Emotion Identification Using Children Facial Images

Final project for BIS 557

Bo Qin

Emotion Identification Using Children Facial Images

Abstract

Facial expressions of emotion is a well studied topic in psychology and increasingly important topic in computer vision. Research community has put substantial efforts to build databases of emotional images but almost all of them contain exclusively adult faces. The rarity of children facial images proposes difficulty in applying deep learning techniques in developmental science. This project brings together three publicly available children facial images databases and fits deep learning models (CNN, VGG-16 and Inception V3) on them for classification task. Some modifications are made to the original models in order to reduce over fitting and generate appropriate results. Results show VGG-16 with a dropout layer best classifies the combined database of children facial images.

1. Introduction

Facial expression is a powerful body language that can convey emotions and intentions. Facial expression has been well studied in the field of psychology, where scientists tried to decipher the real meaning and patterns of facial expression in human. As early as the twentieth century, Ekman and Friesen (Ekman, 1972) discovered that the facial expressions associated with emotions are both universal and culture-specific. They identified six basic emotions based on cross-culture studes (Ekman & Friesen, 1980). These emotions have strong universality among humans. They are anger, disgust, fear, happiness, sadness and surprise. Utilizing the Facial Action Coding System (FACS), Ekman and Friesen were able to give a detailed description for each expression based on the facial muscle movement. For example, if the person raises his/her cheek and pull

the lip corner, then his/her facial expression is defined as happy. Admittedly, recent psychology researches suggest that there may be more universal facial expressions than previously mentioned, such as contempt. Some researches also argue against the strong universality of these facial expressions, claiming that they can be cultural specific too(Jack et al., 2012).

Even though the E&F model has its own limitations, it is still the most popular and straight forward criterion for studies concerning facial expressions of emotions, since it provides convenient categories and objective description for each categories. For the study of computer vision and deep learning, such convenience and objectivity is very important. There are many databases built based on the E&F model. The most famous one is called Affectnet which contains more than 1,000,000 facial images and around 450,000 of its images were labelled manually (Mollahosseini et al., 2019) . We now have a vast literature focusing on how to build deep learning models to classify these images based on the EF model. Some exceptional architectures include AlexNet, VGGNet, GoogleNet (or Inception V1) and ResNet. However, given the variability between human faces and the subtlety of facial expression of emotions, achieving a high accuracy rate for universal facial expression classification remain a difficult challenge.

In this project, I will consider only the facial images from children. Most of the publicly available databases, like Affectnet, are absent of children images. Research community has put a lot of efforts to build databases that concern only adult faces. There is a lack of exploitation for children faces. children faces are systematically different from adult faces by many general facial features like skin texture and chin shape. Extending research and models to contain stimuli from Children facial images may lead to possible application in developmental science. Utilizing the E&F model, I hope to be able to apply advanced architectures (VGG16 and InceptionV3) to a complied database that concerns only children facial images and compare their results. Transfer learning is a technique taught in class and I will use this technique to tackle this task.

2. Method

1. Databases

I will synthesize the following databases to form a larger database:

(a) The Dartmouth Database of Childrens Faces (Age between 6 - 18, mean 9.84) (Dalrymple et al., 2013)

(b) The NIMH Child Emotional Faces Picture Set (NIMH-ChEFS; age between 13- 17, mean 13.7) (Egger et al., 2011)

(c) The Child Affective Facial Expression (CAFE; age between 2 - 8, mean 5.5) (LoBue & Thrasher, 2015)

These databases contain high resolution images from children faces. All images have resolution higher than (1000,1000). NIMH-ChEFS only has 5 categories for facial expression of emotions (afraid, angry, happy, neutral and sad). The other two databases both contain these 5 categories. Therefore, I merged these databases on these 5 categories and truncated images from other categories. After the merging, there are 3266 images in the general database: Afraid(520); Angry(866); Happy(640); Neutral(662); Sad(378). All the image data are pre-processed to be (400,400) with RBG 3 color modules. We will further process each image to fit the requirement of each deep learning model during training. We randomly divide the full database into a train set and a validation set with a 4:1 ratio stratified by each category. The training set was also augmented during training using common data augmentation techniques including: shift, sheer, zoom, rotate and horizontal flip. Please note that since these databases contain personal information, I can not upload them to Github. One should contact original authors for access if interested. A selected sample is presented in Figure 1.

