Chris Kaschner 2016-08-8

I. Definition

Project Overview

An estimated 2 trillion photos were shared online in 2015, with that number expected to continue to grow. Lacking efficient methods to examine and identify objects in these images they can only be understood by the text that is used to summarize or tage them not the actual image content.

Being able to identify different aspects of an image would provide considerable value to a brand seeking to connect with its users. This type of structured data is already available for text but what about images? There is significant value in a brand or company being able to understand when and where it's products and services are being shown in images online.

Is there a way to be able to identify the objects in an image to be able to know what brands are present in an image? What if a customer had an issue and posted an image without your company name in the text description? How would you could you identify and find such images?

In this project I created a Convolutional Neural Network that is capable of identifying brands in untagged/ unlabeled photos from a social media feed. The model I use implements a previously trained network and transfer learning to speed training. This project was inspired by a number of different sources.

Problem Statement

Can you teach a computer to recognize the brand logos of Nike and Altra in an unlabeled feed of images from Instagram?

In order to solve this problem I undertake the following:

1. Identify suitable existing Neural Networks trained for image recognition and classification
2. Implement the existing models locally and load pretrained weights
3. Obtain a dataset of images to retrain the network on the brands of interest
4. Reconfigure the final layers of the pre-existing models to classify the brands
5. Retrain the final layers of the new network using my dataset
6. Measure resulting performance
7. Optimize the model based on results

This project solves a classification problem. In its first incarnation discerning between 2 brands represents a *supervised binary classification* problem. In this case, selecting either Nike or Altra products are present.

Later, when a third category is introduced it becomes a *supervised multiclass classification* problem. This website offers a great layman's introduction to these concepts.

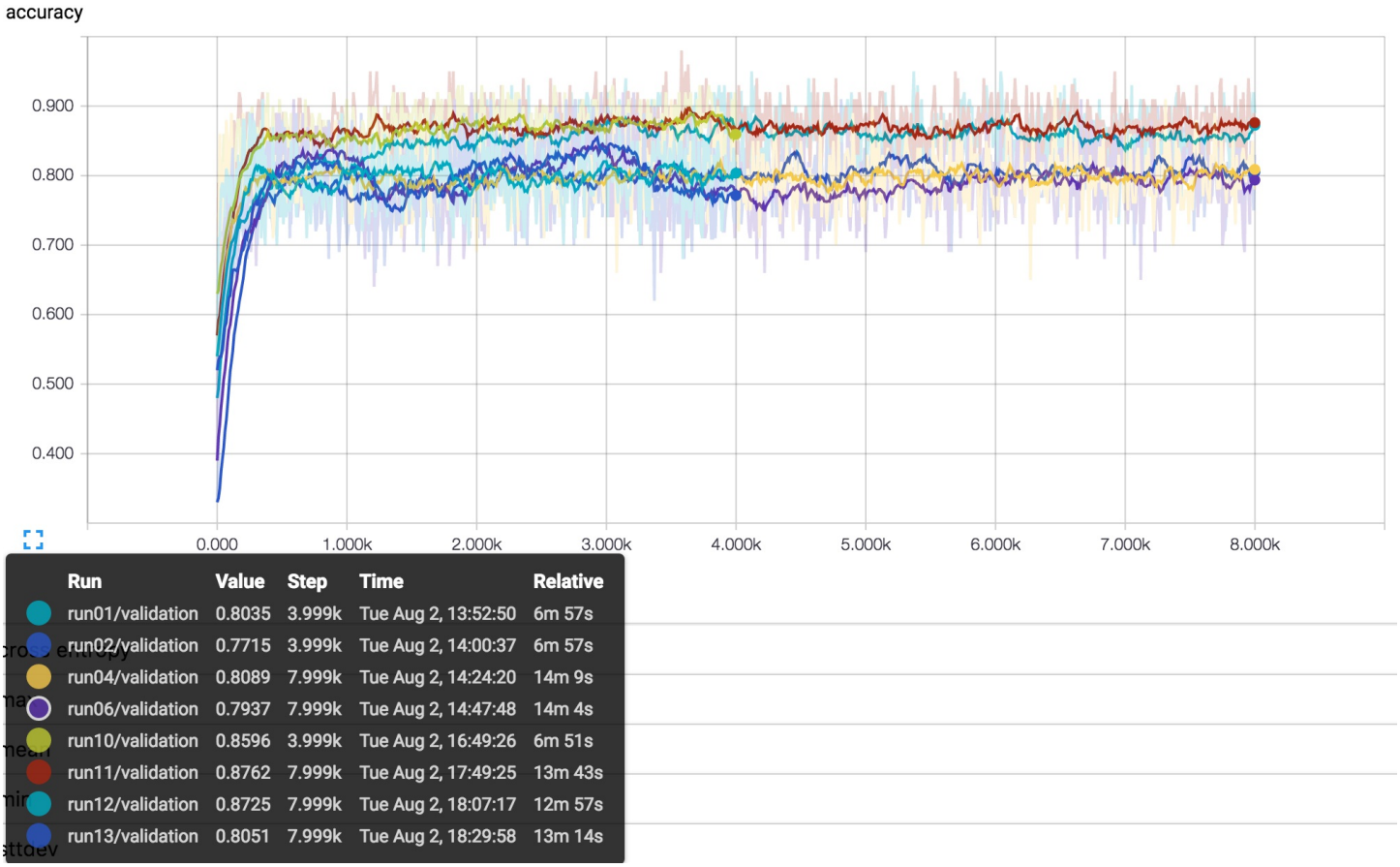
Metrics

There are 2 metrics I use to gauge the performance of my model: accuracy and cross-entropy.

Accuracy

The accuracy is the primary measure of overall model performance. It is represented as a percentage of the number of images the model correctly identified divided by the total number of attempts.

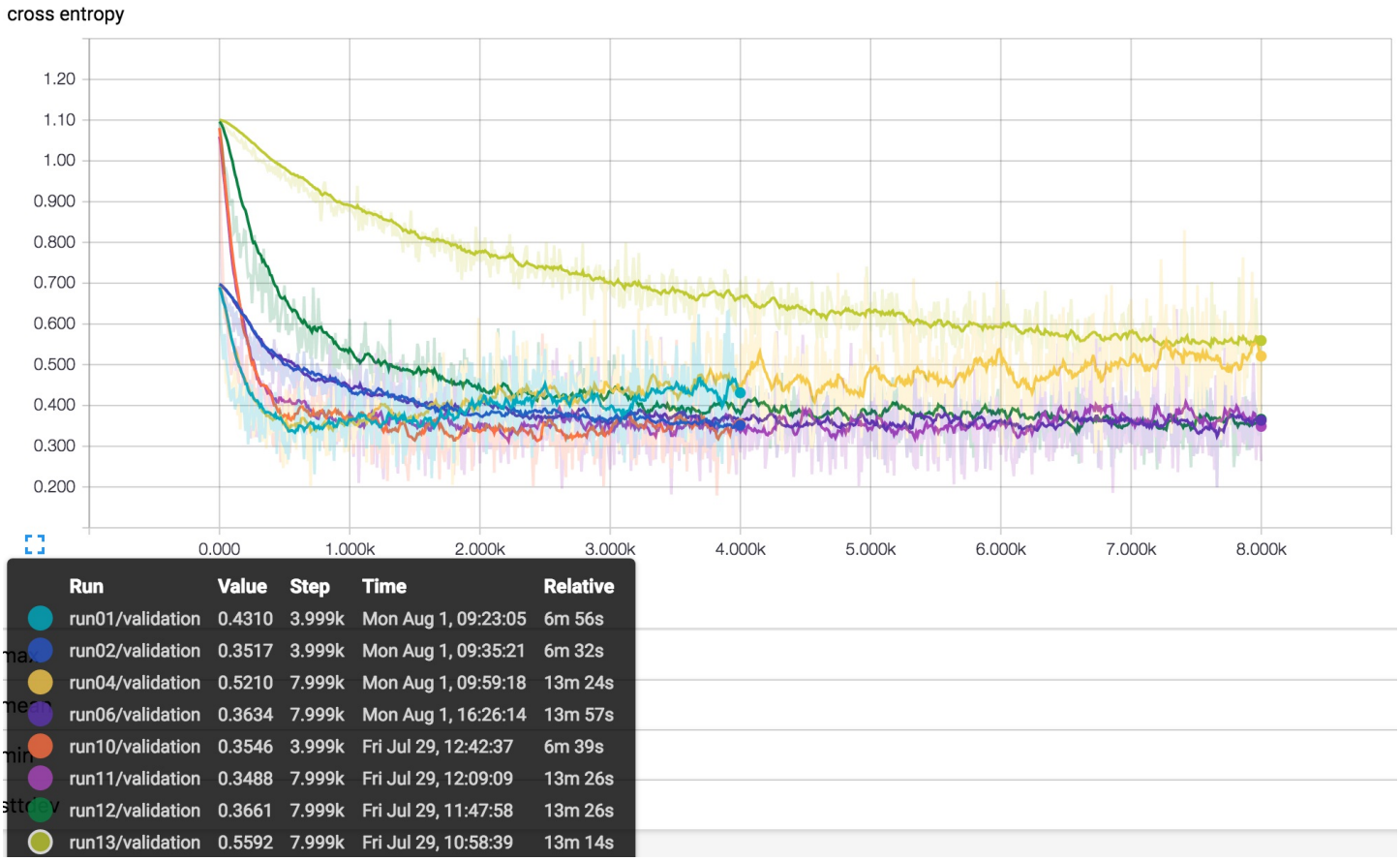
Accuracy should improve as our model iterates, as shown below:



