Training Neural Networks for Robotic Inference

Stephan Hohne

Abstract—Two project tasks for object classification are carried out. First, a given data set containing images of grocery store items is used to create a model that meets criteria for inference time and accuracy. For the second task, image data for classification of three types of fruit is collected. For each task, the NVIDIA DIGITS GPU infrastructure is used to train a chosen neural network architecture for image classification. The resulting models are applied to infer the class of unknown data samples. The accuracy and time for inference is evaluated. Applications of these models to robotic systems such as object sorting on a conveyor belt are discussed.

Index Terms—Neural Network Training, Image Classification, Inference.

1 Introduction

Deep neural networks with convolution layers have become a standard tool for object classification [1]. Originally, they were used for recognizing handwritten digits, with the classic MNIST data set [2]. The AlexNet [3] and GoogleNet [4] architectures were trained on the larger ImageNet data set. These deep neural networks rely on a large database of labeled samples, and are able to detect a variety of objects in images.

In this project, classification of objects on a conveyor belt is studied. Two different tasks are specified.

Task 1 Classify grocery items on the conveyor belt in one of three categories `Bottle`, `Candy_Box`, or `Nothing`.

Task 2 Classify food items on the conveyor belt in one of three categories `apricot`, `egg`, or `tomato`.

The training data for the first task are provided in the Udacity workspace. The data set consists of 10094 color images in PNG format. The training data for the second task is collected by taking pictures with an iPhone 6, resulting in 507 color images in JPG format.

The model training is performed using a GPU powered NVIDIA DIGITS environment, which also provides standard network architectures with predefined hyperparameter settings. For the first task, the AlexNet architecture [3] is employed for model training. For the second task, AlexNet [3] and GoogleNet [4] are used.

This report is organized as follows. In section 2, the choice and configuration of network architectures for this project is discussed. In section 3, the acquisition of image data and the training of the networks is described. The inference results of the trained models are stated in section 4 and discussed in section 5. A conclusion and an outlook to future work is given in section 6.

2 Background: Neural Networks for Classification

In order to accomplish the given object recognition tasks, the AlexNet and GoogleNet architectures were chosen for model training. The AlexNet architecture [3] won the 2012 ImageNet challenge, where it achieved a 17% top-5 error rate [1]. The GoogleNet architecture [4] won the ILSRVC 2014 challenge, with a top-5 error rate below 7% [1]. It features inception modules which allow for a much more efficient use of parameters. GoogleNet uses 10 times fewer parameters than AlexNet which reflects in the model size. The proven performance of these architectures makes them good choices for attempting the given classification tasks. The NVIDIA DIGITS environment provides these architectures to the user, implemented using the Caffe framework.

2.1 Task 1

The model training was performed using the AlexNet architecture over six episodes. The batch size was kept to network defaults. The stochastic gradient descent solver was used. The hyperparameter settings for the learning rate were kept at their default values. This means that the initial learning rate was 0.01 for episodes one and two, 0.001 for episodes three and four, and 0.0001 for episodes five and six.

2.2 Task 2

For the task of food item classification, training was performed on two different network architectures, AlexNet [3] and GoogleNet [4]. For both architectures, the same parameter setting was used. the number of episodes was set to 30. The batch size was kept at network defaults, and the stochastic gradient descent solver was used. The hyperparameter settings for the learning rate were kept at their default values. This means that the initial learning rate was 0.01 for the first ten episodes, 0.001 for the next ten episodes, and 0.0001 for the last ten episodes.

3 Data Acquisition

3.1 Task 1

The first project task was to practice the DIGITS workflow for model generation using a supplied data set. The raw training data is located in the workspace in the folder `/data/P1_data`. The data set consists of 4568 images in the class `Bottle`, 2495 images of type `Candy_Box`, and 3031 images in the class `Nothing`, for a total of 10094 color images. The files are in PNG format. Figure 1 shows the

S2H.Mobile@gmail.com

properties of the corresponding database creation job. The job was run with a train/val/test split of 60/20/20, so there were 6057 images for training, 2019 images for validation and 2018 images for testing. In order to match the size of the initial layer in the network, the images were squashed to a dimension of 256×256 pixels.

