Visual Recognition of Human Emotion via Neural Network Classification

Arnold Kompaniyets

Robotic Software Engineering, Term 2

Project 1

Abstract

Progress in the field of robotics continues to expand the role robots have, and will continue to have, in the everyday life of the average person. Thus, aside from computational and process improvement, equal attention must be paid to improving human-robot communication, allowing interactions to be seamless and natural. A crucial component of this is implementing emotional comprehension to provide proper understanding and responses. In this project, a series of 20 neural networks were trained (8 AlexNet, 8 GoogLeNet, and 4 Modified AlexNet) using two variations of starting learning rates (0.01 and 0.001) and four variations of a 3,200 image data set (consisting of 400 front-view, whole face images, each image rotated three times by 90°and each of the subsequent images color inverted) to classify four emotional states (angry, happy, sad, and neutral). With all neural network types, grayscale data sets performed superiorly compared to color sets, with the lower starting learning rate generally performing slightly better. Negligible variation in success was seen between the three network types, with a maximum of 63% accuracy achieved on a 16 image test set. Further work would need to be done to either expand the data set, if using a similar training structure or, preferably, train individual neural networks to recognize specific facial features as part of an overall emotional expression.

Introduction

Undoubtedly, humanity has long since held a deep fascination with all matters of robotics. From mega-computers capable of controlling armies to android super-humans, much literature and thought has gone to the odds and ends of potential apocalyptic results of robotic advancement. However, more attention should perhaps be paid to the short-term realities of robotics, and just as with virtually all other creations of the human race, the crux of this will lie in robots' interactions with the people around them. Long before robots are venturing the nether regions of space on self-guided adventures, they will be in constantly close proximity to people. And even in this latter scenario, many might already see images of humanoid-like robots interacting with people around them in the most awkward and artificial of ways. This most certainly does not need to be the case. At least in the current day and age, all robots are a product of human manufacture, so there is simply no need for the stereotype of, well, “robotic” interactions to persist as the field of robotics continues. Be it self-driving cars, home-cleaning devices, robot chefs, etc., it is easy to see that perhaps all too soon, most individuals will interact with plethora robots on a daily basis. How will, or rather how should, these interactions proceed?

It can quite easily be agreed upon that the fundamentals of human communication lie in emotions. Whereas some animals may choose to interact via basic chemical signals transferred from one membrane to another, humans are much more subjective communicators. Almost more than any other species, humans are incredibly expressive in communication. In fact, rarely is the content of a conversation (i.e. the words) nearly as significant as to the means of delivery (i.e. the emotional content involved). Furthermore, communication can even proceed with emotional content alone, as a great deal of information can be transferred via only emotional expression. However, even with this important realization, human-robot interaction is still largely only done with the basic content of communication, irrespective of emotion; inevitably, this leads to immediate disconnect, mistrust, and feelings of inadequate robot performance.

Of course, the primary reason for the lack of emotional content in robotic interaction with humans has not been due to a lack of trying, per se. Instead, the difficulty of integrating emotional interpretation and communication into robotic technology has been the issue at hand. With human communication being primarily visual and/or auditory in nature, a robot would not only need to be capable of properly taking in visual/auditory information, but then adequately interpret and classify the emotional content of said information. This added depth provides an incredible challenge, and can understandably be seen as unnecessary for accomplishing many tasks of a robot. Nevertheless, over time, it cannot be denied that as the technological capabilities of robots improve, so should their communication skills.

Ultimately, the biggest impediment to successful human-robot communication will be in the realization (or perhaps the better term is innate instinct) that truly successful communication requires the ability for both parties involved to feel, to actually experience the emotional component of communication. While this is an altogether different topic, the focus here will remain in the fact that successful emotional interpretation, especially visually, will be a must as the field of robotics continues to grow.

With the rise in prominence of neural networks, it is the hope that they can also be used in this context. Just as a child is guided in learning emotions through extensive and continual observation, perhaps neural networks can too. This is, undoubtedly, a very extensive topic, and will therefore require extensive research and experimentation. Therefore, it appears logical to simply start with the most basic concept: training a neural network with fontal-view, whole-face images and seeing if success in emotional interpretation can be attained.

Background/Formulation

Supplied Data Set:

For the supplied data set, the default GoogLeNet architecture was chosen using the Nvidia Digits interface. From the three default network architectures to choose from, AlexNet and GoogLeNet were narrowed down as the two potential choices, simply owing to the fact that 256 X 256 pixel images are expected, which is exactly the size the sample data set images were already formatted as. After this decision, the choice to run GoogLeNet was simply based on superior average network performance (as observed from classification competitions). While it is true that the two aforementioned network architectures do vary in their overall energy and CPU (or GPU) utilization, as well as overall architecture depth and complexity, these factors were not needed to be considered for this rather quick task. Furthermore, the overall architecture layers were not altered in the GoogLeNet used here, nor were the details, so the kernel size, strides, etc., were left as is. In training the network, 5 epochs were chosen as sufficient, with one run through the training data set per epoch and one subsequent validation dataset pass per epoch. Lastly, given the larger dataset and resulting more rapid learning and weight adjustments happening after each epoch, the starting learning rate was lowered by a factor of 10, form the default 0.01 to 0.001. During the training itself, the learning rate was allowed to lower in a step-wise manner by DIGITS.

Robotic Inference Project:

Given the difficulty of the subject matter, both AlexNet and GoogLeNet networks were used in attempts to correctly classify the desired emotions. Initially, both networks were used in their default states, identical to that described in the “Supplied Data Set” section above. In subsequent attempts at attaining superior results, various alteration were performed. First, the number of epochs used for training was altered extensively. For both network architectures, a range from 10 to 150 epochs was implemented. Within these different epoch ranges, a wide variety of learning rates were also experimented with. The default 0.01 starting learning rate was used first in both networks, choosing a step-wise descent. Following this, a 0.001 starting learning rate was used, also with step-wise descent. A 0.01 starting learning rate was attempted with exponential descent, but inconsistent and poor results prevented this from being used to train the dataset.

Additionally, a modified AlexNet architecture was also used with all of the above epoch variations and a starting learning rate of 0.01. The two modifications done on said architecture were to lower the stride to 2 and the kernel size to 3, with the desire to see if more intricate and detailed passes over the training data would lead to superior results. Overall, these networks were chosen because of their ease of use and accessibility in Nvidia DIGITS, along with their previously discussed general success in classification tasks. Also, the expected image size of both networks was appropriate for adequate definition of the study images, without significantly increasing runtime.

