CS230 Project:

"Who is This Person on Your Screen?" Classification of Celebrities Using CNN models

Yuzhou Wang

yuzhouw@stanford.edu

Graduate School of Business, Stanford University

Abstract

Identifying the celebrity in a movie/TV show has a big impact to help video platforms monetize its traffic and make the video watching experience more fun and interactive. In this project, I worked towards this goal by building a multi-class celebrity classification system that can recognize 6 different celebrities: Emma Watson, Nicole Kidman, David Beckham, Micheal Jordan, Michelle Obama, Barack Obama. I trained the model using transfer learning with 500 images for each of the celebrity with four different models pretained on the ImageNet dataset: InceptionV3, VGG19 and ResNet, SqueezeNet. ResNet produces the best results from my model search. I then conduct various hyperparameter tuning experiments, and use dropout regularization and data augmentation to overcome overfitting problems to produce a multi-class celebrity classifier with accuracy rate of 98.64%.

Introduction

When we are watching videos online, there are many moments we want to find out who is the actor/actress that is playing the role that we saw in a movie/TV show, but it is usually hard to find out this information (we can not Google it). What if we can just click on the face of the actor/actress and their name will be shown instantly? This will help making the video watching experience more interactive and extract more commercial value from the video (e.g., we can show ads that is relevant to this celebrity so the video platform can monetize on the traffic in addition to the pre-roll ads).

To predict output category from input images, I deployed transfer learning on InceptionV3, VGG19 and ResNet, SqueezeNet. I trained the models on 3000 images and compared performance across the four scenarios. ResNet performs the best. So, I tuned its hyperparameters to further improve accuracy and achieved 98.64% validation accuracy.

Related work

In 2016, Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao published a paper called "MS-Celeb 1M: Challenge of Recognizing One Million Celebrities in the Read World",[1] and this paper provided a measurement dataset to evaluate the performance of recognizing the one million celebrities, who are the real persons in the world and have/had public attentions. The measurement set is constructed by blending a set of carefully labeled images and a set of distractor images. The motivation of developing the dataset is to inspire more researchers to develop new algorithms to classify the face images. However, the paper does not provide any recommendation on which CNN models are the best for training on this dataset.

Dataset and features

Although Y. Guo. L. Zhang provided a large dataset of the celebrities in their paper " MS-Celeb-1M: Challenge of Recognizing One Million Celebrities in the Real World" [1] , I decided to not use this dataset for this project. The reason is 1) The number of training data for each celebrity is too small (only ~30 images per person), and this would be hard to achieve good accuracy of the algorithm. 2) The training images are of various sizes, which required a lot of cropping of the images if I want to have a training dataset of identical image size 2) The MS-Celeb-1M dataset is optimized for inputting one image and identify which celebrity it is from a long name list of celebrities (~1million celebrities), however, for this project, I decided to limit the scope of my research to write a CNN model that is able to classify only 6 different celebrities. The reason of reducing the scope is because I realized that my understanding of computer vision algorithm at this moment is not enough to handle a complex task of classifying 1 million celebrities, therefore, I decided to start with classification of 6 different celebrities for this project and expand the scope of this project in other classes about deep learning.

To collect the dataset necessary for model training, I used Google Image search to find the images of Emma Watson, Nicole Kidman, David Beckham, Michael Jordan, Michelle Obama, Barack Obama. The reason I choose these 6 celebrities for this project is because I want to test if my algorithm will be able to tell similar faces from each other. For example, based on my personal opinion, I find Emma Watson's face looks similar to Nicole Kidman's, while Michelle Obama's face looks similar to Barack Obama's. Therefore, such similarity would add difficulty to achieve certain accuracy of classification, and this makes the project more challenging and interesting. Additionally, I intentionally have chosen 3 celebrities that has white skin color and 3 celebrities that has black skin color, so that I can compare the result within each group and between each group and see if the skin color will make if easier/hard for machine to classify accurately.

Before setting up or training models, I cropped all images to size 200*200 pixels to align with the models' input requirements. The data was divided in train, dev and test: 1947 for training, 486 for dev and 533 for test.

Method

I first conducted a model search across the four families of the InceptionV3, VGG19 and ResNet, SqueezeNet, and then used hyperparameter tuning to optimize the model for the celebrity classifier.

The reasons for me to choose the four models to begin with are as follows:

For ResNet 50, I chose it because it can prevent the vanishing gradient problem and overfitting, and it is also the winner of ImageNet challenge in 2015. Therefore, it is pretty suitable for this task.