Figure 1: sample images

2. **4 layers CNN**

   For the purpose of comparison, we start with a straight forward neural network by stacking 4 convolutional layers and pooling layers. Figure 2 illustrates the architecture of this model.

3. **Transfer Learning**

   I will transfer the trained models to fit this database.

   (a) VGG-16 (Simonyan & Zisserman, 2015)

   Briefly mentioned in class, VGG-16 is an award winning model capable of classifying large image databases.VGG-16 has a simple structure as one can think of simply stacking 13 convolutional layers with 3 dense layers (Figure 3). This model achieves 92.7% top-5 test accuracy in ImageNet, which is a database of over 14 million images belonging to 1000 classes. In an 2015 paper, it achieves an 54.56% average accuracy rate in 7 emotions classification (Levi & Hassner, 2015). I will adapt this model by using all of its convolutional layers

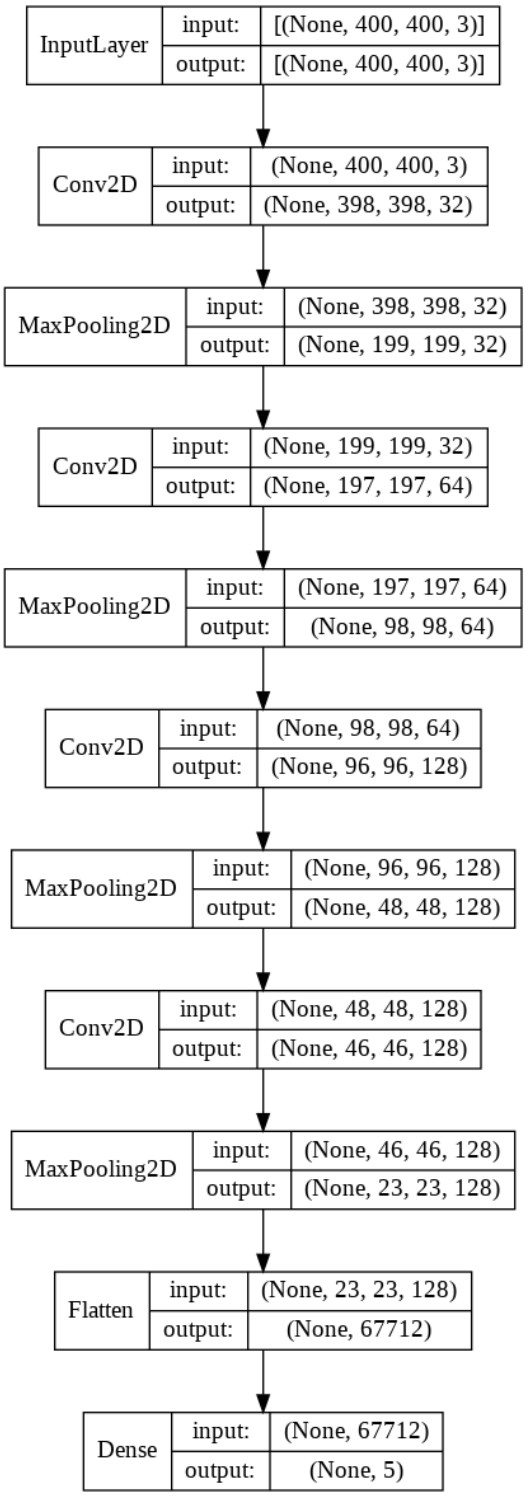

Figure 2: 4-layers CNN structure

plus an output layer using softmax activation function to classify the output into 5 classes.