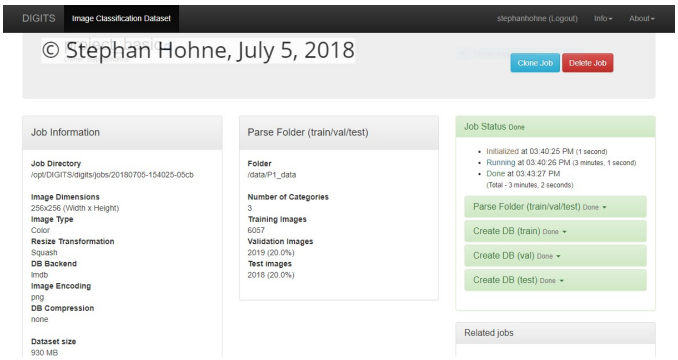


Fig. 1. Properties of database creation job 20180705-154025-05cb. Screenshot taken in the DIGITS environment.

3.2 Task 2

The training data for classifying food items was collected by taking pictures with an iPhone 6. The items were placed on a desk with a surface structure that resembles a conveyor belt. A variation in orientation and lighting was achieved by rotating the camera. In total, 507 color images in JPG format were created. The image dimensions are 2448×3264, and each file has a size of approximately $2 - 3$ MB. There are 159 images of eggs, 166 images of apricots, and 182 images of tomatoes used for training. The picture folders can be downloaded from Google Drive as shared files for eggs, tomatoes, and apricots. Example images are shown in figures 2, 3, 4.

The images used for inference were created in the same way. The inference data set consists of 11 images of food items that are not present in the training data, 3 pictures of an apricot, 4 images of a tomato, and 4 images of an egg. The inference images can be found in the `data/inference` folder in the project archive.

The database creation job was run with a train/val/test split of 60/20/20, so there were 304 images for training, 101 images for validation and 102 images for testing. In order to match the size of the initial layer in the network, the images were squashed to a dimension of 256×256 pixels.

4 Results

4.1 Task 1

The model 20180705-160040-e827 resulted from training the AlexNet architecture with the parameter setting described in section 2 on the data set described in section 3. The corresponding model files can be found in the `models/` folder in this archive. Since the AlexNet architecture was used, the `.caffemodel` file is about 222 MB in size.

Figure 5 shows the inference performance of this model as determined by using the `evaluate` command. The average inference time is approximately $4.6ms$. The accuracy is approximately 75.41%.

Fig. 2. Example image of a tomato in the training data set.

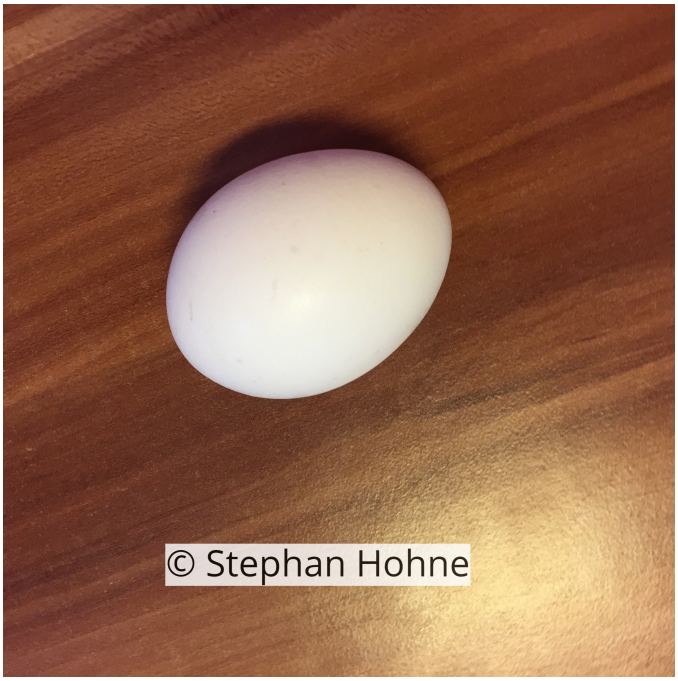


Fig. 3. Example image of an egg in the training data set.

Fig. 4. Example image of an apricot in the training data set.

4.2 Task 2

The data set used for inference can be found in folder `/data/inference`. The corresponding image list can be found in the file `data/inference.list`. Both the AlexNet and the GoogleNet architectures were trained, as described in section 2. This resulted in two different models named `new_alex_30` and `new_large_google_30`. Each model achieved an accuracy of over 80% on the inference data set. Since the `evaluate` command could not be applied to models generated with custom data, the inference time could not be measured. The performance of both models is summarized in table 1.

TABLE 1
Accuracy of the models performing task 2.

