CS230

EmotiveLens: face detection with focal loss training criterion and small data

Kevin Song
Biomedical Informatics Graduate Program
Stanford University School of Medicine
kmsong@stanford.edu

Anthony Van
Department of Electrical Engineering
Stanford University
avan994@stanford.edu

Abstract

Computer-vision-based facial emotional classifiers have valuable applications in telemedicine (specifically, psychiatry and mental healthcare in the digital age). In addition, innovations from the field of computer vision have not been readily applied to the field of telemedicine. Here, we partially implemented a webcam-enabled emotional expression classifier by training a face detection algorithm on a small dataset of frontal face images (N = 450). Our trained RetinaNet convolutional neural network (composed of a ResNet CNN trained using a novel focal loss criterion) achieved a mean intersection over union training accuracy of 94.3%.

1 Introduction

Computer-vision-based facial emotional classifiers have valuable applications in telemedicine (specifically, psychiatry and mental healthcare in the digital age). In addition, computer vision innovations have not been readily applied to the field of telemedicine.

Rather than replacing human psychiatrists and mental healthcare professionals altogether, we initially proposed building a computer-vision-based classifier that would assist trained mental healthcare professionals in their abilities to accurately interpret their patients' affective states. In Japanese culture, it is said that individuals have "three faces": 1) one for their societies, 2) another for their family, and 3) one for just themselves. Current mental healthcare readily allows for professionals to easily interpret first-layer affective states. We had hoped that a computer-vision-aided classifier could augment healthcare professionals' abilities to peer into the second—and possibly the innermost third—affective states of individuals under their care.

Here, we partially implemented a webcam-enabled emotional expression classifier. In our project, we proposed building a webcam-enabled emotional expression classifier (along the happy/non-happy affective axis), with an emphasis on deep learning methods involving convolutional-neural-network-based (CNN-based) object detection (RetinaNet is a network equivalent to a ResNet CNN trained using a novel loss criterion) and transfer learning approaches (i.e., training a VGG-19 network with ImageNet: weights) [1]. Due to time constraints during the course of this project, we ended up finalizing only our object detector's implementation, which takes in images as inputs and returns detected, boundary-boxed faces as outputs.

We hope that our partial, open source implementation will exist as a proof-of-concept that leads to potential exploration and optimization of this technology space for future development.

CS230: Deep Learning, Winter 2018, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

2 Related Work

Innovations from the field of computer vision (i.e., object detection and deep transfer learning) have not been widely applied to the field of telemedicine (specifically, in this case, telepsychiatry). Startups such as Brain Power have previously applied machine learning to build wearable applications that combine semi-real-time emotional image classifiers with neurofeedback to train children to better recognize their family members' emotional expressions [2].

With respect to emotional and mental healthcare applications, the field of computer vision has made tremendous progress in building neural networks that accurately identify arousal and valence in a two-dimensional affective feature space [3]. The field of clinical informatics has not entirely caught on to such advances made by computer vision researchers. Nevertheless, several industrial and academic players have expressed interest in integrating accurate emotional classifiers for applications in telemedicine [4]. In the academic medical literature, one previously published method has enabled physicians to infer patients' emotional states (i.e., stress level, disappointment, euphoria, and neutral affect) through a support vector machine (SVM) classifier trained on physiological measurements (i.e., skin conductance, facial muscle activity, electrocardiogram signal, and respiratory rate) [5]. Rather than relying on physiological measurements obtained from not-always-perfectly-calibrated instruments, computer-vision-based algorithms may be better able to capture subtle nonlinearities and model complex representations of raw emotional expressions, thereby interpreting individuals' emotions more accurately and more conveniently than via physiology-based classifiers.

3 Dataset and Features

Object detection dataset: frontal face images (N = 450, each 896 x 592 pixels in jpeg format) were collected by Markus Weber at Caltech [6]. Color images of 27 unique people under various lighting, expressions, and backgrounds were collected. Images were previously annotated with individual faces' boundary box coordinates.

Figure 1: Example image taken from the Caltech face dataset.