Data Acquisition:

Forming the data set for this particular project solely involved collecting and editing publically-available images on Google Images. The searches used ranged from the basic, such as, “human face”, to more defined ones, such as, “person face _____”, where the last term was filled by the desired emotion. Since the implemented neural networks both require images of size 256 x 256 pixels, initially the Google search itself was filtered to show only images of that specific, or Icon, size. The number of images available with this search criteria were depleted rather quickly, so the aforementioned search filter was removed. Instead, images with desired faces were found and subsequently cropped to the needed 256 x 256 pixel size using Microsoft Paint. The following is an example of an image used in the data set:

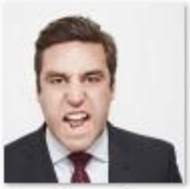
For this data set, four classes were chosen, involving three basic, distinct emotions and a baseline: happy, angry, sad, and neutral. In regards to the images themselves, only full-face, frontal-view images were chosen. The percentage of the entire image space that a face would take up was allowed to vary, as was the orientation of the face itself. Furthermore, faces were chosen irrespective of gender or race, attempting to provide as varied of a dataset as possible; however, variation in this was still expected, since the face images were not explicitly counted for equal gender and race variation. Additionally, the images were selected irrespective of color pattern, which resulted in a mix of both RGB and greyscale images present in the data set.

Using the search method described above, 100 images were found for each of the four classes. In order to expand the dataset, each image was then consecutively turned 90 degrees in order to provide an additional three images for each existing image, as such:

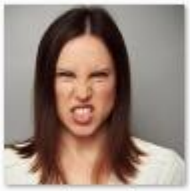
This raised the image count to 400 per category. Following this, the color scheme of each image was inverted, effectively doubling the dataset to 800 per class. In total, the data set consisted of 3,200 images. The following is an example of the inverted color image:

By the end of the project, four different data sets were created in DIGITS, two with image type Color and two with image type Grayscale. Within those pairs, one set was designated to have the default 25% of the total image set used for validation, and one with only 10 % used for validation.

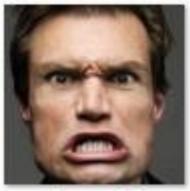
Lastly, a file containing 16 test images was created, containing four images of each class. The following is the 16 test images:



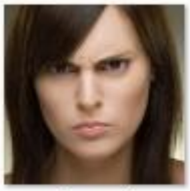
Angry 1



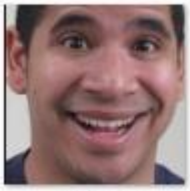
Angry 2



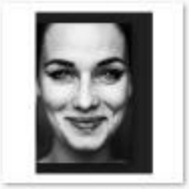
Angry 3



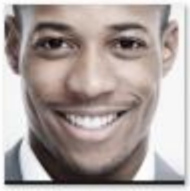
Angry 4



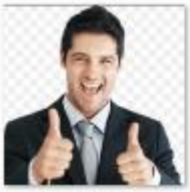
Happy 1



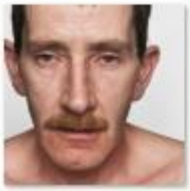
Happy 2



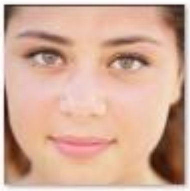
Happy 3



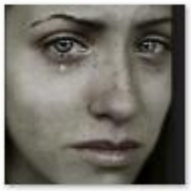
Happy 4



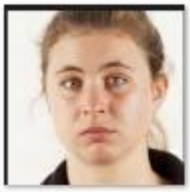
Neutral 1



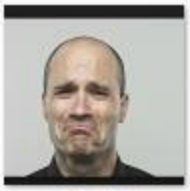
Neutral 2



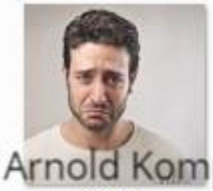
Neutral 3



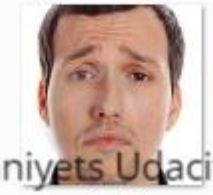
Neutral 4



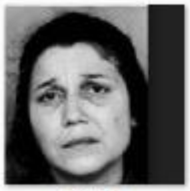
Sad 1



Sad 2



Sad 3

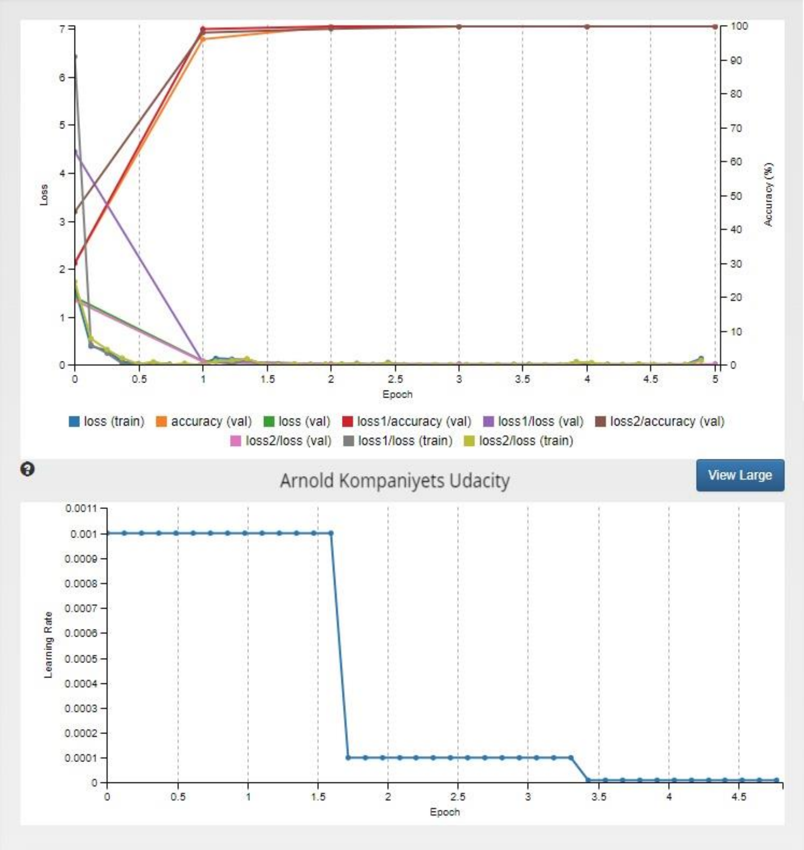


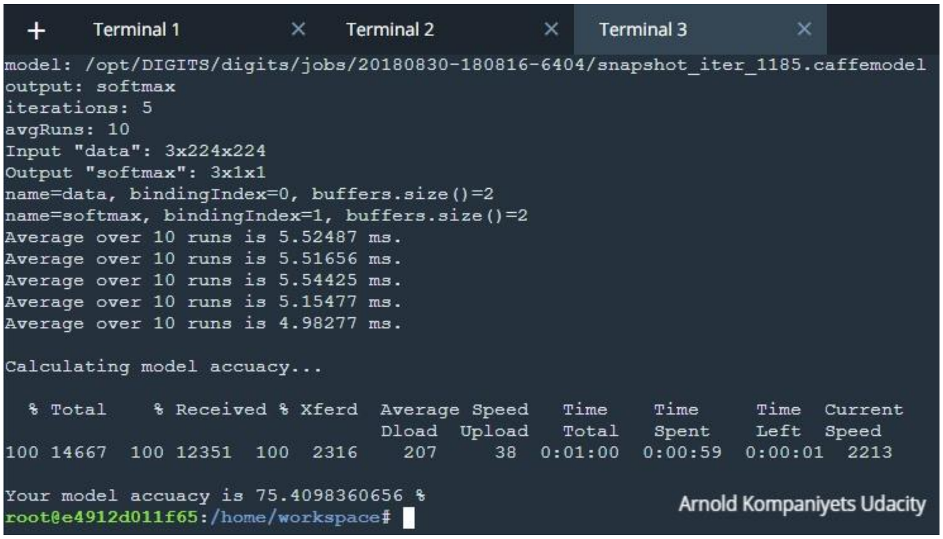
Sad 4

Results:

Provided Data Set:

To start, the supplied data set was tested, with a starting learning rate of 0.001:



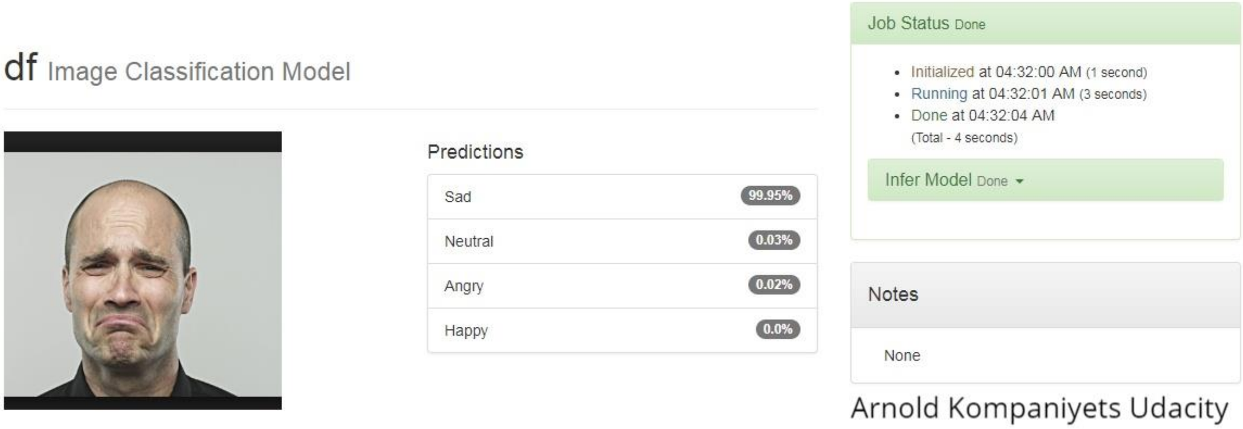


```
model: /opt/DIGITS/digits/jobs/20180830-180816-6404/snapshot_iter_1185.caffemodel
output: softmax
iterations: 5
avgRuns: 10
Input "data": 3x224x224
Output "softmax": 3x1x1
name=data, bindingIndex=0, buffers.size()=2
name=softmax, bindingIndex=1, buffers.size()=2
Average over 10 runs is 5.52487 ms.
Average over 10 runs is 5.51656 ms.
Average over 10 runs is 5.54425 ms.
Average over 10 runs is 5.15477 ms.
Average over 10 runs is 4.98277 ms.

Calculating model accuacy...

  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100 14667  100 12351  100  2316    207     38  0:01:00  0:00:59  0:00:01  2213

Your model accuacy is 75.4098360656 %
root@e4912d011f65:/home/workspace#
```

With the provided data set, five epochs proved sufficient to achieve the needed 75% minimum accuracy. The accuracy of the training data was already at virtually 100% by the second epoch, which produced faster inference times, the accuracy was not up to par. By the end of the fifth epoch, the inference times proved to still be within the desired range, and the accuracy of the test data showed to be above 75%. Lowering the starting learning rate to 0.001 still resulted in very rapid learning of the training data, but overfitting was not observed.

Project Data Set:

Next, attaining results for the emotion classification project was performed. After training a given neural network, all of the 16 test images were classified at four different intervals of neural network training: 10, 40, 80, and 115 epochs. The result for one such image would appear like the following:



df Image Classification Model

Predictions

Sad	99.95%
Neutral	0.03%
Angry	0.02%
Happy	0.0%

Job Status Done

- Initialized at 04:32:00 AM (1 second)
- Running at 04:32:01 AM (3 seconds)
- Done at 04:32:04 AM (Total - 4 seconds)

Infer Model Done

Notes

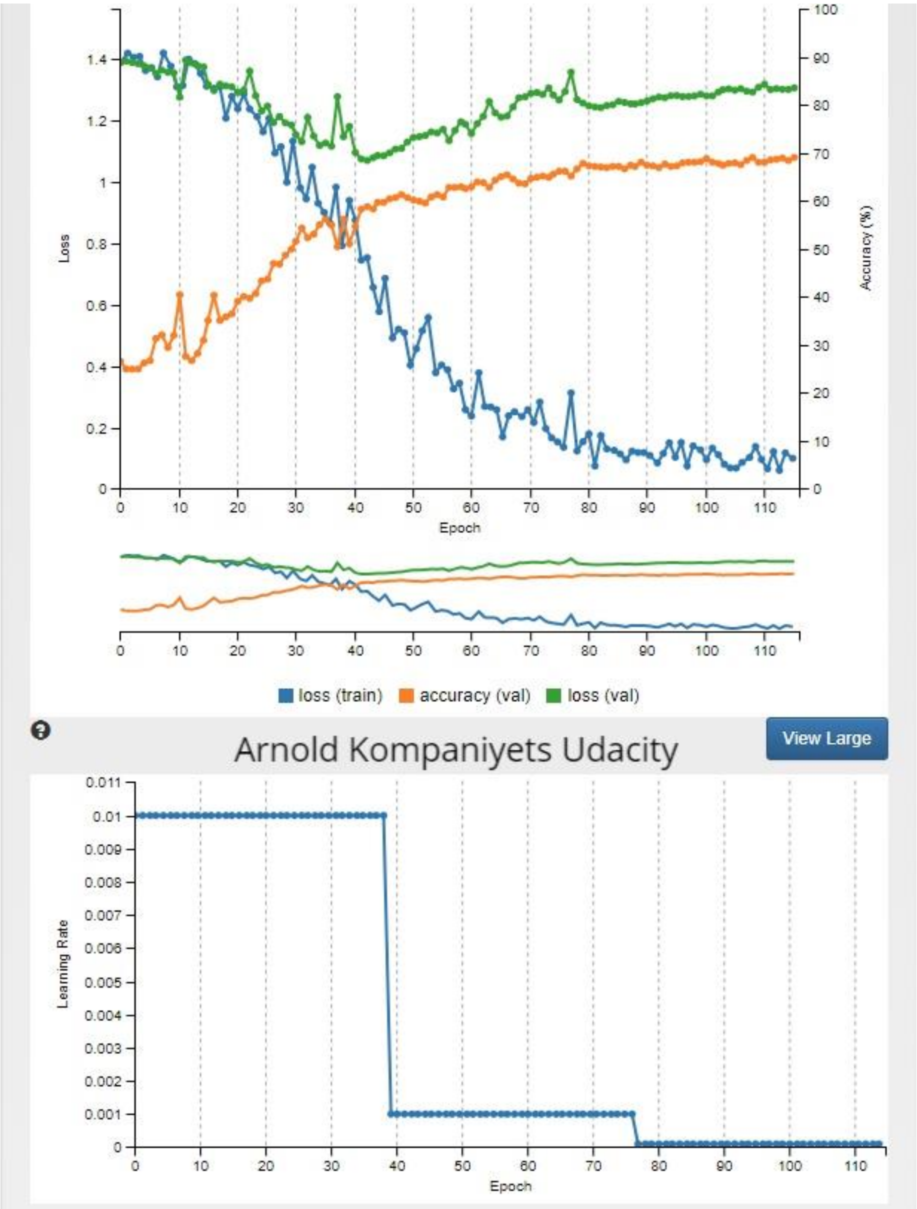
None

Arnold Kompaniyets Udacity

With each of the 16 images classified for each interval, the percentage correct was recorded (for each emotion, as well as in total).