Cross-Entropy

Cross-entropy is a measure of a model's loss or cost. A high value represents a model doing a poor job. We seek to minimize the cross-entropy. A more in-depth discussion of cross-entropy is available here

In general, we expect to see cross-entropy decreasing as our model iterates, like this:



Run	Value	Step	Time	Relative
run01/validation	0.4310	3.999k	Mon Aug 1, 09:23:05	6m 56s
run02/validation	0.3517	3.999k	Mon Aug 1, 09:35:21	6m 32s
run04/validation	0.5210	7.999k	Mon Aug 1, 09:59:18	13m 24s
run06/validation	0.3634	7.999k	Mon Aug 1, 16:26:14	13m 57s
run10/validation	0.3546	3.999k	Fri Jul 29, 12:42:37	6m 39s
run11/validation	0.3488	7.999k	Fri Jul 29, 12:09:09	13m 26s
run12/validation	0.3661	7.999k	Fri Jul 29, 11:47:58	13m 26s
run13/validation	0.5592	7.999k	Fri Jul 29, 10:58:39	13m 14s

By using these 2 metrics we'll be able to quantify how well our model is performing via the accuracy but also how well the model is *learning* via the cross entropy.

II. Analysis

Data Exploration

2 datasets were used in this project. The majority of the images were downloaded from Instagram and represent the typical input image that could be expected from a direct feed of images flowing into Instagram at any given time. Tags were used to search for the Altra & Nike images. An example search and results can be seen here.

Examples of Altra and Nike logos

First data set

Altra - 110 images (no cropped/ transformed images) Nike - 470 images (200 full size and 270 cropped images)

Second data set

Altra - 481 images (240 full size and 241 crops) Nike - 608 images (260 full size and 348 cropped images) Neither - 453 images (452 full size)

Similar transformation were applied to this data set for cropping images. They were applied to both the Altra and Nike data sets.

The 'neither' category of images is comprised of downloading the bulk "firehose" of images from Instagram.

To supplement this data set, as well as test my models' performance on perturbations in input image shape, approximately 10% of the images are from alternative sources such as Pinboard, Facebook, or Google Image Search. Thus not all of the input images are square, as would be expected from Instagram alone, and their size varies.

Examples of difficult images

Here are some examples and explanations of difficult images in the dataset.

Logos tend to be relatively small in actual images

Logos can be occluded by other objects in the scene

Both brands of interest could be in the same photo

Logos can be morphed to match product designs

Multiple brands in a single image

Neither brand of interest is present

Algorithms and Techniques

[### todo fix this so it doesn't suck]

1. CNN
2. Existing Network
3. Transfer learning
4. Retraining

The core of this project is the use of a Convolutional Neural Network.

CNNs and you, "No one teaches a child how to see"

Instead of defining algorithms to explicitly define every permutation of logo that we would potentially see,

machine learning and neural networks allow us to feed training images to a and allow it to define for itself what the critical features are.

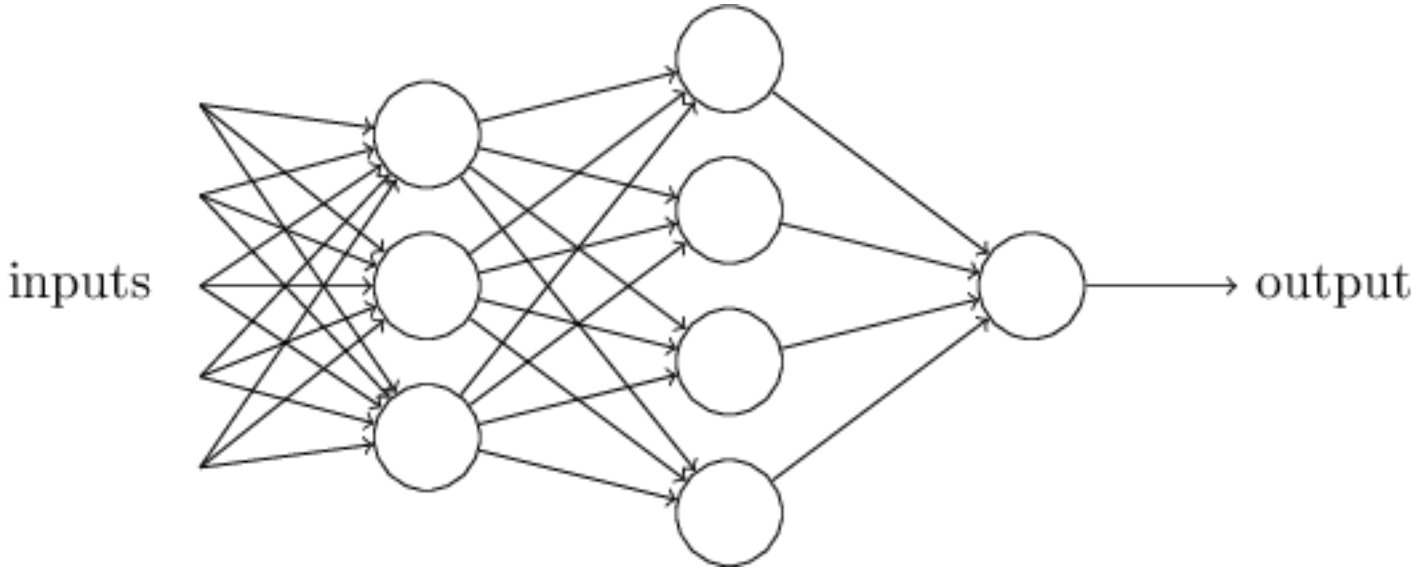
There's clearly a lot of information here to be learned and a variety of resources, this is a good place to start.

Convolutional neural networks are a subtype of neural network and have been shown to be effective at image classification.

ILSVR or ImageNet are both examples where CNNs achieve great performance and their performance continues improving.

They approach, or occasionally exceed, human performance on similar benchmarks.
<!-- The first building block used in my solution was a neural network and deep learning.