Emotional classification dataset: though not included in our final implementation, we planned on passing detected faces to a binary emotional image classifier that would have been trained using the AR Face Database [7]. The AR Face Database (collected by Aleix Martinez and Robert Benavente at Ohio State) consists of over 4,000 color images (of 768 x 576 dimensions) from 126 people. Images were previously annotated according to the following feature labels: neutral expression, smile, anger, scream, lighting conditions, and presence/absence of occlusions.

4 Methods

Object detection task: we trained a RetinaNet model (i.e., a ResNet CNN trained using a focal loss criterion) on our object detection dataset. Focal loss was previously proposed by Lin, et al. at Facebook AI Research (FAIR) [1].

RetinaNet is a CNN-based model composed of several convolutional layers, and is based on the ResNet model previously formulated by the computer vision community. RetinaNet differs from previous ResNet implementations by incorporating an additional term in its loss criterion (from here on referred to as "focal loss"):

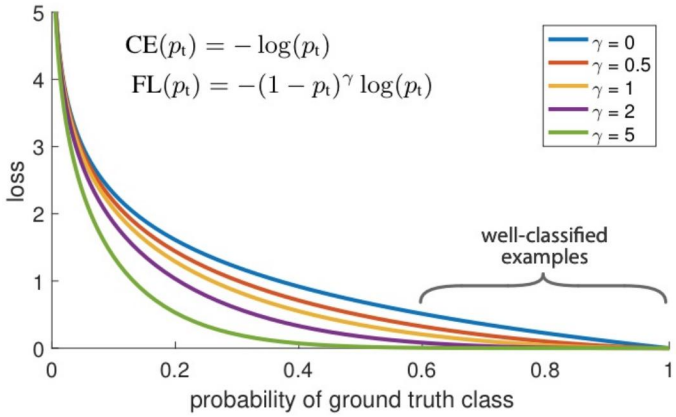

Figure 2: Focal loss. Our RetinaNet model was trained with an additional factor incorporated into the standard cross-entropy training criterion. As gamma increases, more focus is put on harder, misclassified examples [1].

Emotional classification task (not performed): we initially proposed training a VGG-19 CNN with ImageNet weights on our emotional classification dataset, freezing the first five layers. Training would have been performed one-versus-all so that labeled smiling images would have been considered as positive cases, and non-smiling images would have been considered as negative cases.

5 Experiments/Results/Discussion

Our ResNet model was trained across all 450 training images, for one epoch, with default parameter settings [8]. Our primary accuracy metric was the Intersection over Union (IoU) metric of predicted boundary boxes drawn on a subset of the training set. Our mean training IoU accuracy was 94.3%.

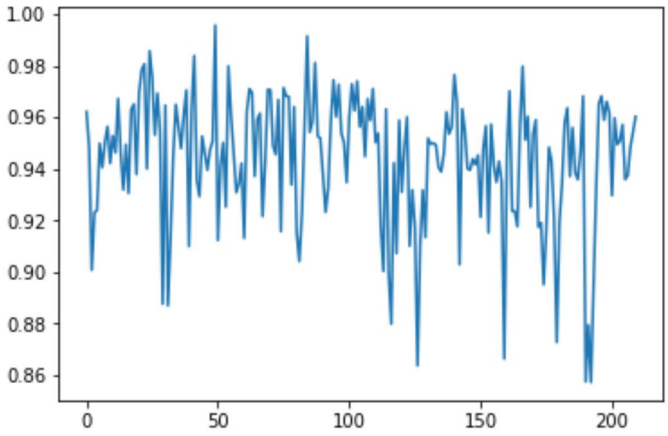

Figure 3: Plot of intersection over union accuracy for a subset of training examples. Our RetinaNet CNN object detector achieved a mean training accuracy of 94.%, a minimum training accuracy of 85.7%, and a maximum training accuracy of 99.5% (training set N = 450).