Model Screeenshot	`new_alex_30` 9	`new_large_google_30` 10
apricot	100%	100%
tomato	50%	50%
egg	100%	100%
total	81.8%	81.8%

The models can be downloaded from Google Drive as shared files `new_alex_30`, `new_large_google_30` in `tar.gz` format.

5 Discussion

Both task 1 and task 2 involve classifying objects on a moving conveyor belt. The robotic inference project (task 2) is to sort food items, which could be commercially applied in logistics to package items separately. For both task 1 and task 2, the accuracy of classification is more important than the inference time. The rate of speed of the conveyor belt can be adjusted to the inference time. If necessary, the belt can be slowed down, which means that a smaller number of items can be sorted per time unit. This is more tolerable than incorrectly packaged food items. A classification model with longer inference time and higher accuracy should be preferred over a model with lower accuracy and shorter inference time.

The models trained for task 2 seem to perform reasonably well at inferring the object classes. For applying these models to real conveyor belts, their accuracy needs to be improved further.

The training curves shown in figures 6, 7 and 8 indicate that the models over fit the training data, since the accuracy goes up to almost 100% after only a few episodes. Since the networks and hyperparameters used for creating the models are proven standard, the main reason seems to be the limited amount of data collected. Therefore the main challenge when improving model performance is to collect more training data with a greater variety in lighting conditions, object orientation, etc.

Several methods of collecting image data were investigated. A laptop camera and the onboard camera of the Jetson TX2 were used. Variation in the images was achieved by moving the objects in between the snapshots. This method turned out to be inaccurate and inefficient, so that a large number of images could not be acquired in a reasonable amount of time. Moving the camera of a mobile phone instead of the object was more efficient.

6 Conclusion and Future Work

Two object classification tasks were performed. In the first task, the AlexNet architecture was trained on a given data set and achieved an average inference time of about $4.6ms$ and an accuracy of about 75.41% at classifying grocery items on a conveyor belt. For the second task, image data of three types of food were collected, and used to train AlexNet and GoogleNet networks. The resulting models achieved overall accuracy of 80.0%, albeit on a small data set.

In order to deploy the models on a commercially viable product, more training data needs to be collected, which increases the accuracy and overall performance of the models. The choice of award wining architectures is a safe starting point for hyperparameter tuning. Therefore future work should include the following tasks.

- Collect more image data with a greater variety, possibly in an automated procedure.
- Develop a means to determine the inference time fro custom models, where the `evaluate` command can not be applied.
- Evaluate the models further. Create the confusion matrix. Calculate precision, recall and F1 score.
- Explore the hyperparameter space of the convolution architectures. For instance, use the AdaGrad solver instead of stochastic gradient descent.
- Deploy the inference models on the Jetson TX2. Mount the Jetson over a conveyor belt.

