 6 animal photos (horse, dog, and cat faces) that are misclassified, again suggesting the network's weakness in reading non-human faces. We also notice 5 images with text that clearly denote sadness, such as a story of a cat who lost a fight to cancer, suggesting our network is not learning from the text in images or there are too few images with text from which to learn.



1. VGG19 Base
2. Removing FC layers, adding our own
3. Iterative Unfreezing

MODEL

1. Transfer Learning Initialization
We use a VGG19 base model pre-trained on ImageNet and apply transfer learning to initialize our weights.

2. Training FC Layers
We remove the top 4 layers (the 3 FC layers and the softmax output) of VGG19 and add 1 FC layer of size 8 with a softmax activation. We freeze all layers except this and use Adam with LR=1e-4 and Categorical Cross Entropy Loss to train.

$$f(s)_i = \frac{e^{s_i}}{\sum_j^C e^{s_j}} \qquad CE = -\sum_i^C t_i log(f(s)_i)$$

3. Iterative Unfreezing
We unfreeze the top 4 convolutional layers one at a time and train them with the FC layers over 4 iterations. The first iteration uses Adam with LR= 1e-5; subsequent iterations use Adam with LR=1e-6. We then unfreeze the entire model and train overall using Adam w LR=1e-6.

FUTURE WORK

As humans rarely respond with singular emotions, we can allow for multiple types of emotional responses to images. We can extend our model to accommodate multi-hot encodings so images can be labelled, for example, both amusing and exciting. We can also train on more images with text to help the network better interpret language.