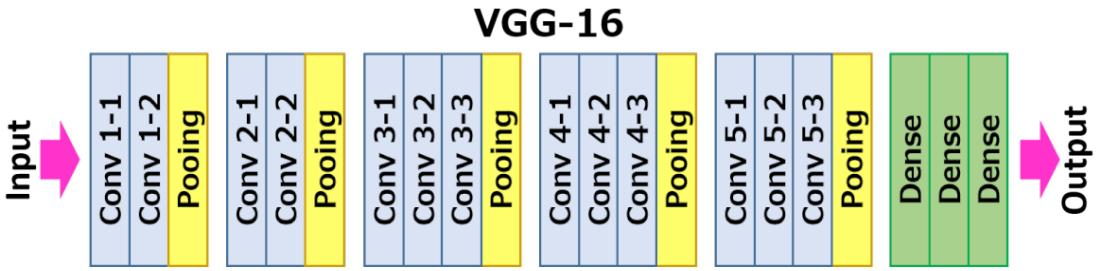

Figure 3: VGG 16 architecture (ul Hassan, 2018)

(b) VGG-16 + dropout layer

One common problem for fitting small datasets with complex neural network is overfitting.VGG-16 has 138 million parameters. Our database only has 3000 images with the smallest category 300 images, which leads to a high possibility of over fitting. Anticipating the problem, I will also try to add a dropout layer with a 0.75 probability of dropping each nodes right before the output layer.

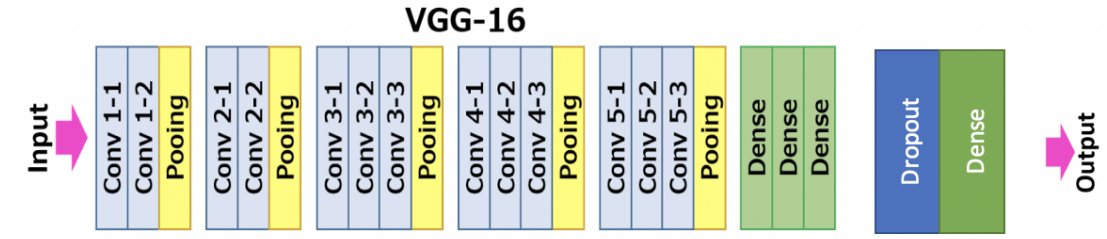

Figure 4: VGG 16 + dropout architecture

(c) Inception V3 (Szegedy et al., 2015)

I also want to try a different accomplished model just for comparison. To achieve this end, I choose Inception v3 which is a smaller

network with 24 million parameters. It has shown superior performance than VGG-16 in classifying the ImageNet. The structure for this model is fairly complex as it introduces a new type of layers called Inception layers. It achieves higher efficiency in optimization and computation by (1) factorizing convolutional net (e.g. connect three 3x3 convolutional nets with one 1x1 convolutional net to replace 5x5 convolutional nets) (2) efficient grid size reduction (e.g. reduce input dimensions by using convolution and pooling simultaneously). Figure 5 illustrates the architecture of Inception V3. Each cluster of points represents an inception layer. Detailed exploitation of inception layers exceeds the scope of this project. Interested researchers should read the original paper by Christian Szegedy.

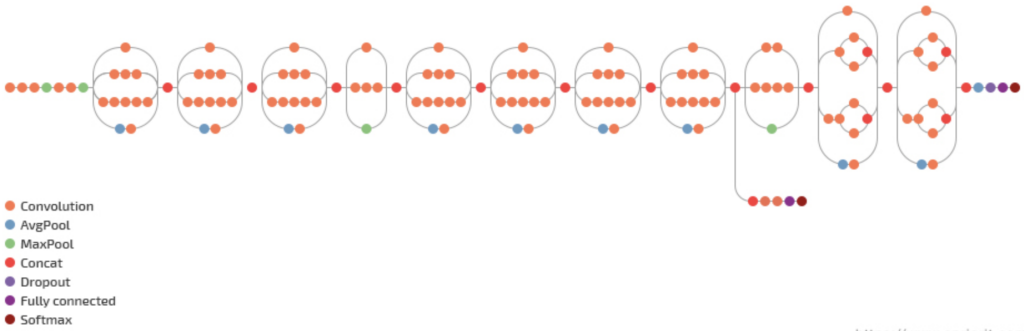

Figure 5: Inception V3 architecture (Boyko, 2019)

4. **Computing resources and Code**

   All of the codes were written in Python language and models were built primarily through Tensorflow API. Since most of the models require extensive training, I leveraged Yale high performance computing (HPC) clusters for training and data storage. Codes were recompiled into a jupyter notebook document attached to the project package. Bash scripts for HPC are available upon requests.

3. Results

I have tried 4 models to fit this dataset. The accuracy rate and training progress for each model will be presented below.