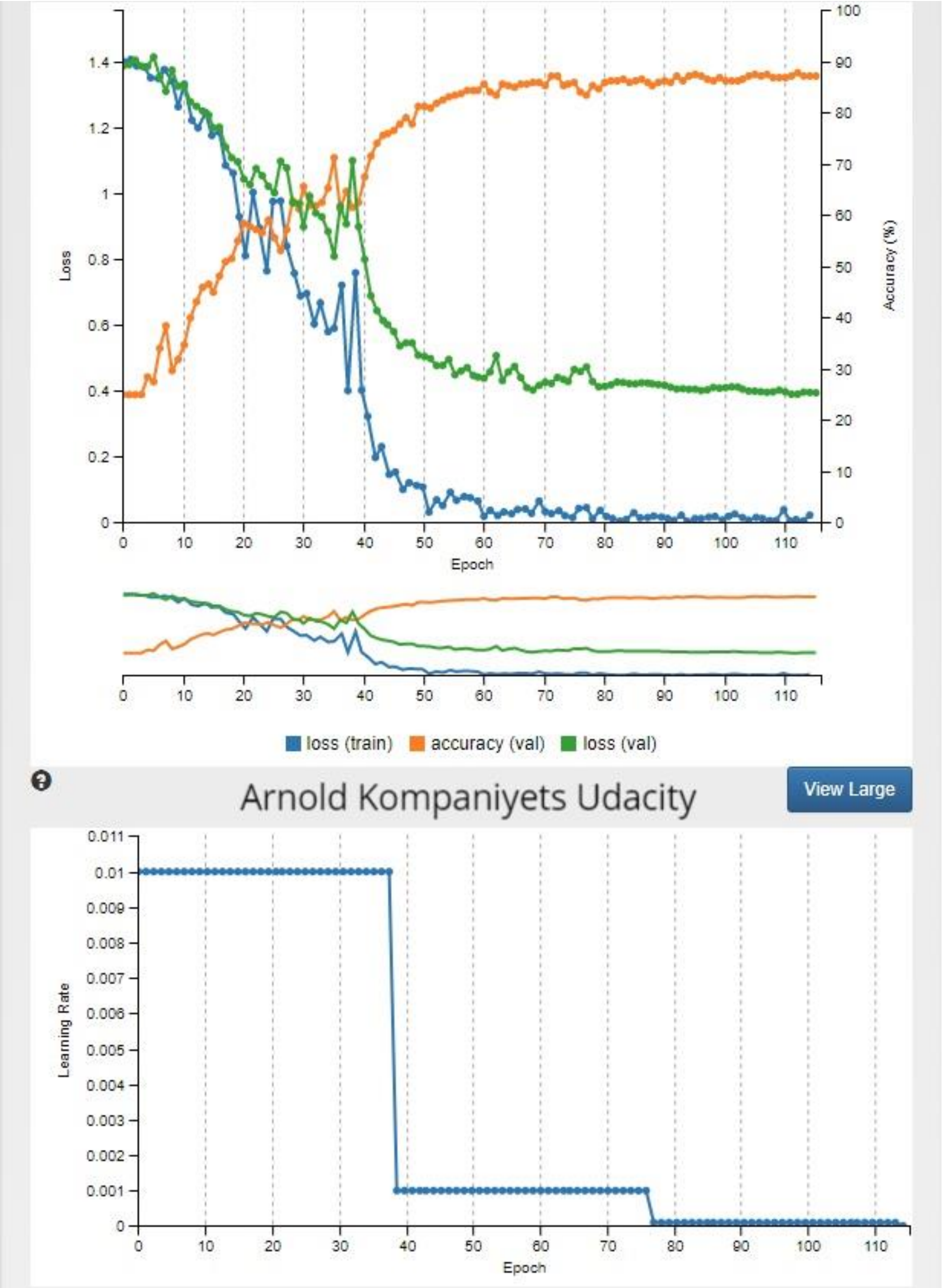
To start, the data sets were run through AlexNet, with 0.01 starting learning rate:

Data Set: Color (25% used for validation)



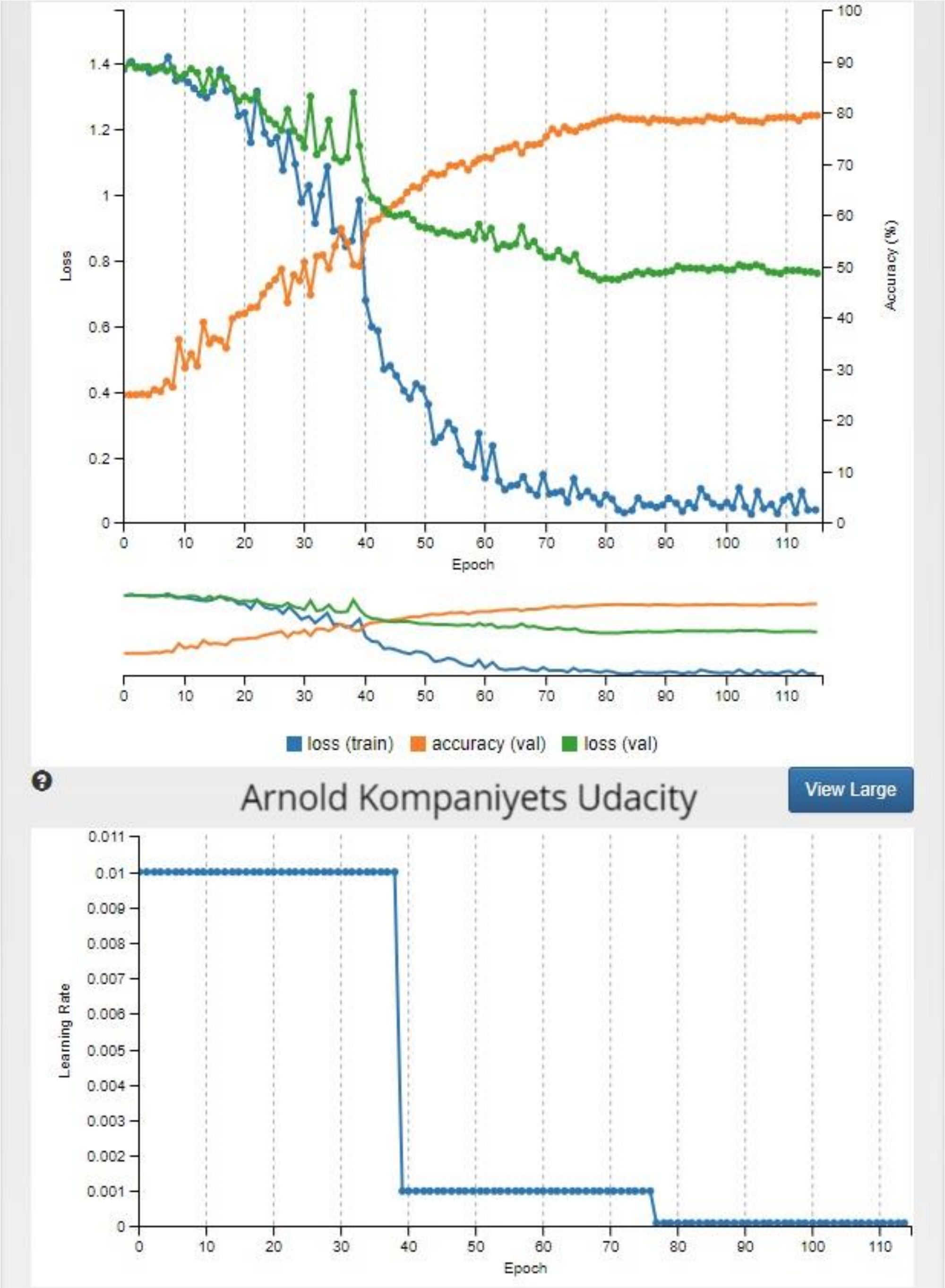
Epoch →	10	40	80	115
Angry	0%	25%	25%	25%
Happy	25%	0%	25%	25%
Sad	50%	25%	50%	50%
Neutral	25%	25%	0%	0%
Total	25%	19%	25%	25%