[###todo] Explanation of a network

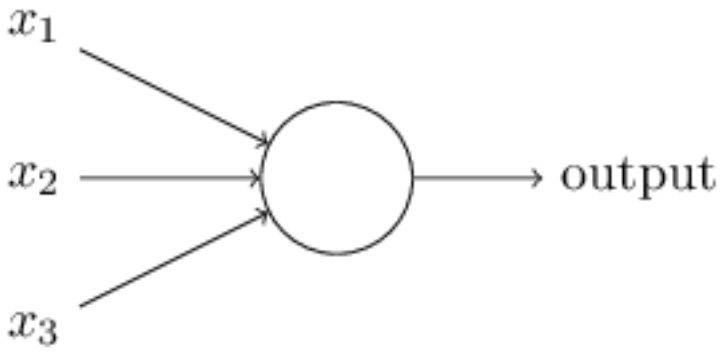


Convolutional neural networks are at the heart of these models. A way to reduce the dimensionality of the mathematics/ reduce computational complexity. Likely to have overfitting in convolved features. Pooling/ aggregating over a convolved feature, depends on mean/ max pooling

image layer -> convolved feature layer ->

Explanation of a perceptron [###todo]

Basic unit of a neural network is a neuron like node. It takes inputs and uses them along with weights to



determine whether to "fire" its output or not.

$$\text{output} = \begin{cases} 0 & \text{if } \sum_j w_j x_j \leq \text{ threshold} \\ 1 & \text{if } \sum_j w_j x_j > \text{ threshold} \end{cases}$$

Layers - > input, hidden, output layers --> <!-- [###todo images that are difficult for humans to differentiate but computers get correct]

Breaking it down: [###todo - images of each section of the graph]

1. Input layers - decoding, resizing, flattening, etc to prepare images for the network
2. Decode the jpeg file, resize to a standardized size/ shape of 299x299x3 representing 299 pixels wide by 299 pixels tall by 3 layers (red, green, blue) deep
3. Convolutional layer
4. Pooling layer
5. Mixed layer
6. Softmax
7. Output layers -->

Existing networks

In the case of the inception v3 network would take about 2 weeks to train on 8x GPUs (~$3k Nvidia K40s). Once trained the model and associated weights can be transferred to a smaller/ less expensive computer (in this case a MacbookPro) and used to classify images. The Inception-v3 model has over 27 million parameters and runs 5.7 billion floating point operations per inference

This approach allows us to "leverage" the resources/ computing power/ model complexity available to Google but used in our local environment.

The method I used to approach this problem utilized the significant work that large companies and research institutions have invested into this problem. I used 2 different networks to attempt this, VGG16 and Inception

v3.

The VGG model was created by the Visual Geometry Group (VGG) at Oxford University and was first described in 2015. Although there are multiple versions in , the one used in this work is 16-layers.

The Inception architecture first appeared in a model called "GoogLeNet" in 2014. Version 3 was described in 2015 and is the model used in this work.

Transfer Learning Pre-trained networks

Of particular interest to this project is transfer learning.

Transfer learning is the ability to train a neural network in one scenario, say at Google in Mountain View, and then downloading the model and its associated weights to be reused in a different location/ time such as Austin, TX.

Additionally, pre-trained networks can be used. They have the ability to distinguish important features in images despite being trained on images that differ from the images that will be used in our implementation of the network.

The ability to distinguish edges in images is universally applicable to classification tasks.

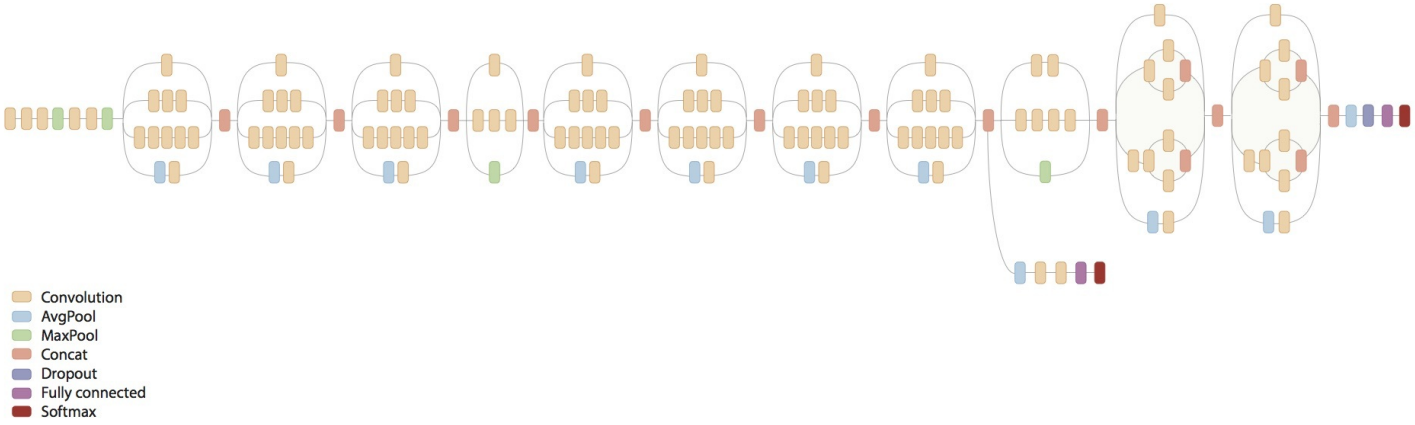
Layers of Inception- Pooling, 2D Convolution, etc.

1. Pooling layers look like this [###todo add more layer info]
2. Conv2D layers look like this

[###todo image of first few layers of a CNN]

Replacing the final layer of a network and replacing it with our preferred classifications allows us to leverage the power of a network that may have been trained for weeks on an array of high-powered GPUs and retraining it in a few minutes for our task.

This is what the Inception v3 network looks like:



1. Identify a suitable framework or library to build a pre-trained network

   There are a number of different options for building and and loading existing network architectures. Caffe, Keras, and TensorFlow are all excellent options.

   For the purpose of my project I limited my choices to Keras & TensorFlow. Their descriptions from their developers are provided below:

   ## Keras

   > Keras is a minimalist, highly modular neural networks library, written in Python and capable of running on top of either TensorFlow or Theano. It was developed with a focus on enabling fast experimentation. Being able to go from idea to result with the least possible delay is key to doing good research.

   ## TensorFlow

   > TensorFlow is an open source software library for numerical computation using data flow graphs. Nodes in the graph represent mathematical operations, while the graph edges represent the multidimensional data arrays (tensors) that flow between them. This flexible architecture lets you deploy computation to one or more CPUs or GPUs in a desktop, server, or mobile device without rewriting code. TensorFlow also includes TensorBoard, a data visualization toolkit. TensorFlow was originally developed by researchers and engineers working on the Google Brain team within Google's Machine Intelligence research organization for the purposes of conducting machine learning and deep neural networks research

2. Construct an existing network

   Keras provides a tutorial on loading and utilizing the VGG16 architecture.

TensorFlow provides instructions for the reuse of the Inception v3 network.

3. Load pre-trained weights into the previously built network

   The weights that effectively "remember" the training of a neural network are loaded. In some cases these weight files can be quite large. For the VGG16 model the weight file topped 550 mb

4. Modify the pre-existing model for our classification tasks

   Remove the original model's top layers and replace with ones suitable to perform our classification tasks.

5. Update the bottlenecks

   the penultimate layers of the model that perform the actual classification.

6. Retrain the network on my images of interest

   I investigate 2 options- binary classification or a closed world problem where images of only Nike or Altra shoes are expected and and open world multiclass classification of images that could be of Nike or Altra shoes or some third category I've cleverly titled 'Neither'.

7. Evaluate model performance

   Using training, validation, and testing values for accuracy and cross-entropy

8. Optimize hyperparameters and compare performance

1. Select the optimum solution

2. download existing pre-trained model and weights Keras and TensorFlow both offer frameworks for creating the CNNs used, transferring weights, and popping layers to retrain.