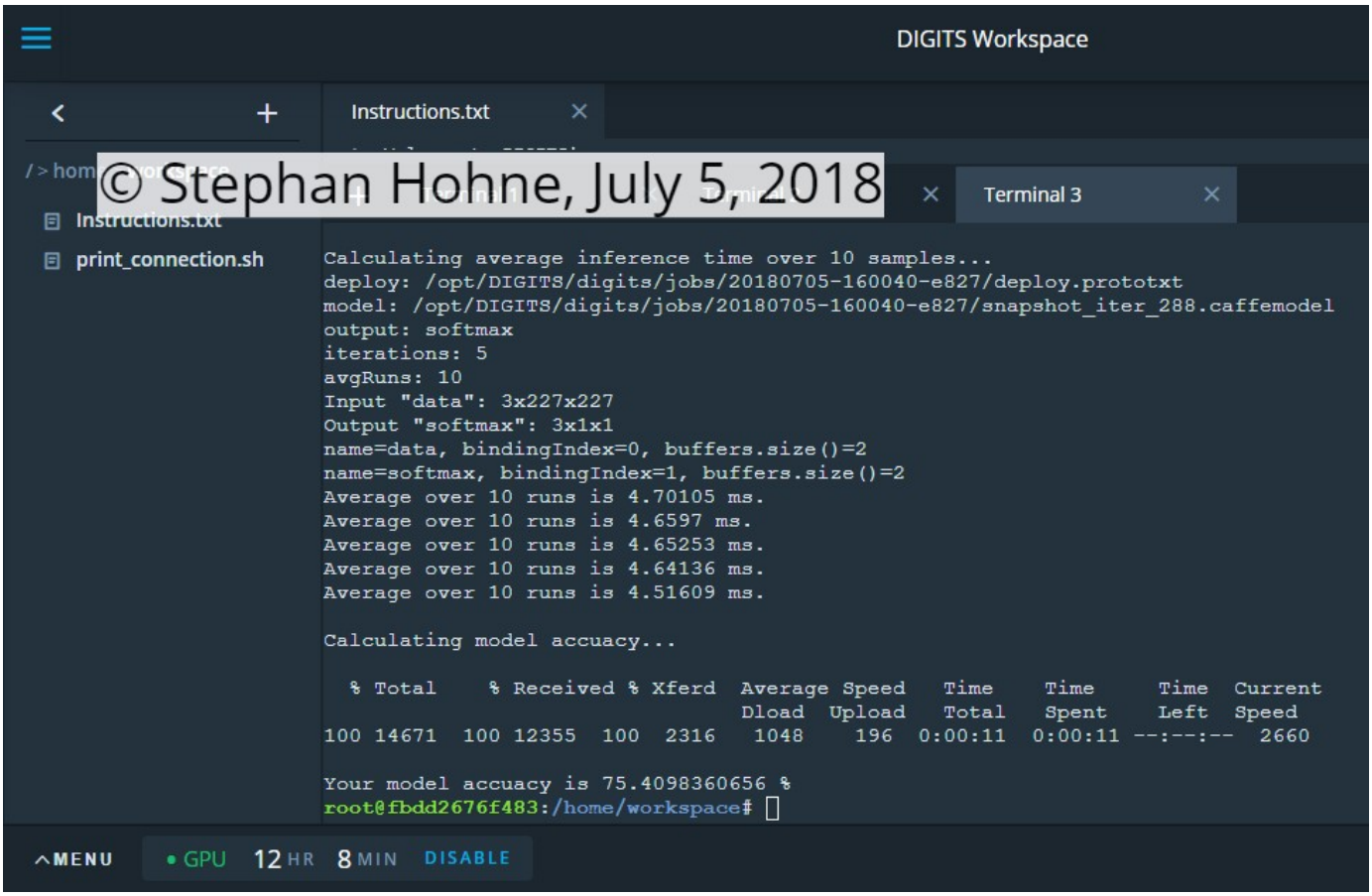


Fig. 5. Performance of model 20180705-160040-e827 on data set 20180705-154025-05cb. Output from the `evaluate` command. Screenshot taken in the Udacity workspace.

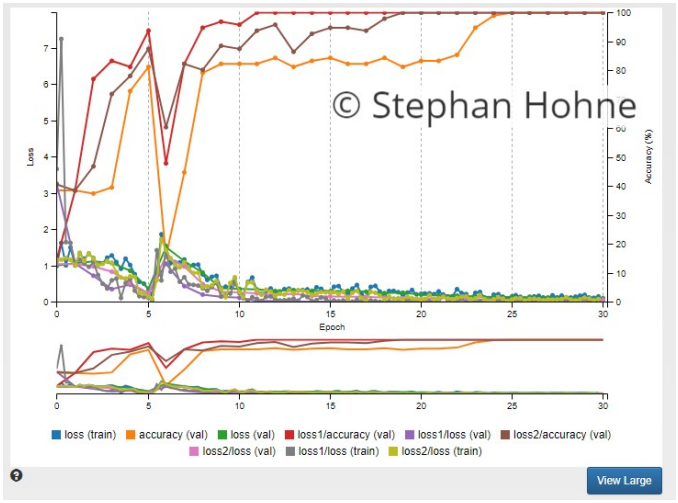


Fig. 6. Training curve over 30 episodes for the model `large_google_30` with job id 20180707-133949-aae5.