<u>Data Set:</u> Color (10% used for validation)



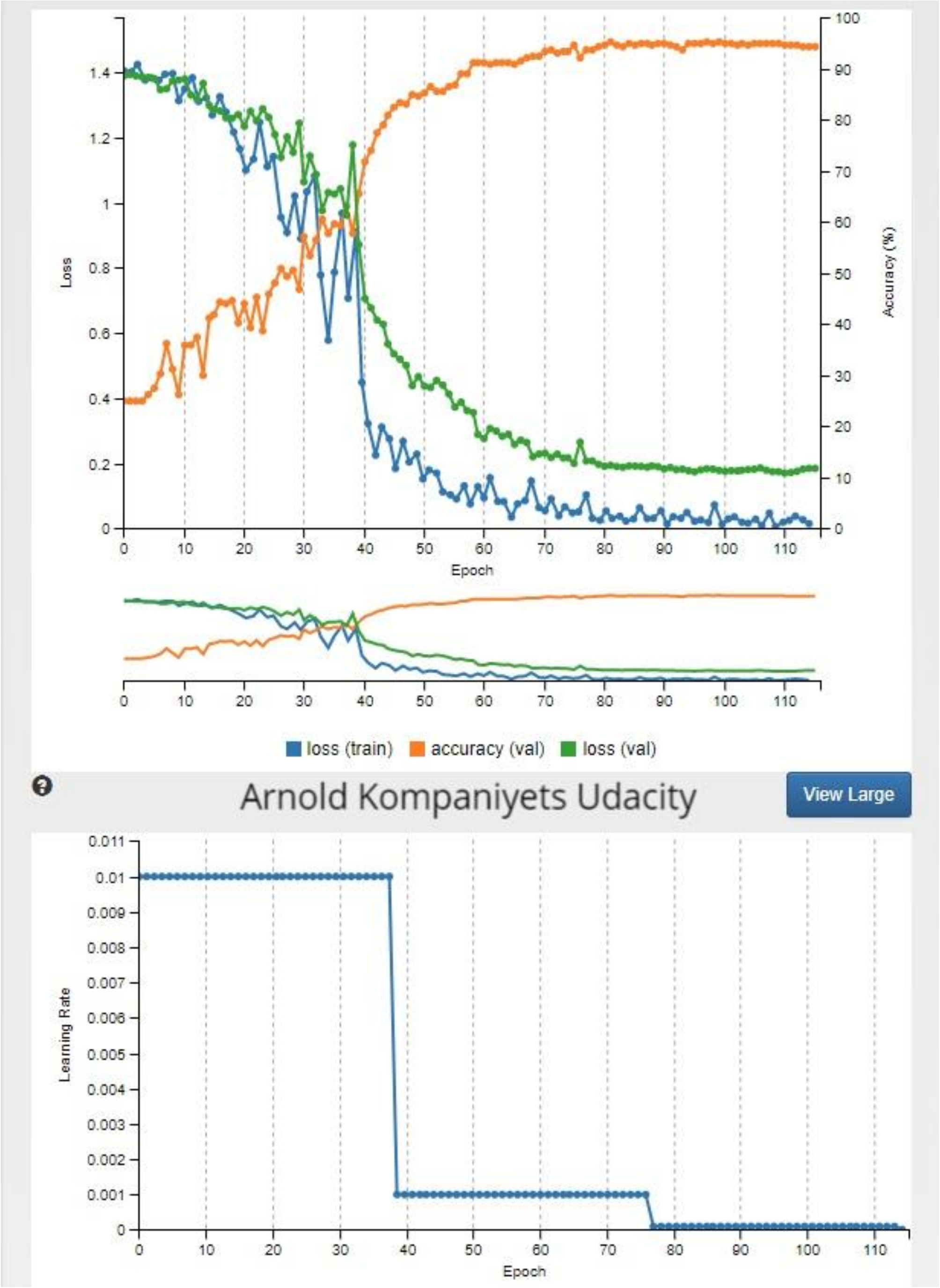
Epoch →	10	40	80	115
Angry	0%	25%	0%	0%
Happy	0%	25%	50%	50%
Sad	100%	25%	50%	50%
Neutral	25%	0%	25%	25%
Total	31%	19%	31%	31%

Data Set: Grayscale (25% used for validation)



Epoch →	10	40	80	115
Angry	0%	25%	50%	50%
Happy	50%	25%	25%	25%
Sad	25%	50%	75%	50%
Neutral	100%	50%	75%	75%
Total	44%	38%	56%	50%