3. remove/pop the final layers that are used to classify ImageNet images (1000 classes)
4. Replace final layers with ones that suit my needs, in this case 3 classes
5. Retrain the network on images that I provide
6. Measure performance and vary hyperparameters seeking to maximize the models performance

Splitting Data into Training, Validation, and Test sets

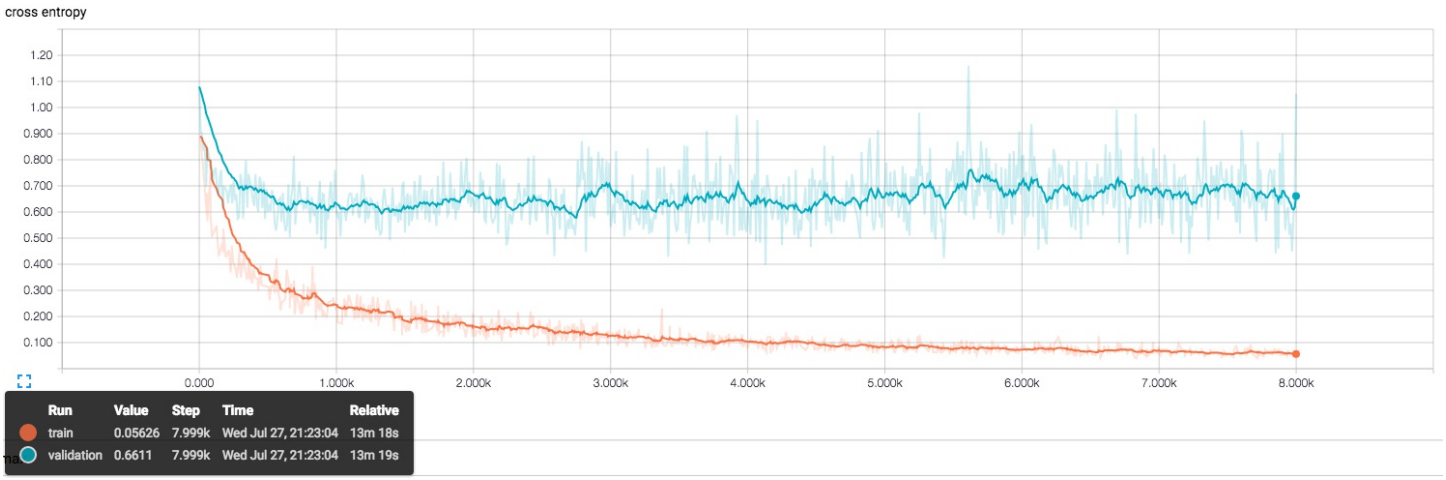
To reduce and mitigate overfitting, we split our images into 3 categories.

1. Training Data - 80% of the images are used for the model to learn from
2. Validation Data - 10% of the images are used for periodic "checking" of our model during training
3. Test Data - 10% of the images are used only once to predict real world results of our model against

images it has never seen.

Training and Validation information can be viewed during training runs via the TensorBoard application and are shown throughout this report. The test accuracy is reported as a single percentage datapoint at the conclusion of a successful training run of the model.

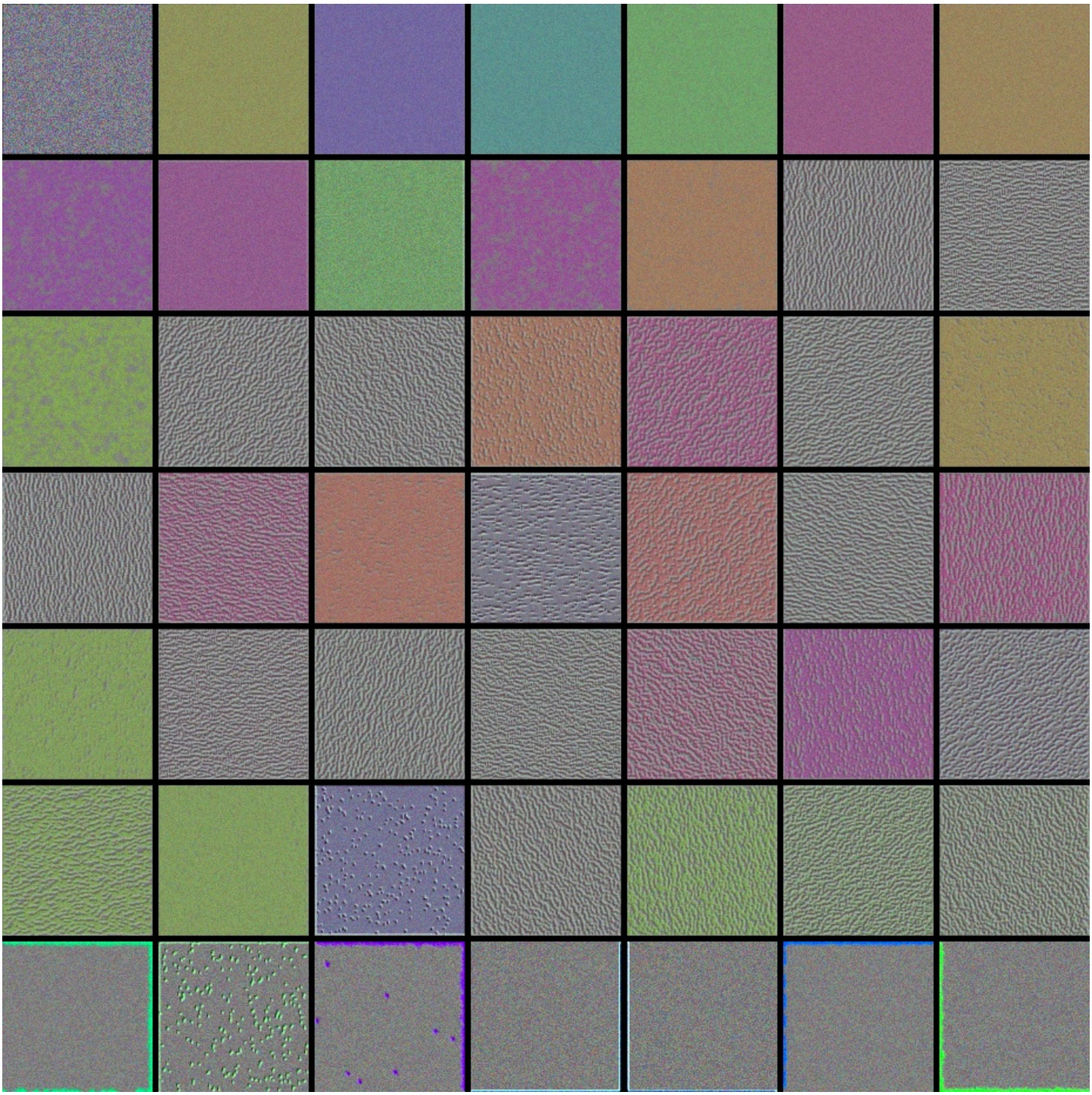
Overfitting, a model's "over-learning" of features in the training set.