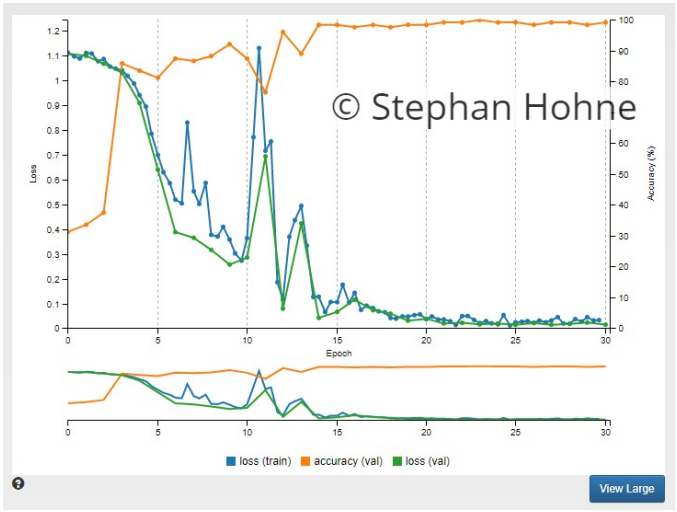


Fig. 7. Training curve over 30 episodes for the model `new_alex_30` with job id 20180707-143126-1d4a.

REFERENCES

[1] A. Geron, *Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems*. Sebastopol, CA: O'Reilly Media, 2017.

[2] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, pp. 2278–2324, November 1998.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25* (F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds.), pp. 1097–1105, Curran Associates, Inc., 2012.

[4] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *CoRR*, vol. abs/1409.4842, 2014.