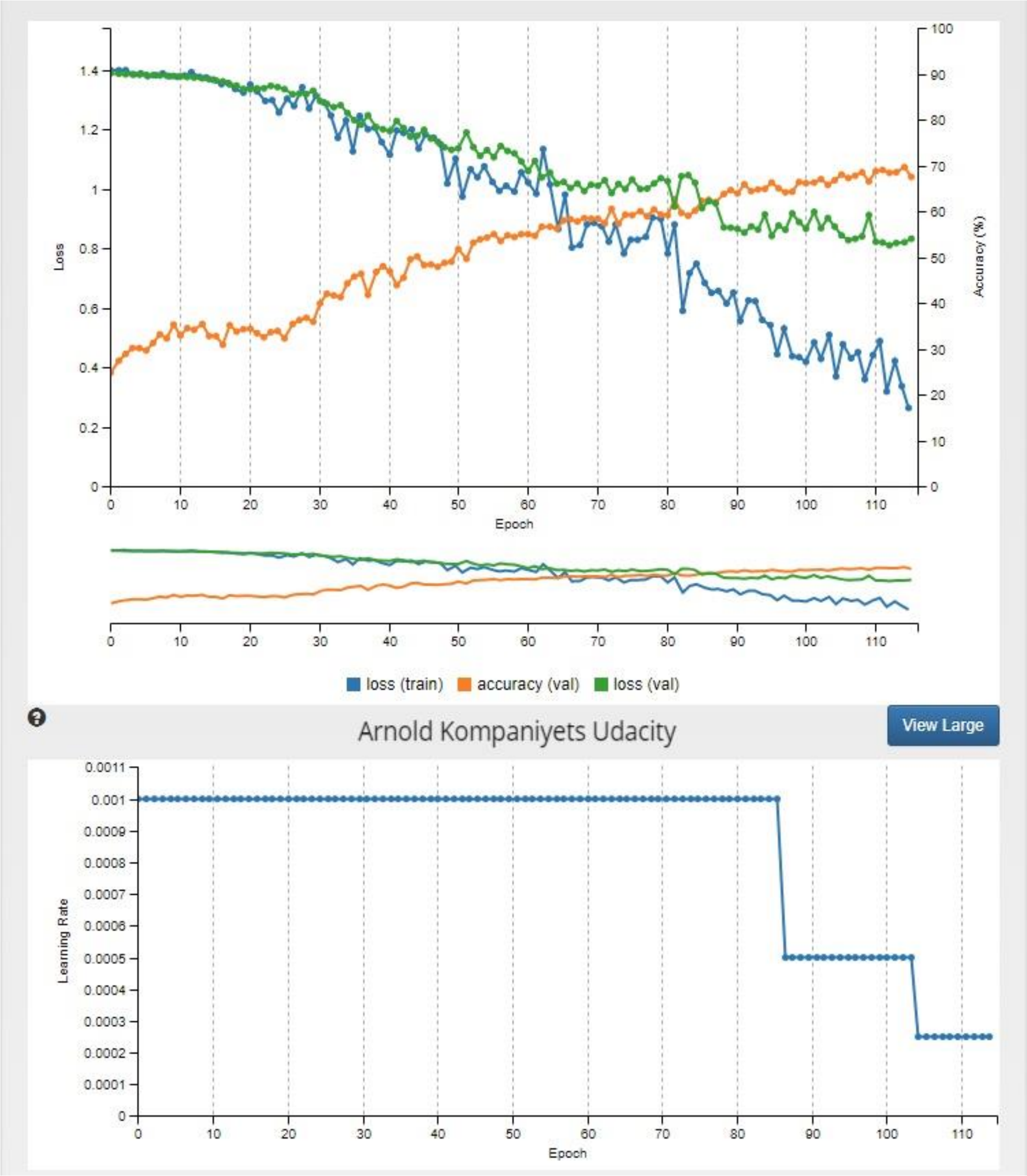
Data Set: Grayscale (10% used for validation)



Epoch →	10	40	80	115
Angry	0%	25%	25%	25%
Happy	100%	25%	75%	75%
Sad	0%	50%	75%	75%
Neutral	0%	50%	25%	25%
Total	25%	38%	50%	50%

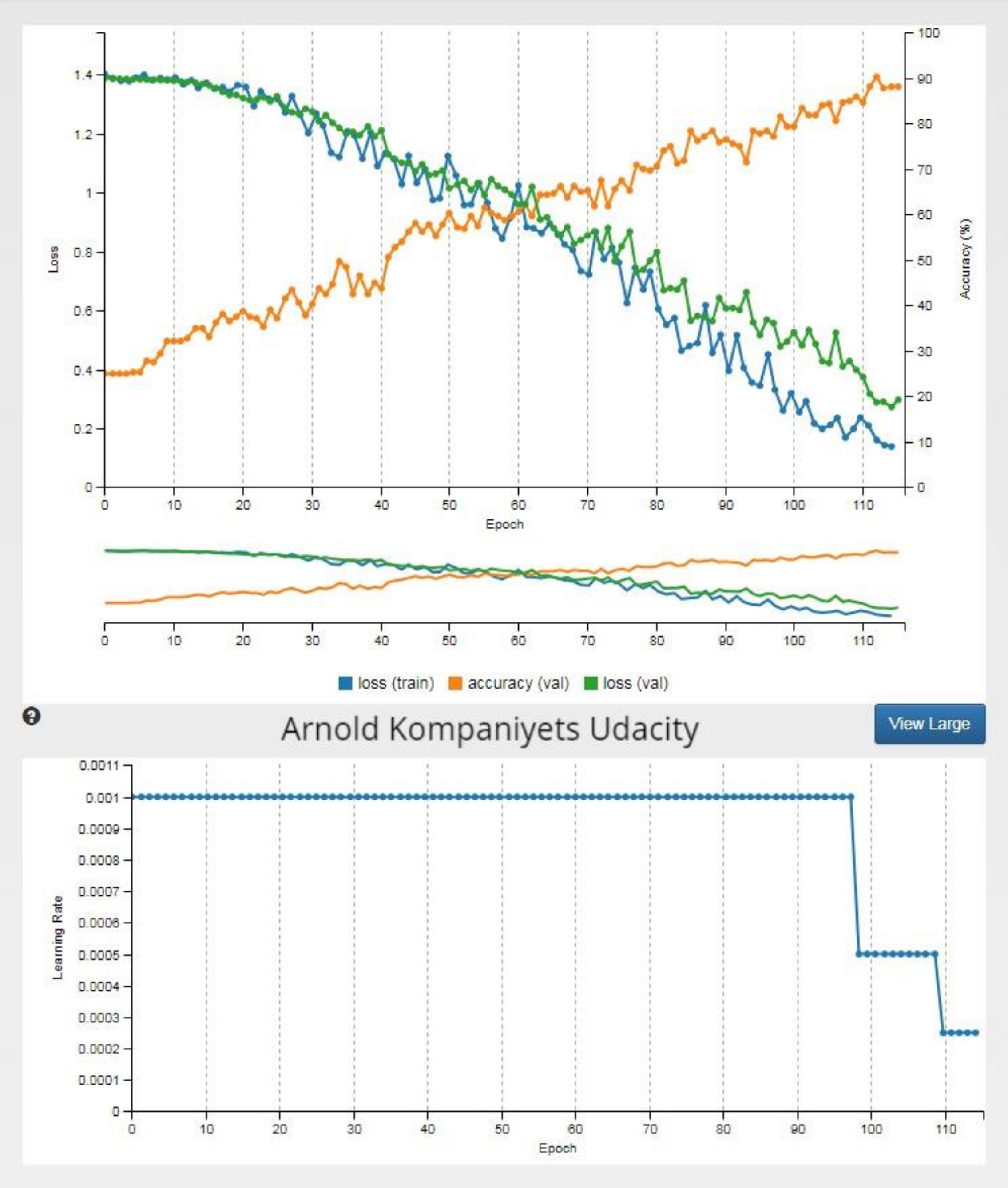
Next, each of the four data sets were once more run through AlexNet, only this time with a 0.001 starting learning rate:

Data Set: Color (25% used for validation)



Epoch →	10	40	80	115
Angry	0%	0%	25%	0%
Happy	50%	0%	50%	25%
Sad	25%	25%	25%	25%
Neutral	75%	50%	50%	50%
Total	38%	19%	38%	25%

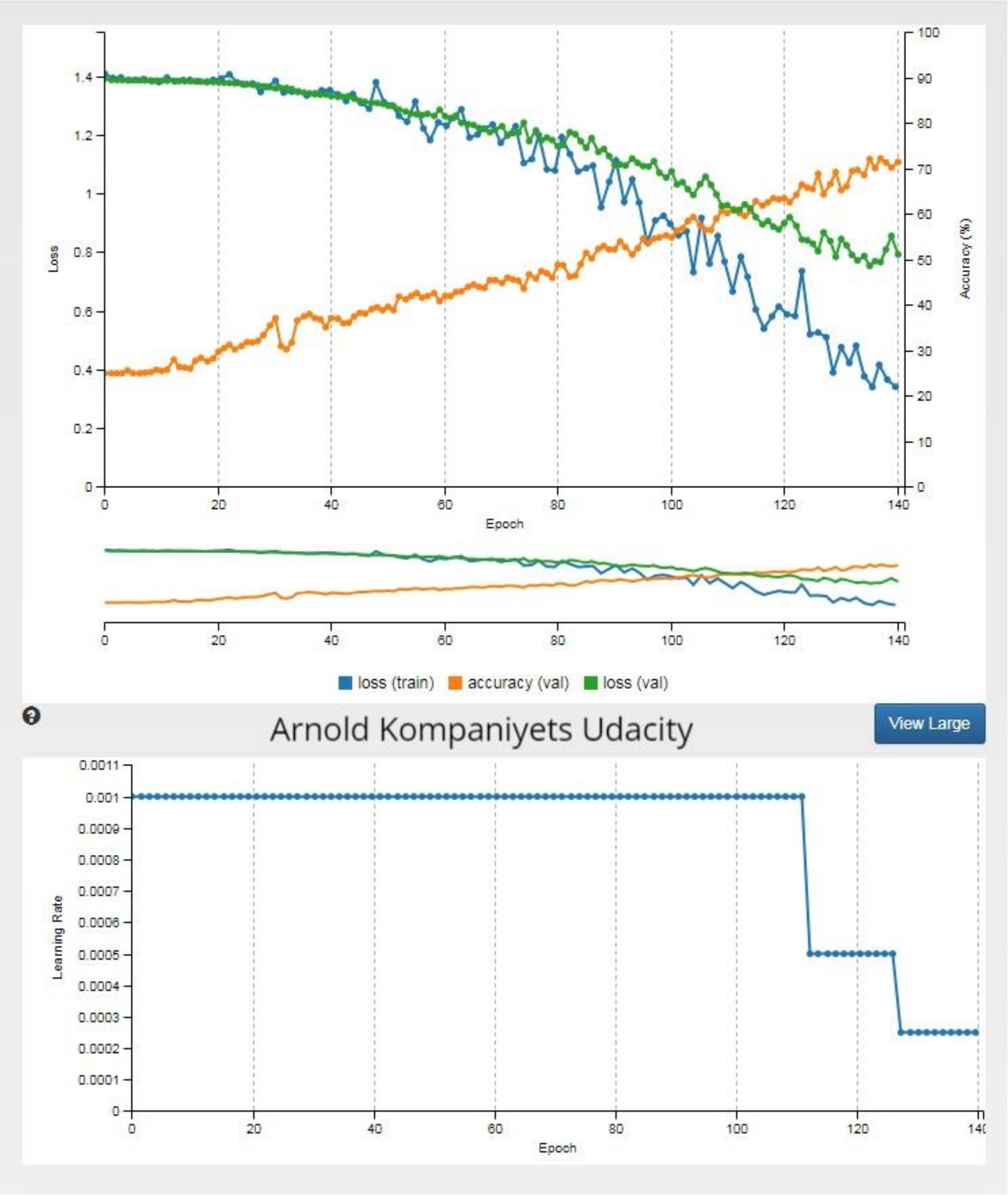
Data Set: Color (10% used for validation)



Epoch →	10	40	80	115
Angry	75%	0%	25%	50%
Happy	0%	25%	25%	50%
Sad	50%	50%	25%	50%
Neutral	50%	50%	25%	0%
Total	44%	25%	25%	38%

Data Set: Grayscale (25% used for validation)

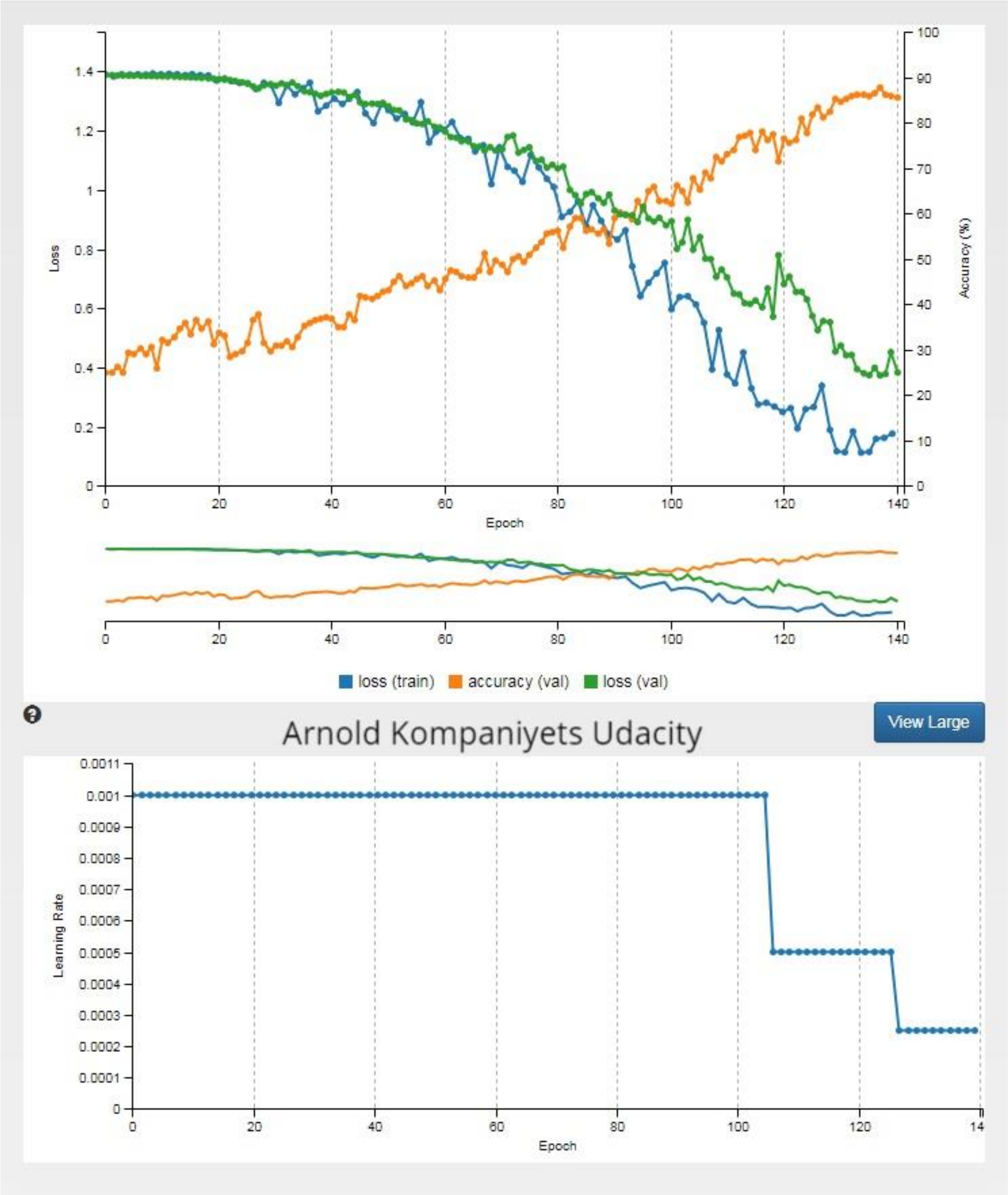
* Neural network was allowed to keep training past 115 epochs to achieve higher accuracy, but no improved scores were achieved past 115 epochs.