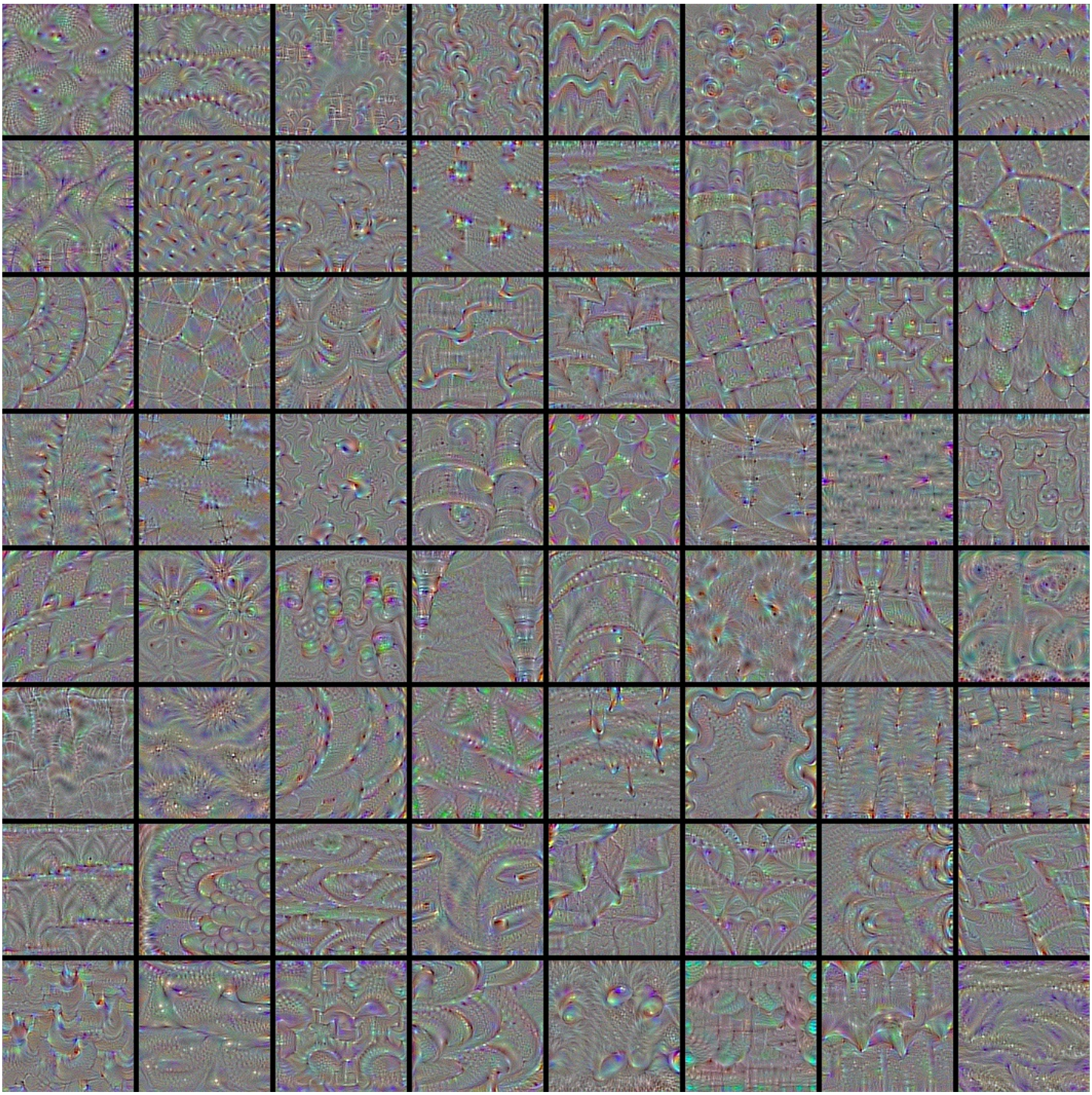
Exploratory Visualization

It is important to remember that computers do not "see" images in the same way as humans. One way to help us to understand how a CNN is "viewing" an image is to visualize the inputs that would maximize different layers of our network. Using the VGG16 model and the process described here we can see what a neural network would characterize as an ideal input.

From the first convolutional layer:



From the last convolutional layer:

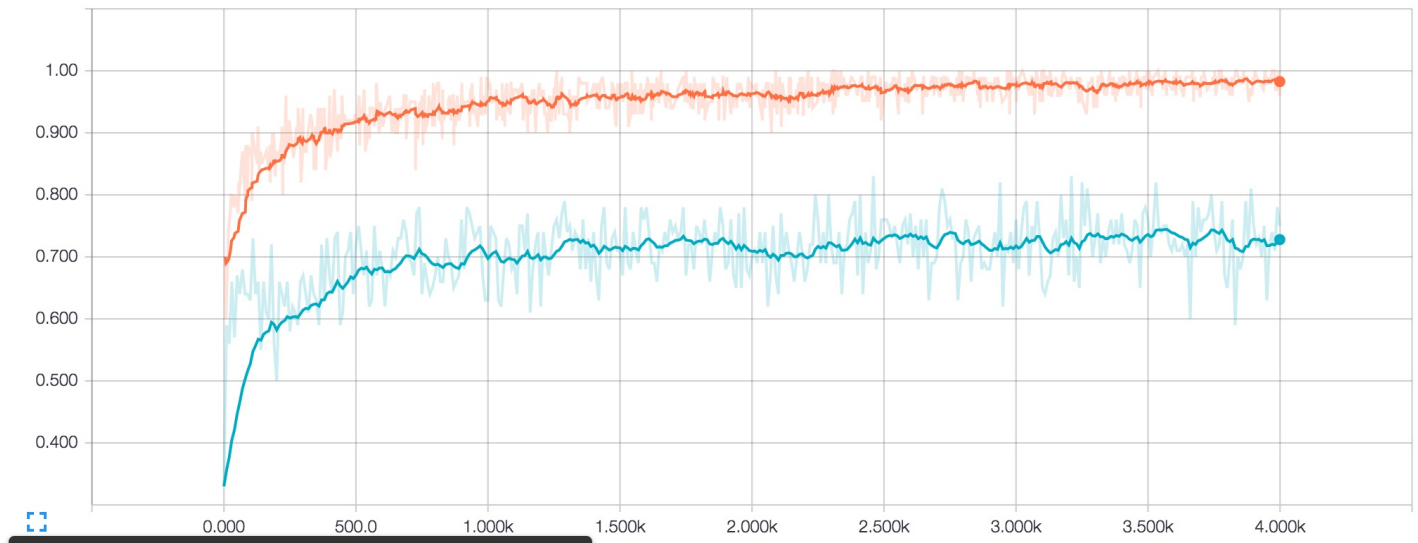


Examining different layers within our neural network, we can begin to appreciate how our network is building up features to identify objects. We can also start to understand what a network can "see".

Benchmark

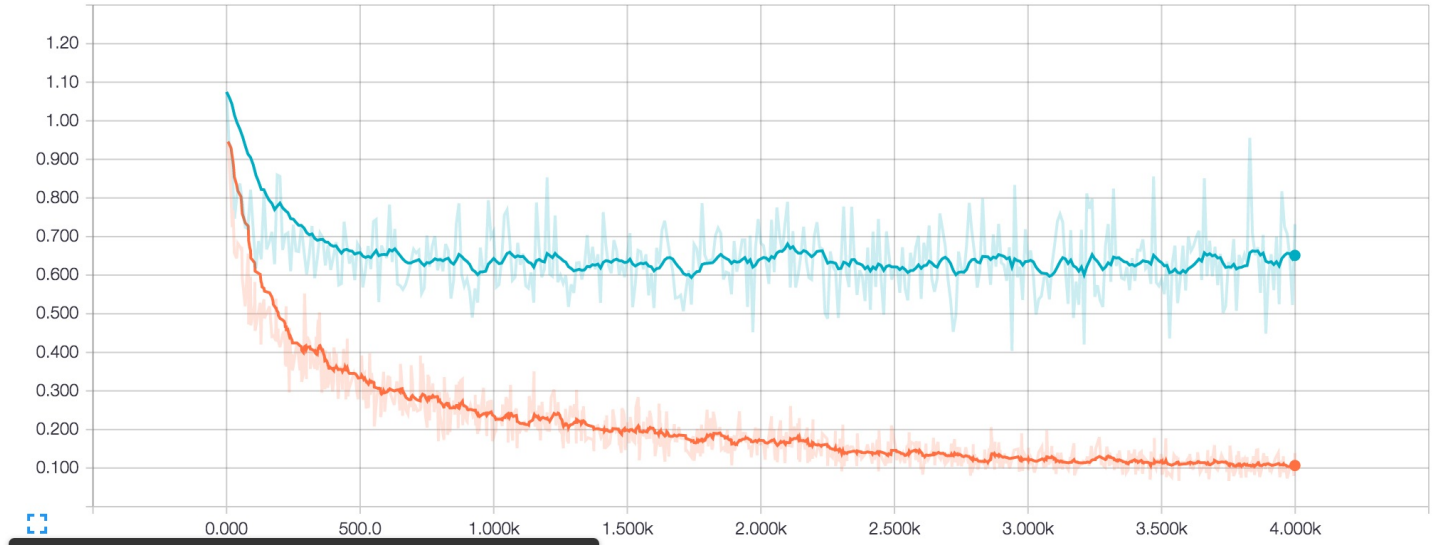
Running the initial CNN with default values as seen from this tutorial yields a result of 87.8% test accuracy on the simplest form of binary classification- only selecting Nike or Altra and the smaller our two datsets

accuracy



Run	Value	Step	Time	Relative
train	0.9826	3.999k	Wed Jul 27, 09:46:22	7m 15s
validation	0.7278	3.999k	Wed Jul 27, 09:46:22	7m 16s

cross entropy



Run	Value	Step	Time	Relative
train	0.1067	3.999k	Wed Jul 27, 09:46:22	7m 15s
validation	0.6512	3.999k	Wed Jul 27, 09:46:22	7m 16s

How does it do on the difficult images we identified previously?

altra (score = 0.98203)
nike (score = 0.01797)

altra (score = 0.99800)
nike (score = 0.00200)

altra (score = 0.96446)
nike (score = 0.03554)

nike (score = 0.98963)
altra (score = 0.01037)

altra (score = 0.85372)
nike (score = 0.14628)

nike (score = 0.80665)
altra (score = 0.19335)

III. Methodology

Data Preprocessing

Premodel Level - Image transformations

One way to improve my data set is to crop the images in a way that highlights the features that I'm interested in extracting. Cropping the images of shoes so that the logo is the most prominent aspect of the image helps clearly define the features that need to be recognized in the images.