1. **4 layers CNN**(Figure 6)

   In 50 epochs of training, 4-layers CNN has a highest average training accu-

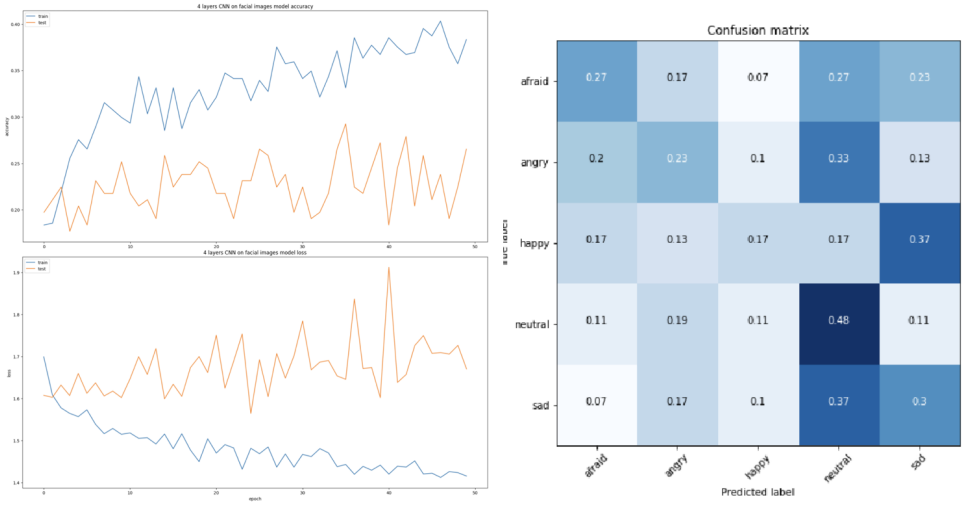

Figure 6: 4 layers CNN. Top left figure tracks the average accuracy rate; bottom left figure tracks the average loss; right figure is the confusion matrix with specific classification rate. Darker the blue higher the rate

racy of 40%, highest average validation accuracy of 28%, average training loss of 1.54 and average validation loss of 1.72. The confusion matrix shows that the highest per class validation accuracy achieved for neutral category (46%) and the lowest for happy category (13%). Since we only have 5 classes in this data set, a 28% out of sample accuracy is far from desirable. The gap between training accuracy and testing accuracy also suggests that this model does not have good generalizability.

2. **VGG-16**(Figure 7)

In 50 epochs of training, pure VGG-16 has a highest average training

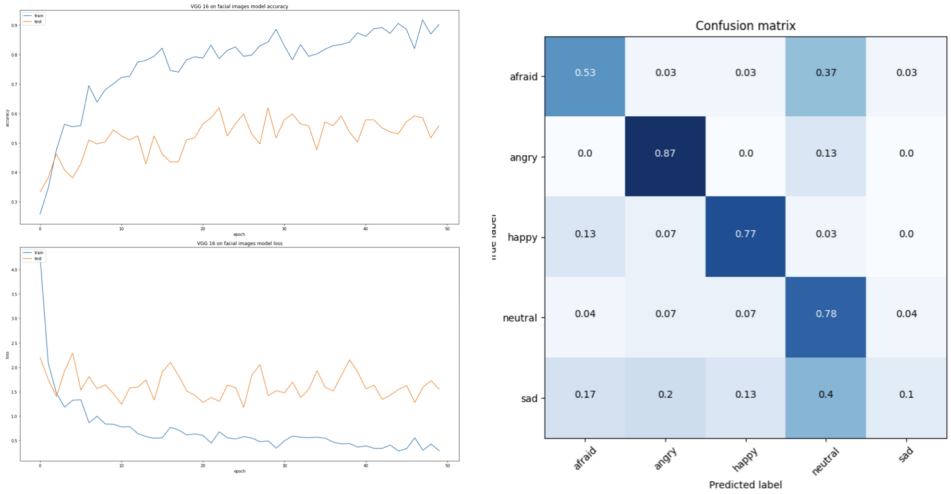

Figure 7: VGG-16 with training details and confusion matrix

accuracy of 92%, highest average validation accuracy of 60%, average training loss of 0.54 and average validation loss of 1.54. The confusion matrix shows the highest per class validation accuracy achieved for happy category (87%) and the lowest for sad category (10%). VGG performs well as hypothesized. However, it makes significantly less inference in the sad category as compared with other categories. Additionally, there is a 30% gap between training and validation accuracy, which suggests that over fitting exists.

