Comparing Inception v3 to Xception Based on the Kaggle Contest: Yelp Restaurant Photo Classification Task

Sebastian Muhle
sebastian.muhle@tum.de

Arda Ozdere
arda.oezdere@tum.de

Haydar Sahin
haydar.sahin@tum.de

Arman Garip
arman.garip@tum.de

Project Proposal

1. Introduction

With our project, we want to compare the Inception v3 architecture to the Xception architecture, performance-wise.

As Francois Chollet showed in his Xception paper, while only marginally better on ImageNet, Xception[3] was 4.3% better on Google's internal dataset JFT. He suggests that "This may be due to the fact that Inception V3 was developed with a focus on ImageNet and may thus be by design over-fit to this specific task."

In our project, we want to use the Kaggle Yelp Restaurant Photo Classification[4] to compare both architectures and see how big the performance gap is on this task.

Additionally, as an optional task, we had the idea to use the generated multiple labels to write fake restaurant reviews using an RNN trained on a dataset of Yelp reviews.

1.1. Related Works

1.1.1 Yelp Restaurant Photo Classification

Since the Kaggle Yelp Restaurant Photo Classification was a Kaggle Competition there already exist some work and benchmarks. However, due to the fact that the challenge was finished in April 2016, the participants only used ResNet and Inception v3 architectures[2]. We didn't find any later examples on the internet of people using Xception architecture on this task. Also, this Kaggle Competition was part of a Stanford CS231 report[1] but they used VGG16.

1.1.2 RNNs for fake review generation

Researchers from the University of Chicago have already written a paper on how to write very convincing fake reviews for restaurants using RNNs[6]. In their method, they used specific information about the restaurant like the name of their dishes to produce the fake review. With our approach we want to use the generated multiple labels from our CNN to write these reviews. In this way, we hope we can use more specific information about the restaurant like it looks and interior to make the fake reviews even more convincing.

2. Dataset

We will be working on the dataset that is provided by the Kaggle competition at the link.

The dataset contains approximately 240,000 RGB images of 2000 different restaurants. Image count per restaurant varies from 1 to 3000 and the average image count for a restaurant is around 120. The restaurants have 9 different attributes as follows:

- 0: good_for_lunch
- 1: good_for_dinner
- 2: takes_reservations
- 3: outdoor_seating
- 4: is_expensive
- 5: has_alcohol
- 6: has_table_service
- 7: ambience_is_classy
- 8: good_for_kids

The dataset is very large, but some of these attributes may not be associated directly with the features that the images represent. Especially attributes 2 and 8 can cause some issues regarding this aspect.

Also for our optional task, we are planning to use the dataset provided in the following link.

The dataset includes detailed data about the business and the review. It contains different JSON objects that store

locations, descriptions, categories etc. for the businesses, whereas they store star rating, review text, user info, voting data etc. for the reviews. It consists of nearly 3 million reviews for approximately 100k businesses. These reviews are written by around 700k different users.

3. Methodology

In order to train our models to solve this problem, we will use transfer learning with pre-trained models of Inception v3[5] and Xception[3] models.

Since Inception v3 models were used by the participants of the challenge in April 2016, we will firstly try to have a similar score to the best scores of the challenge with Inception v3. Afterwards, we will train a Xception model, and compare its results with our results in Inception v3.

Since our problem is a multi-label classification problem, in order to evaluate our models we will use Mean F1 Score, which is basically mean of each F1 score of each label. Mean F1 Score was also used for Kaggle challenge. Moreover we will use a loss function for multi-label classification other than single label loss functions such as softmax or cross entropy loss.

For training, we need a GPUs with a lot of VRAM to get a decent batch size. We plan to use K80 in Google Cloud. They have 12 GB GDDR5 per GPU. From past experience in transfer learning with an Inception v3 model, we should fit a batch size of 16 images onto the card.

4. Outcome

The main outcome of the project is comparing the performance of two different architectures on a problem that is related to the daily life. The classification of these images on Yelp can save people some time choosing restaurants, where the technical aspect of the project can bring us a better perspective on how different architectures work on a large dataset like this one. Also the optional task of creating fake reviews presents a certain challenge to convincingly simulate human behavior, making the topic more interesting.

References

[1] P. Agrawal and R. Gupta. Miml learning with cnns: Yelp restaurant photo classification. 1

[2] K. Blog. Yelp restaurant photo classification, winner's interview: 1st place, dmitrii tsybulevskii. 1

[3] F. Chollet. Xception: Deep learning with depthwise separable convolutions. *CoRR*, abs/1610.02357, 2016. 1, 2

[4] Kaggle. Yelp restaurant photo classification. 1

[5] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015. 2

[6] Y. Yao, B. Viswanath, J. Cryan, H. Zheng, and B. Y. Zhao. Automated crowdturfing attacks and defenses in online review systems. *CoRR*, abs/1708.08151, 2017. 1