The following image provides an example- the original image contains 3 samples of the Nike logo. By training on the original image as well as the 2 cropped versions of just the logo increases the number of training images as well as our models' ability to differentiate the Nike logo in images.

Original Image

Cropped Areas
#1 & #2 from Original

Adobe Lightroom was used to load, crop, and export images in both the Altra and Nike categories. The resulting images can be seen in the second dataset.

Model Level

The model itself handles data preprocessing steps including:

1. Images randomly split into training, validation, and testing sets

2. Resized so that the longest edge is 299 pixels wide/ tall
3. Reshaped into 299x299x3 array for Preprocessing

Implementation

The following steps were performed to create the final model:

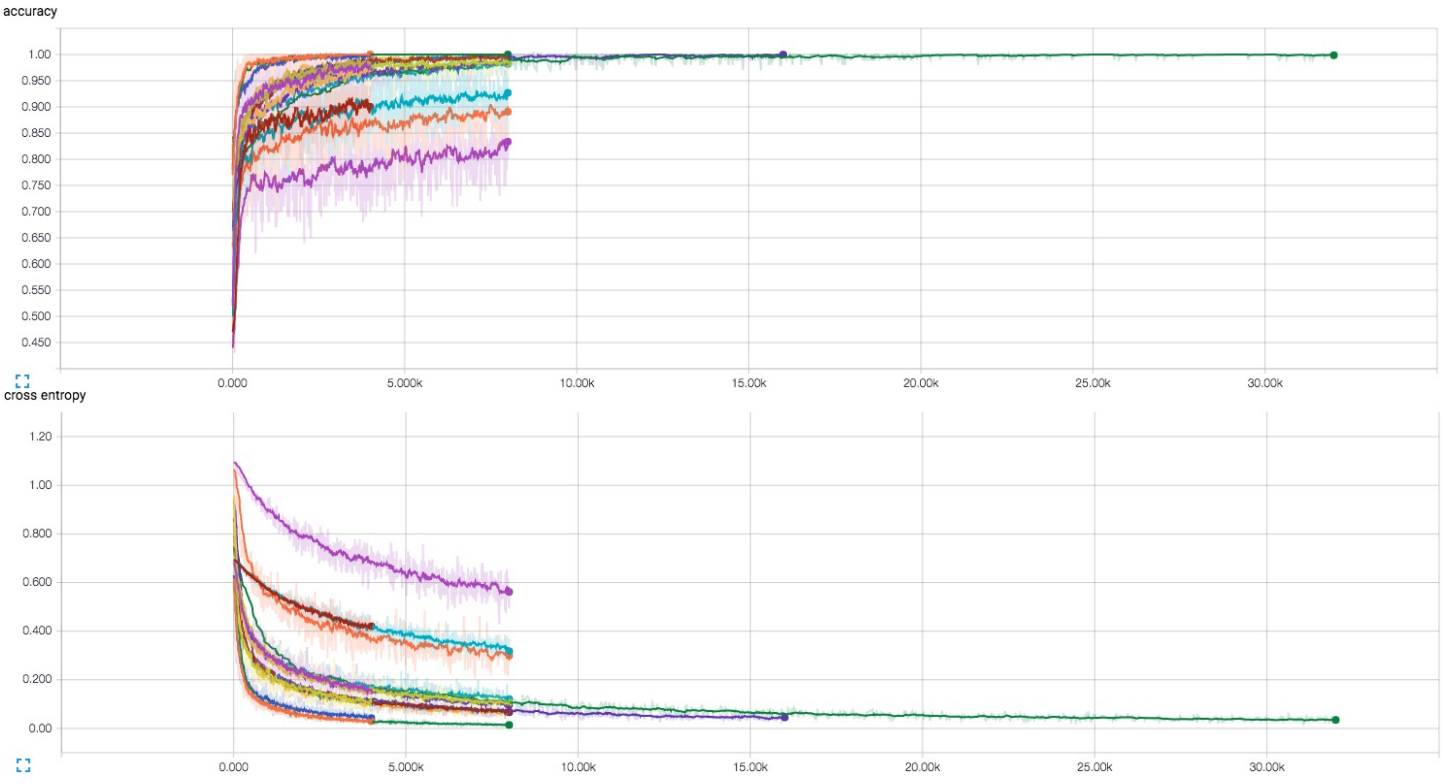
1. Construct the Inception v3 network in TensorFlow
2. Load pre-trained weights into the previously built network
3. Modify the pre-existing model for our classification tasks
4. Update the bottlenecks
5. Retrain the network on images of interest
6. Record & Evaluate model performance

In order to reconstruct these results locally, follow the TensorFlow tutorial on transfer learning and retraining. Once you have that setup, replace the existing `retrain.py` file with the version here. [###todo github link for retrain.py] and the dataset available here[###todo link to dataset].

Initial Results

18 runs total are shown below. Some underperforming trials are omitted for clarity and brevity.

Training



Validation

Refinement

Our benchmark test accuracy of 87.8% only describes the case of binary classification. By adding the items below my model achieves a similar level of accuracy with a more complex multiclass classification problem set before it.

Not just binary classification:

Importantly I also wanted to differentiate between the brands of interest and other shoes/ brands. I.E. I didn't want to categorize everything as either Nike or Altra, but rather be able to know that an image did not contain one of the logos I was tracking. To implement, I added a third image category title "neither" comprised of a random selection of images that represented anythings that was not a Nike or Altra item.

Additional images

The most straightforaward way to improve results from a neural network is to increase the amount of data being used to train that network. The second dataset was added to increase the number of images in the training pool.

Hyperparameter Search

Once my initial network was functioning there are a number of different hyperparamters available that can be varied to potentially improve model performance. They include:

- Learning Rate (η)

    - How big of a step to take along the gradient of descent.
- Number of iterations
    - How many training steps to take.
- Optimizer Type
    - Stochastic Gradient Descent (SGD) - most common.
    - Adaptive Gradient or "Adagrad" that adapts the learning rate and utilizes an accumulator value (δ) that is typically between 0.01 and 0.1