3. **VGG-16 + dropout layer**(Figure 8)

   In 50 epochs of training, VGG-16 with a dropout layer has a highest average training accuracy of 68%, highest average validation accuracy of 62%, average training loss of 0.54 and average validation loss of 1.54. The confusion matrix shows the highest per class validation accuracy achieved for happy category (90%) and the lowest for afraid category (2%). After adding a dropout layer, the training accuracy and validation accuracy

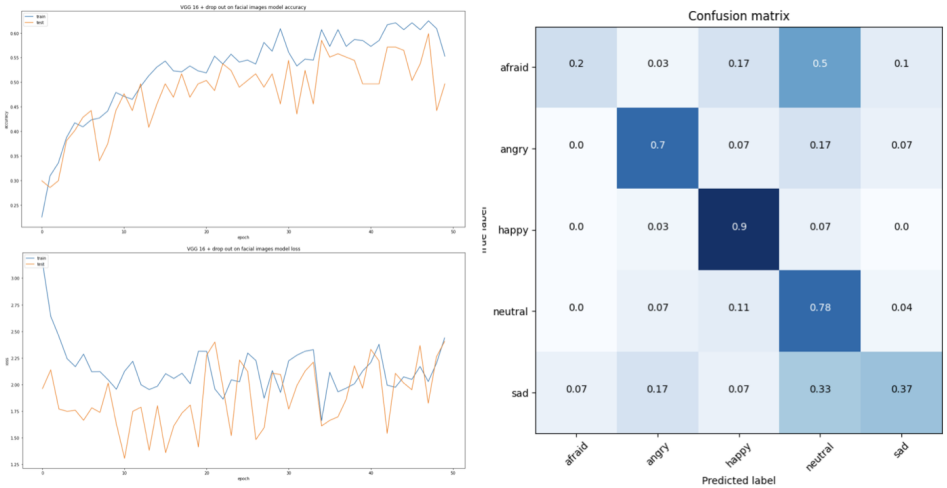

Figure 8: VGG-16 + dropout layer with training details and confusion matrix

trace each other closely. However, the training accuracy has been significantly reduced as compared with the pure VGG-16 model.

4. **Inception V3**(Figure 9)

   In 50 epochs of training, pure VGG-16 has a highest average training accuracy of 25%, highest average validation accuracy of 28%, average training loss of 3.24 and average validation loss of $2 * 10^5$. The confusion matrix shows the highest per class validation accuracy achieved for neutral category (48%) and made very few inference on angry, happy and sad categories.

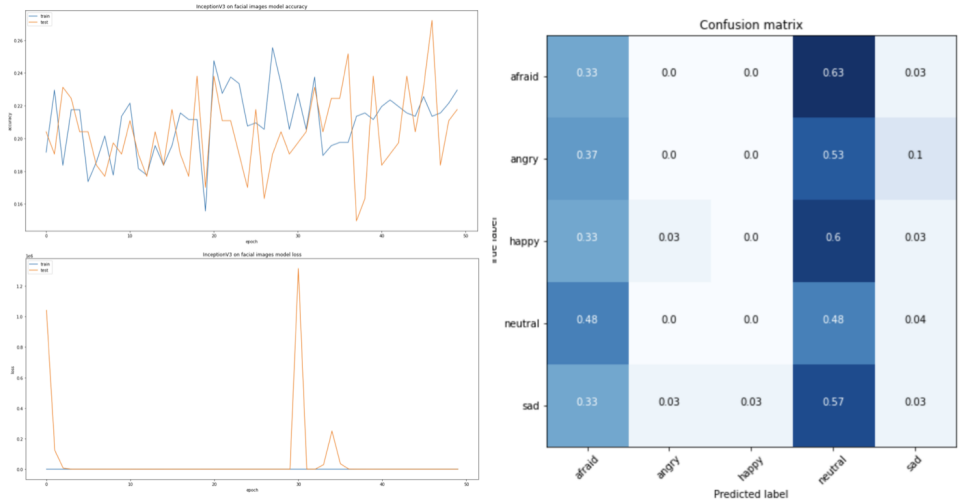

Figure 9: Inception V3 with training details and confusion matrix

Discussion

Of the 4 models built, we see that VGG-16 with a dropout layer has the best overall out of sample performance. Pure VGG-16 model has better in sample prediction accuracy but exhibits over fitting behaviors. A general 4 convolutional layers neural network does not seem to be able to capture the complex dynamics between facial expression of emotions. And to my surprise, the Inception V3 did not demonstrate learning behavior and showed worse performance than 4 layers CNN.