For Inception V3, I chose it because it's good for image classification tasks where the object to classify is of varying size, and this suitable for the context of this project, since not all the training images are cropped from the face and centered.

For SqueezeNet, I chose it because it can achieve very similar accuracy like ResNet but can be fit onto embedded devices due to its small size and faster computation time.

For VGG19, I chose it because it only has 19 layers (compared to the 50 layers that ResNet has), so it can reduce the chances of overfitting and the vanishing gradient problem. This improves the chance that the model gets better with each epoch.

Experiments/Results/Discussion

Phase 1: I trained and evaluated the four baseline models (InceptionV3, VGG19 and ResNet, SqueezeNet) on the training and dev/validation dataset. For each of these models, I start with each of these pretrained ImageNet models and then use transfer learning on the last 8 layers of each model for training. As shown in the results table below, XXX performed the best. When calculating accuracies, I used "Top-1" accuracy, counting an image as accurately predicted if the top returned softmax percentage matched with the input image.

Model	Training Accuracy	Training Loss	Validation Accuracy	Validation Loss
ResNet	0.9574	0.138	0.9864	0.02
Inception V3	0.9582	0.1878	0.9250	0.3321
VGG19	0.17	13.37	0.1828	13.17
SqueezeNet	0.1625	13.49	0.1719	13.34

Phase 2: After conducting model search and settling on ResNet, I ran a number of experiments to do hyperparameter tuning. I experimented with number of training layers, number of passes with variable numbers of epochs. Given the ResNet model had a problem with overfitting on the training set, I decided to experiment with various regularization techniques like dropout regularization and data augmentation in order to reduce the generality issues in my results. Also, dropout regularization helped reduce overfitting of the model, and data augmentation helps to improve the accuracy of the model without the need to collect additional data for training sets. After the hyperparameter tuning, the best performing model has 98.64% accuracy on the test set and the following features:

Error Analysis: I produced a confusion matrix for all the 6 celebrities I classified (test dataset). Examples are shown below. Based on the confusion matrix, it seems that the classifier systematically classify Michelle Obama as Michael Jordan, and classify Michael Jordan as David Beckham, and the error rate is much higher than other categories, the potential reasons for such systematically bias may be due to ethnical similarities (both Michael Jordan and Michelle Obama are African American).

		Predicted Class					
		David Beckham	Emma Watson	Nicole Kidman	Michael Jordan	Michelle Obama	Barack Obama
True Class	David Beckham	66	0	0	1	0	0
	Emma Watson	0	88	0	0	0	0
	Nicole Kidman	0	2	98	0	0	0
	Michael Jordan	4	1	0	91	0	2
	Michelle Obama	0	0	0	4	95	1
	Barack Obama	0	0	0	0	0	80

Conclusion/Future Work

Identifying celebrities in TV Shows and movies can improve the user experience of many video platforms like YouTube and Netflix, and with the rise of modern technologies like machine learning, such interactive experience for watching videos is not only a use case that exist in science fictions, but will become a reality. This paper is a baby step toward this exciting future, and with the transfer learning to train the 4 models, the best model is ResNet with validation accuracy of 98.64%.

For future work, I would like to do three types of development. First, to further improve accuracy, I would train the model on more data. In this project I only have 500 images for each of the 10 celebrities, and hopefully in the future I will be able to get at least 1000 images for 50 different celebrities. Second, the system can be extended to recognize multiple celebrities in an image rather than only one. That would entail a two step process of detection and then recognition. Lastly, right now I only deploy a simple accuracy and confusion matrix approach (top-1 score) to compare different models. For a better comparison, I would include top-3 score. Eventually, I hope to match the identified celebrity with their ad sponsors so that I can help the video platforms to further monetize their online traffic (imagine when you see David Beckham show up on the screen, and Budweiser's ad also show up because David was signed by Budweiser to promote the product for the brand).

Reference

[1] Y. Guo, L. Zhang, Y. Hu, X. He & J. Gao. (2016) "MS-Celeb-1M: Challenge of Recognizing One Million Celebrities in the Real World" , IS&T International Symposium on Electronic Imaging

[2] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2818-2826)

[3] (2017, March 23). Neural Network Architectures – Towards Data Science. Retrieved from https://towardsdatascience.com/neural-network-architectures-156e5bad51ba

[4] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large scale image recognition. arXiv preprint arXiv 1409.1556

[5] He, K., Zhang, X., Ren, S., & Sun, J (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778)

[6] Iandola, F., Han, S., Moskewicz, M., Ashraf, K., Dally, W., Keutzer, K (2016). SquuezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size.