IV. Results

Model Evaluation and Validation

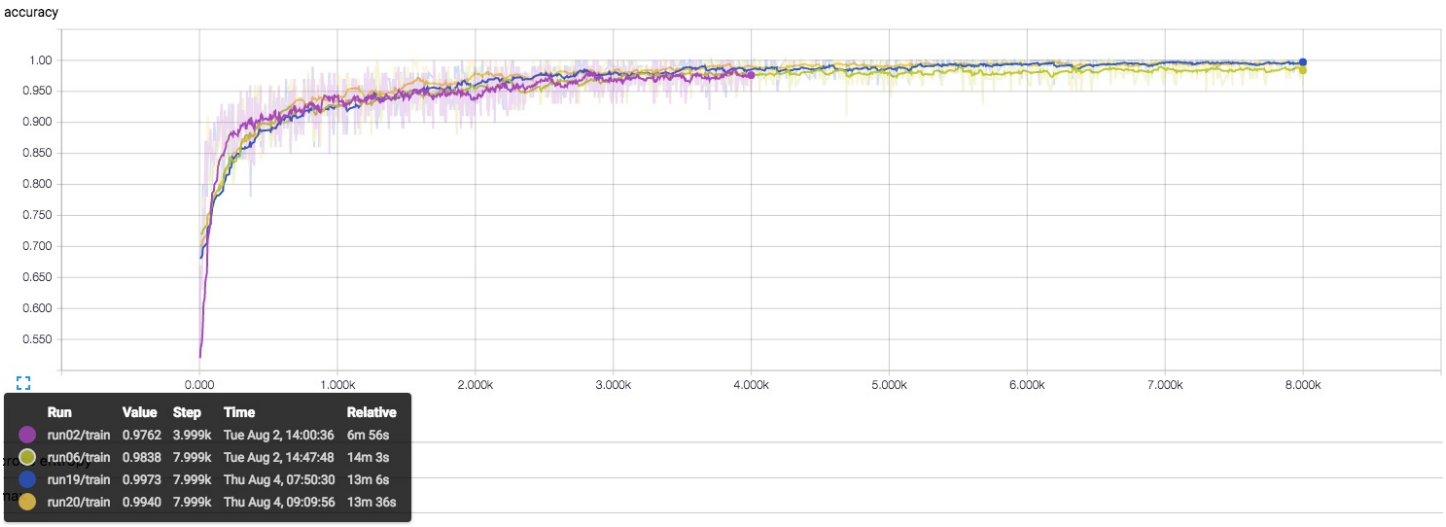
Approximately 20 different runs were conducted across a number of different hyperparameters.

Ranked by test accuracy, here are the top results.

Top Results					
Run #	Test Accuracy	Iterations	Learning Rate (η)	Data Set	Classificatio
6	97.4%	8,000	0.001	First	Binary
2	96.8%	4,000	0.001	First	Binary
20	88.6%	8,000	0.001	Second	Multiclass
19	87.4%	8,000	0.001	Second	Multiclass

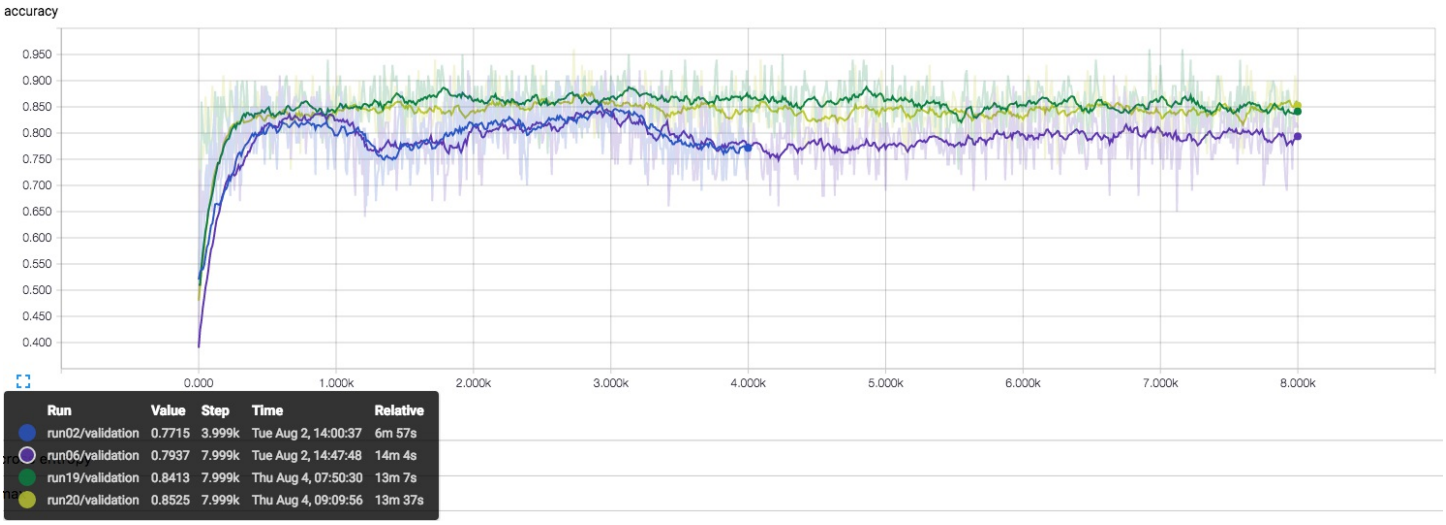
Comparing the training and validation scores amongst the top results:

Training Accuracy



Run	Value	Step	Time	Relative
run02/train	0.9762	3.999k	Tue Aug 2, 14:00:36	6m 56s
run06/train	0.9838	7.999k	Tue Aug 2, 14:47:48	14m 3s
run19/train	0.9973	7.999k	Thu Aug 4, 07:50:30	13m 6s
run20/train	0.9940	7.999k	Thu Aug 4, 09:09:56	13m 36s

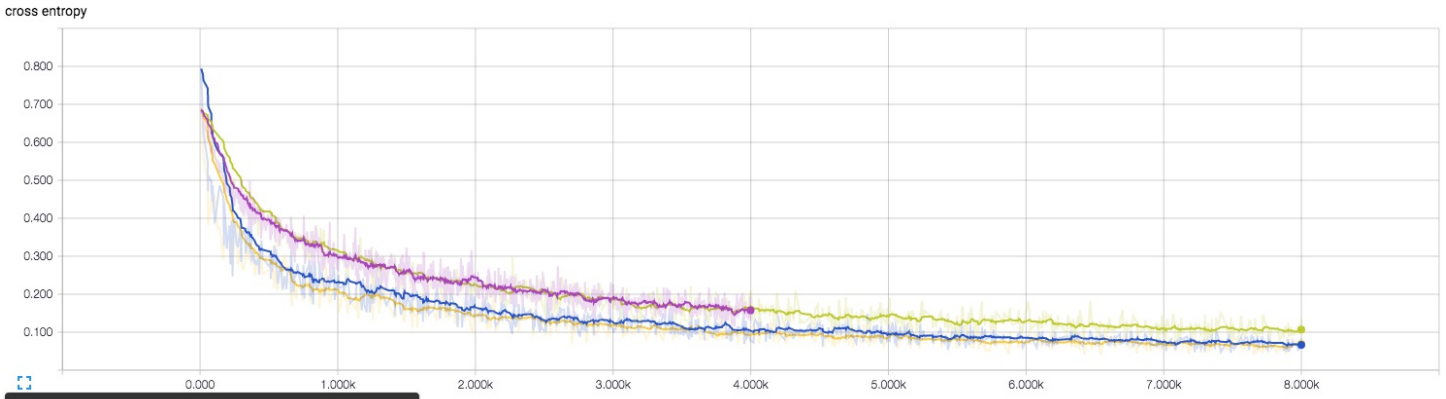
Validation Accuracy



Run	Value	Step	Time	Relative
run02/validation	0.7715	3.999k	Tue Aug 2, 14:00:37	6m 57s
run06/validation	0.7937	7.999k	Tue Aug 2, 14:47:48	14m 4s
run19/validation	0.8413	7.999k	Thu Aug 4, 07:50:30	13m 7s
run20/validation	0.8525	7.999k	Thu Aug 4, 09:09:56	13m 37s