This project utilized a synthesized database to achieve the goal of finding a deep learning model for children facial expressions classification. By doing this, we greatly enhanced the opportunity of learning for our deep learning models. Unfortunately, the combined database does not present a balanced distribution among categories. For example the sad category only constitutes 10% of the total data set. The imbalance among categories may explain why our models share high validation accuracy rate in neutral category but low accuracy rate in the sad category.

The four models presented in this project can serve as a good starting point for the further exploitation of better architecture or hyperparamters tuning. Image classification is generally difficult as the data points are often lie in a high dimensional space. It is necessary to experiment and learn through trials and errors. This project shows that complex deep learning models may perform utterly different when presented with different dataset. Inception V3 is preferred over VGG-16 when classifying ImageNet data. However, its ability is fairly limited when classifying smaller datasets like our children facial expressions data sets. We saw from the results section that Inception V3 primarily make inference on afraid and neutral categories. Further research can investigate why this pattern happens. Is this a result of optimization or unsatisfactory parameters tuning? We also see that adding drop out layer reduces overfitting of VGG-16 but also significantly reduces the training accuracy. This is not a desirable solution. One possible explanation is that the position of our dropout layer is very close to the output layer, making our model harder to adjust for the effect of missing nodes. Further research should also look into adding drop out layers among earlier layers of the model.

This concludes my project. Thank you for reading!

References

Boyko, S. (2019). Using Modified Inception V3 CNN for Video Processing and Video Classification. See also URL `https://www.apriorit.com/+`.

Dalrymple, K. A., Gomez, J., & Duchaine, B. (2013). The dartmouth database of children's faces: Acquisition and validation of a new face stimulus set. *PLoS ONE*, (p. 11). URL: `https://doi.org/10.1371/journal.pone.0079131`.

Egger, H. L., Pine, D. S., Nelson, E., Leibenluft, E., Ernst, M., Towbin, K. E., & Angold, A. (2011). The nimh child emotional faces picture set (nimh-chefs): a new set of children's facial emotion stimuli. *Int J Methods Psychiatr Res*, (pp. 145–56). doi:`10.1002/mpr.343`.

Ekman, P. (1972). Universals and cultural differences in facial expressions of emotion. *Nebraska Symposium on Motivation*, *19*.

Ekman, P., & Friesen, W. V. (1980). Facial signs of emotional experience. *Journal of Personality and Social Psychology*, *39*.

ul Hassan, M. (2018). VGG16 – Convolutional Network for Classification and Detection. See also URL `https://neurohive.io/en/popular-networks/vgg16/+`.

Jack, R. E., Garrod, O. G. B., Yu, H., Caldara, R., & Schyns, P. G. (2012). Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, *109*, 7241–7244. URL: `https://www.pnas.org/content/109/19/7241`. doi:`10.1073/pnas.1200155109`. `arXiv:https://www.pnas.org/content/109/19/7241.full.pdf`.

Levi, G., & Hassner, T. (2015). Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In *Proc. ACM International Conference on Multimodal Interaction (ICMI)*.

LoBue, V., & Thrasher, C. (2015). The child affective facial expression (cafe) set: validity and reliability from untrained adults. *Frontiers in Psychology*, *5*, 1532. URL: `https://www.frontiersin.org/article/10.3389/fpsyg.2014.01532`. doi:`10.3389/fpsyg.2014.01532`.

Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2019). Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, *10*, 18–31. URL: `http://dx.doi.org/10.1109/TAFFC.2017.2740923`. doi:`10.1109/taffc.2017.2740923`.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. `arXiv:1409.1556`.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2015). Rethinking the inception architecture for computer vision. `arXiv:1512.00567`.