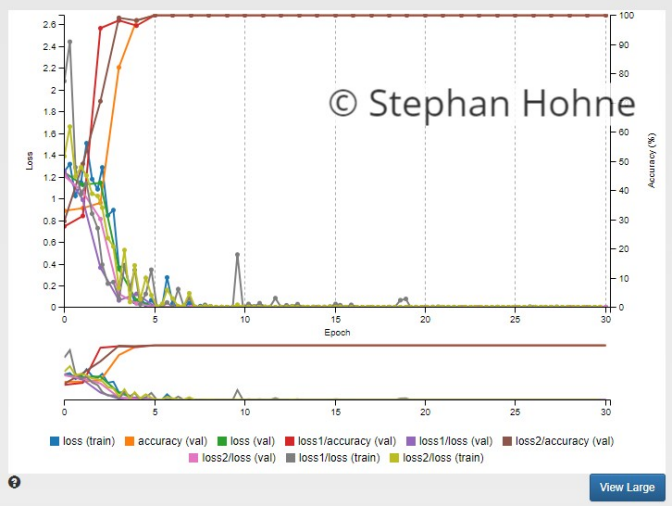
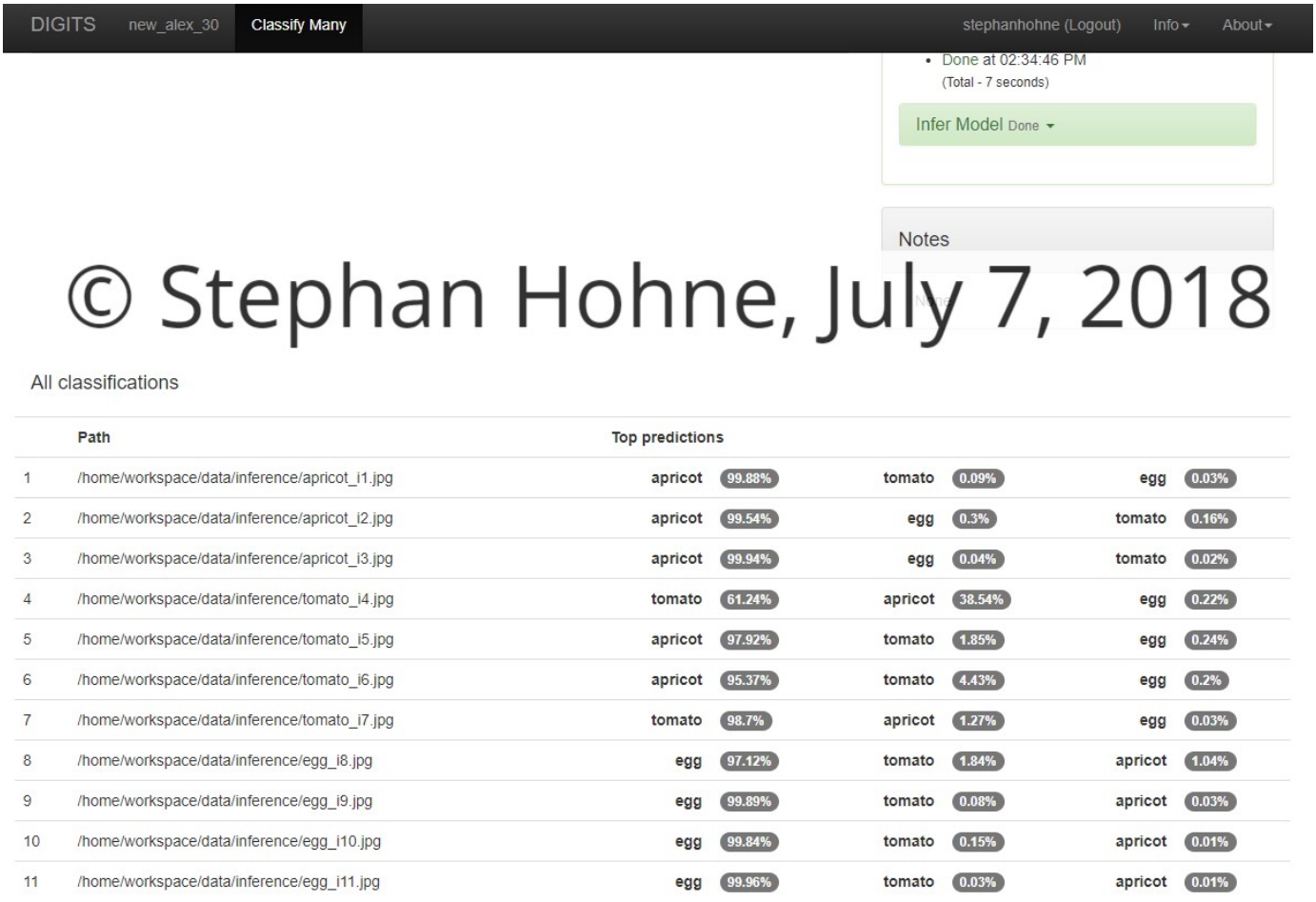
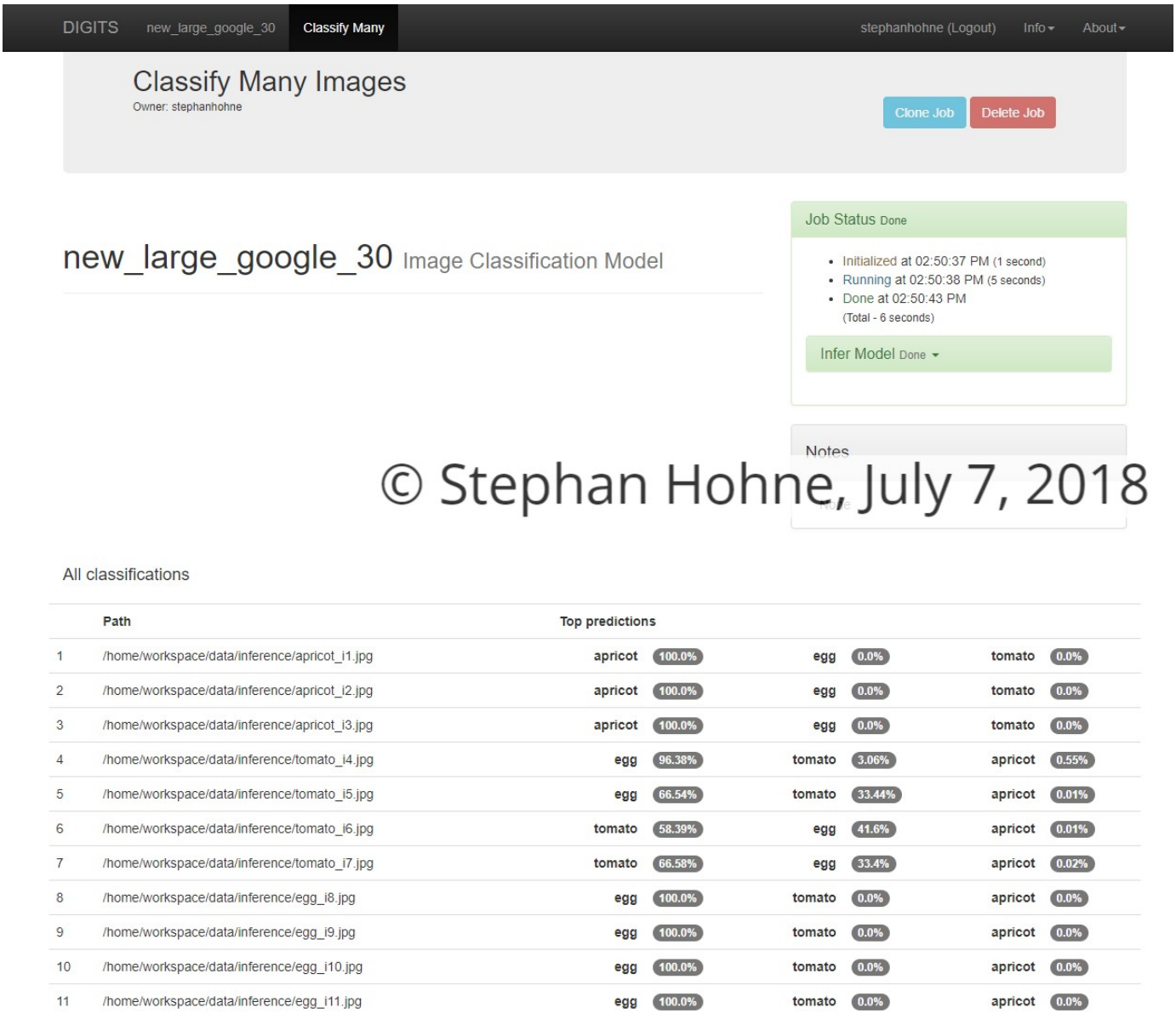