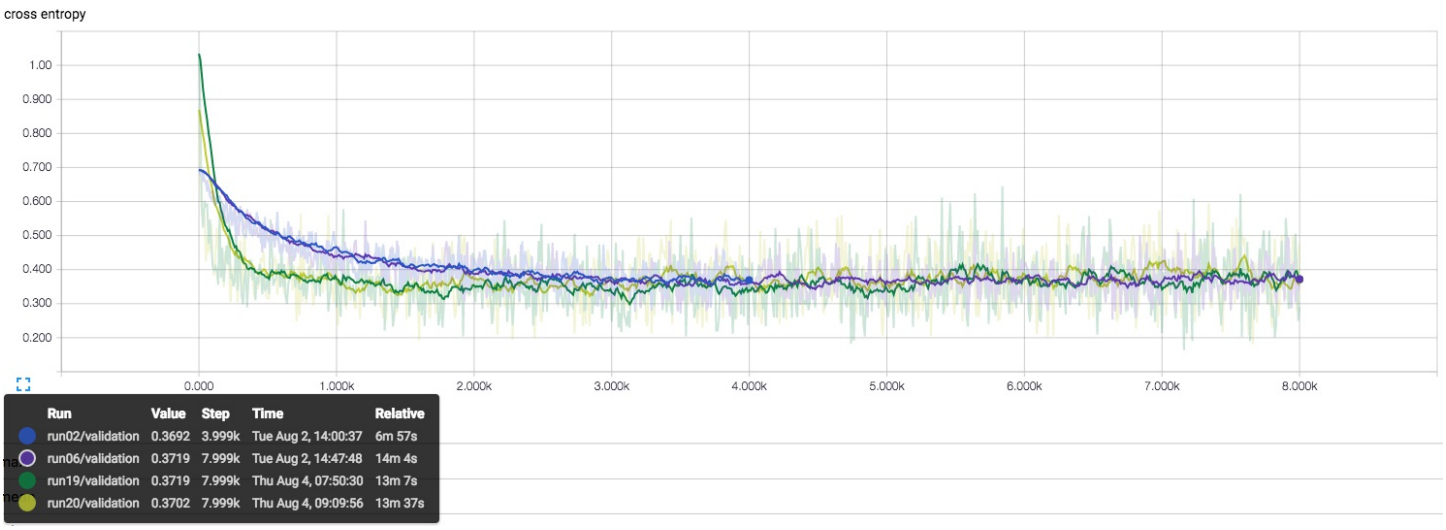
We can also compare cross-entropy between models

Training Cross-Entropy



Run	Value	Step	Time	Relative
run02/train	0.1573	3.999k	Tue Aug 2, 14:00:36	6m 56s
run06/train	0.1069	7.999k	Tue Aug 2, 14:47:48	14m 3s
run19/train	0.06654	7.999k	Thu Aug 4, 07:50:30	13m 6s
run20/train	0.06927	7.999k	Thu Aug 4, 09:09:56	13m 36s

Validation Cross-Entropy



Justification

Top Results Binary

Top Results					
Run #	Test Accuracy	Iterations	Learning Rate (η)	Data Set	Classificatio
20	88.6%	8,000	0.001	Second	Multiclass

Top Results Multiclass

Top Results					
Run #	Test Accuracy	Iterations	Learning Rate (η)	Data Set	Classificatio
20	88.6%	8,000	0.001	Second	Multiclass

Here are the graphs of its performance compared to the previous top-runners. [###todo graph of *the* best model]

Revisiting Difficult Photos

Despite improvements to our model, we can see from the below that there remains images that are difficult to classify. Although these results do not show 100% accuracy, for my purposes I prefer that the model "fails" to the neither category (a false negative). Previously, in the binary classification model, the model would fail to one of the brands of interest (false positive).

Binary

[### todo rebuld model and add binary images]

####Multiclass

neither (score = 0.99816)
nike (score = 0.00158)
altra (score = 0.00025)

neither (score = 0.98079)
altra (score = 0.01201)
nike (score = 0.00720)

neither (score = 0.97332)
altra (score = 0.02171)
nike (score = 0.00497)

neither (score = 0.49912)
nike (score = 0.49754)
altra (score = 0.00334)

neither (score = 0.87823)
nike (score = 0.10636)
altra (score = 0.01541)

neither (score = 0.85841)
nike (score = 0.13508)
altra (score = 0.00651)

V. Conclusion

Free-Form Visualization

We started with a discussion of how being able to identify problems with your products would permit a company to contact and resolve issues for customers. So how does our model perform on images of shoes that may require intervention?

neither (score = 0.58464)
nike (score = 0.23102)
altra (score = 0.18433)

altra (score = 0.92383)
nike (score = 0.07578)
neither (score = 0.00039)

neither (score = 0.55865)
nike (score = 0.35871)
altra (score = 0.08264)

nike (score = 0.87319)
neither (score = 0.11433)
altra (score = 0.01249)

nike (score = 0.52935)
altra (score = 0.46793)
neither (score = 0.00272)

neither (score = 0.74700)
nike (score = 0.25135)
altra (score = 0.00166)

Reflection

Steps to create a solution:

1. Identify a suitable framework, TensorFlow, to build a pre-trained network
2. Construct the existing Inception v3 network
3. Create a dataset of images from Instagram
4. Load pre-trained weights into the previously built network

5. Modify the pre-existing model for our classification tasks
6. Update the bottlenecks
7. Retrain the network on images of interest
8. Evaluate model performance
9. Optimize hyperparameters and compare performance
10. Select the optimum solution

There were a number of new, and now that they have been solved, interesting problems that I had to work through in the course of this project.

Interesting

Machine learning is unique in that the work that is being done at its highest levels- like that be performed at Google or research institutions can has been released for public use. What other field allows for the almost immediate implementation of cutting/ bleeding edge technologies by a new practitioner? The idea of democratizing artificial intelligence is a call to action to its practitioners.

Applying what I have learned to a real world dataset proved immensely empowering and rewarding. Being able to use my coureswork on things found "out in the wild" was a powerful way to implement things that I had studied on an actual problem and not just a "toy" dataset.

TensorFlow & Keras are both popular tools in the machine learning community right now and I'm glad I got the opportunity to apply them to a project and gain a better understanding of them.

Difficulties

TensorFlow and Keras were both new resources for me to work with, and both provided great functionality once I was able to learn and implement their features. Unfortunately they are not without a learning curve and in some cases bugs. Why a retrained network Tracking down and understanding why TensorBoard would not display some of my logged training runs turned out to be an interesting issue with how paths are specified.

Visualizing neural networks proved to be one of the more difficult aspect of the project. There are a number of ways to Attempting to build Caffe for its abilities to create visualizations of different layers of a network was quite difficult. There are a number of dependencies and local customizations needed. While I will continue to sort through them, they ultimately proved too time consuming to utilize in this project.

expectations

My final model performs better than I expected. The accuracy of predictions

In this section, you will summarize the entire end-to-end problem solution and discuss one or two particular aspects of the project you found interesting or difficult. You are expected to reflect on the project as a whole to show that you have a firm understanding of the entire process employed in your work. Questions to ask

yourself when writing this section:

- *Have you thoroughly summarized the entire process you used for this project?*
- *Were there any interesting aspects of the project?*
- *Were there any difficult aspects of the project?*
- *Does the final model and solution fit your expectations for the problem, and should it be used in a general setting to solve these types of problems?*

Improvement

- implement a web based interface to this model. Bonus point if you could provide additional information in the event an image was mis-classified. Remotely using the model would be beneficial in lieu of downloading my Python code and running locally. Additionally, a web based implementation could be extended to mobile devices as well, negating the need for a mobile based version.
- TensorFlow can be run on a mobile device, extending this work to an iOS or Android device would allow the model to be run remotely.
- Comparing my results to another baseline, such as PCANet would be a good gut check of my results.
- *Were there algorithms or techniques you researched that you did not know how to implement, but would consider using if you knew how?*
- The process of removing and re-architecting the VGG16 architecture in Keras became unwieldy.
- Visualizing the outputs/ weights of different levels within the CNN
  - like this
  - or this
- Deeper sensitivity analysis
- The network could always be improved by adding additional training images.

Before submitting, ask yourself. . .

- Does the project report you've written follow a well-organized structure similar to that of the project template?
- Is each section (particularly **Analysis** and **Methodology**) written in a clear, concise and specific fashion? Are there any ambiguous terms or phrases that need clarification?
- Would the intended audience of your project be able to understand your analysis, methods, and results?
- Have you properly proof-read your project report to assure there are minimal grammatical and spelling mistakes?
- Are all the resources used for this project correctly cited and referenced?
- Is the code that implements your solution easily readable and properly commented?
- Does the code execute without error and produce results similar to those reported?