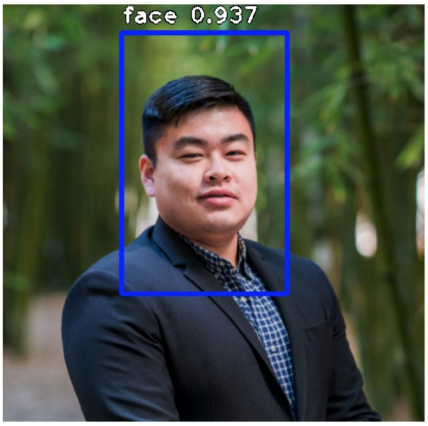

Figure 4: Boundary boxing. Our object detector draws boundary boxes around faces detected in images and assigns a probability of the boxed regions as belonging to the "face" class.

6 Conclusion / Future Work

Our object detector did not have extraordinarily great performance (i.e., 95-99% accuracy and higher), though it did exhibit a reasonably high training accuracy of 94.3%. As we did not have a large enough training set to conduct a train/test split, we did not validate our object detector's performance on a held-out test set. As such, it is quite possible that our object detector may be exhibiting a high degree of overfitting to the training set. In this case, reducing the flexibility of our model (e.g., via regularization methods) may reduce overfitting in future iterations of our model.

Logistically, our primary roadblock in our project was sourcing boundary-box-annotated data (that was also properly formatted) for loading into the RetinaNet model. A larger dataset was publicly available (N > 5000), though its boundary box annotations were presented as elliptical coordinates [9]. These coordinates, when trigonometrically converted to boundary box dimensions, did not lead to viable classifier performance. As our sample size was super-small, we did not expect high performance or accuracy with our trained model. However, our object detector's usage of a focal loss training criterion could have contributed to our non-abysmal observed training accuracy, given the small size of our training set.

Given another six months, access to more time, and larger, better data, we would have ideally been able to implement this project end-to-end. Computational power was not an issue for our project, though it may be a larger issue when training on larger datasets, or for more epochs. With larger, more representative training sets, we also would have been theoretically capable of achieving better model performance.

7 Contributions

Inventory of tasks *performed and completed* during the course of this project:

Project idea and methodology formulation **(Kevin)**

Writing of webcam streaming code **(Anthony)**

Annotations parsing (writing script to extract boundary box coordinates), training object detection model (on Amazon EC2 GPU-based instance) **(Kevin)**

Writing script that tests object detector and obtains an IoU accuracy metric **(Anthony)**

Annotations parsing of AR Face Database and training of VGG-19 network with ImageNet weights (transfer learning). We did not end up including this final model because it did not happen to produce correctly formatted outputs. **(Anthony)**

Creation of write-ups, poster, and paper **(Kevin)**

Acknowledgements

Special thanks to Andrew Ng for instruction and guidance, Kian Katanforoosh for help with project formulation and guidance, Olivier Moindrot for help with project formulation and execution, Ramtin Keramati and Patrick Cho for general instruction, and to the CS 230 teaching staff for general assistance.

References

[1] T.Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, "Focal Loss for Dense Object Detection," arXiv:1708.02002v2 [cs.CV], Feb. 2018.

[2] `http://www.brain-power.com`

[3] H.R. Kim, Y.S Kim, S.J. Kim and I.K. Lee, "Building Emotional Machines: Recognizing Image Emotions through Deep Neural Networks," arXiv:1705.07543v2 [cs.CV], Jul. 2017.

[4] `https://medcitynews.com/2015/07/emotion-recognition-telemedicine/`

[5] C.D. Katsis, G. Ganiatsas, and D.I. Fotiadis, "An integrated telemedicine platform for the assessment of affective physiological states," Diagnostic Pathology. 2006;1:16. doi:10.1186/1746-1596-1-16.

[6] `http://www.vision.caltech.edu/html-files/archive.html`

[7] `http://www2.ece.ohio-state.edu/~aleix/ARdatabase.html`

[8] `https://github.com/fizyr/keras-retinanet`

[9] `http://vis-www.cs.umass.edu/fddb/`

GitHub Repository

Code snippets and formatted annotation files available at `http://github.com/kevinmsong/emotivelens`.