Fig. 8. Training curve over 30 episodes for the model `new_large_google_30` with job id 20180707-144622-eca5.



	Path	Top predictions		
1	/home/workspace/data/inference/apricot_i1.jpg	apricot 99.88%	tomato 0.09%	egg 0.03%
2	/home/workspace/data/inference/apricot_i2.jpg	apricot 99.54%	egg 0.3%	tomato 0.16%
3	/home/workspace/data/inference/apricot_i3.jpg	apricot 99.94%	egg 0.04%	tomato 0.02%
4	/home/workspace/data/inference/tomato_i4.jpg	tomato 61.24%	apricot 38.54%	egg 0.22%
5	/home/workspace/data/inference/tomato_i5.jpg	apricot 97.92%	tomato 1.85%	egg 0.24%
6	/home/workspace/data/inference/tomato_i6.jpg	apricot 95.37%	tomato 4.43%	egg 0.2%
7	/home/workspace/data/inference/tomato_i7.jpg	tomato 98.7%	apricot 1.27%	egg 0.03%
8	/home/workspace/data/inference/egg_i8.jpg	egg 97.12%	tomato 1.84%	apricot 1.04%
9	/home/workspace/data/inference/egg_i9.jpg	egg 99.89%	tomato 0.08%	apricot 0.03%
10	/home/workspace/data/inference/egg_i10.jpg	egg 99.84%	tomato 0.15%	apricot 0.01%
11	/home/workspace/data/inference/egg_i11.jpg	egg 99.96%	tomato 0.03%	apricot 0.01%

Fig. 9. Inference results for the model `new_alex_30` with job id 20180707-143126-1d4a. Screenshot taken in the DIGITS workspace.



	Path	Top predictions		
1	/home/workspace/data/inference/apricot_i1.jpg	apricot 100.0%	egg 0.0%	tomato 0.0%
2	/home/workspace/data/inference/apricot_i2.jpg	apricot 100.0%	egg 0.0%	tomato 0.0%
3	/home/workspace/data/inference/apricot_i3.jpg	apricot 100.0%	egg 0.0%	tomato 0.0%
4	/home/workspace/data/inference/tomato_i4.jpg	egg 96.38%	tomato 3.06%	apricot 0.55%
5	/home/workspace/data/inference/tomato_i5.jpg	egg 66.54%	tomato 33.44%	apricot 0.01%
6	/home/workspace/data/inference/tomato_i6.jpg	tomato 58.39%	egg 41.6%	apricot 0.01%
7	/home/workspace/data/inference/tomato_i7.jpg	tomato 66.58%	egg 33.4%	apricot 0.02%
8	/home/workspace/data/inference/egg_i8.jpg	egg 100.0%	tomato 0.0%	apricot 0.0%
9	/home/workspace/data/inference/egg_i9.jpg	egg 100.0%	tomato 0.0%	apricot 0.0%
10	/home/workspace/data/inference/egg_i10.jpg	egg 100.0%	tomato 0.0%	apricot 0.0%
11	/home/workspace/data/inference/egg_i11.jpg	egg 100.0%	tomato 0.0%	apricot 0.0%

Fig. 10. Inference results for the model `new_large_google_30` with job id 20180707-144622-eca5. Screenshot taken in the DIGITS workspace.