Epoch →	10	40	80	115
Angry	100%	0%	0%	0%
Happy	0%	50%	50%	25%
Sad	25%	25%	50%	50%
Neutral	0%	100%	75%	75%
Total	31%	44%	44%	38%

<u>Data Set:</u> Grayscale (10% used for validation)

* Neural network was allowed to keep training past 115 epochs to achieve higher accuracy, but no improved scores were achieved past 115 epochs.



Epoch →	10	40	80	115
Angry	0%	0%	25%	50%
Happy	0%	0%	0%	25%
Sad	100%	75%	50%	75%
Neutral	0%	50%	75%	50%
Total	25%	31%	38%	50%

Analysis:

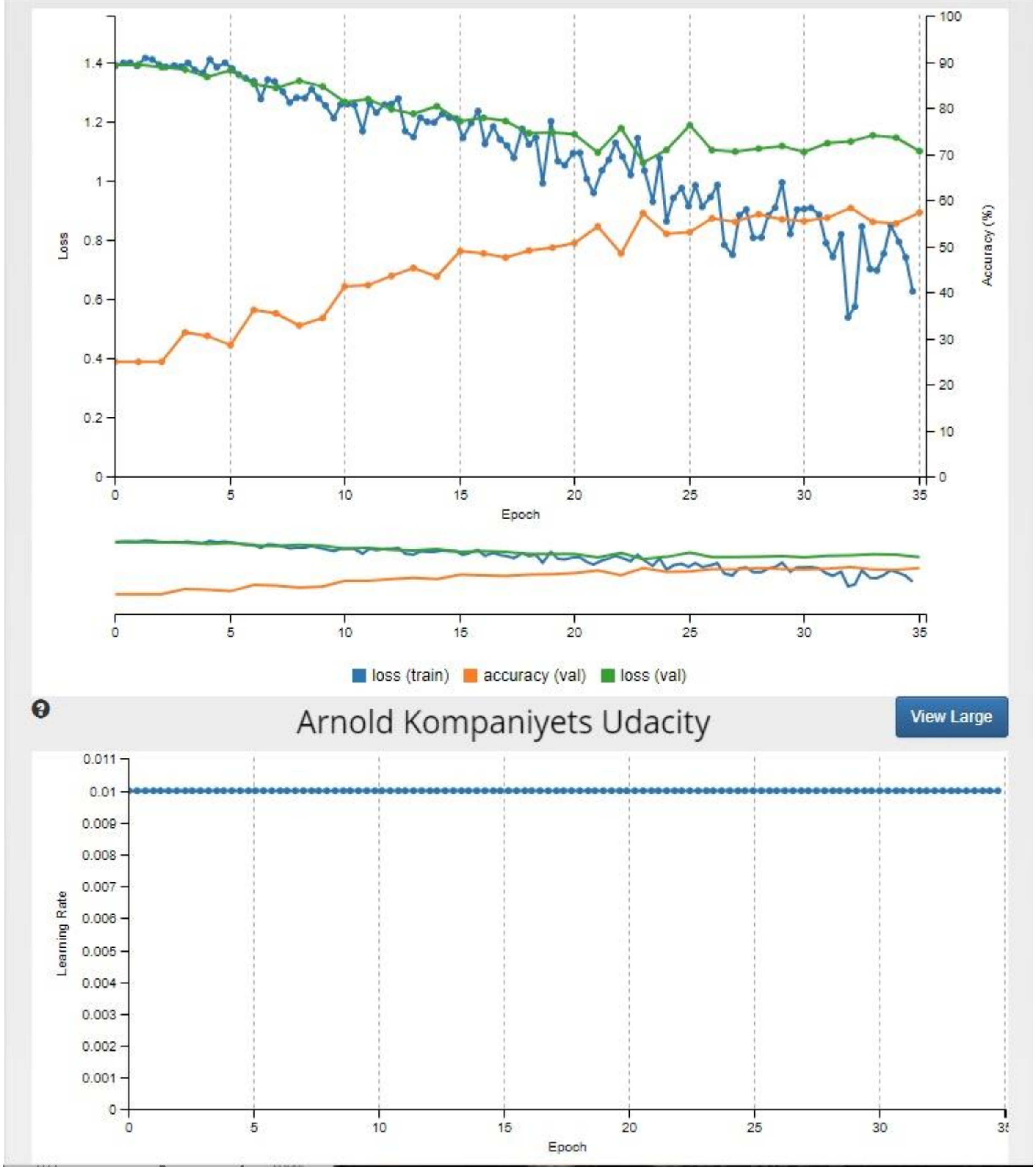
On average, each of the above eight networks took approximately 20 – 25 minutes to train, with a test image taking approximately 2 seconds to classify (according to the Nvidia DIGITS interface). In general, the results during the 10th and 40th epochs seemed to be equivalent to a matter of random chance. With both learning rates, later epoch classifications (80th and 115th) did show improved percentages. Furthermore, the Grayscale data sets outperformed the Color sets, regardless of the starting learning rate. Additionally, with all the data sets, the results achieved for the 80th epoch appeared to closely match the last epoch. The certainty of the choice expressed by the neural network would always increase from the 80th epoch to the 115th, but the ratio of correct images would remain the same. Lastly, using only 10% of total images for the validation component of the data set resulted in improved scores with the Color set, but did not make a significant different in the Grayscale set.

Next, each of the data sets was run through the modified AlexNet neural network, all with a 0.01 starting learning rate:

* Due to higher amounts of required storage, each of the four networks below were trained in sections; as such, the illustrations below reflect only the first 35 epochs

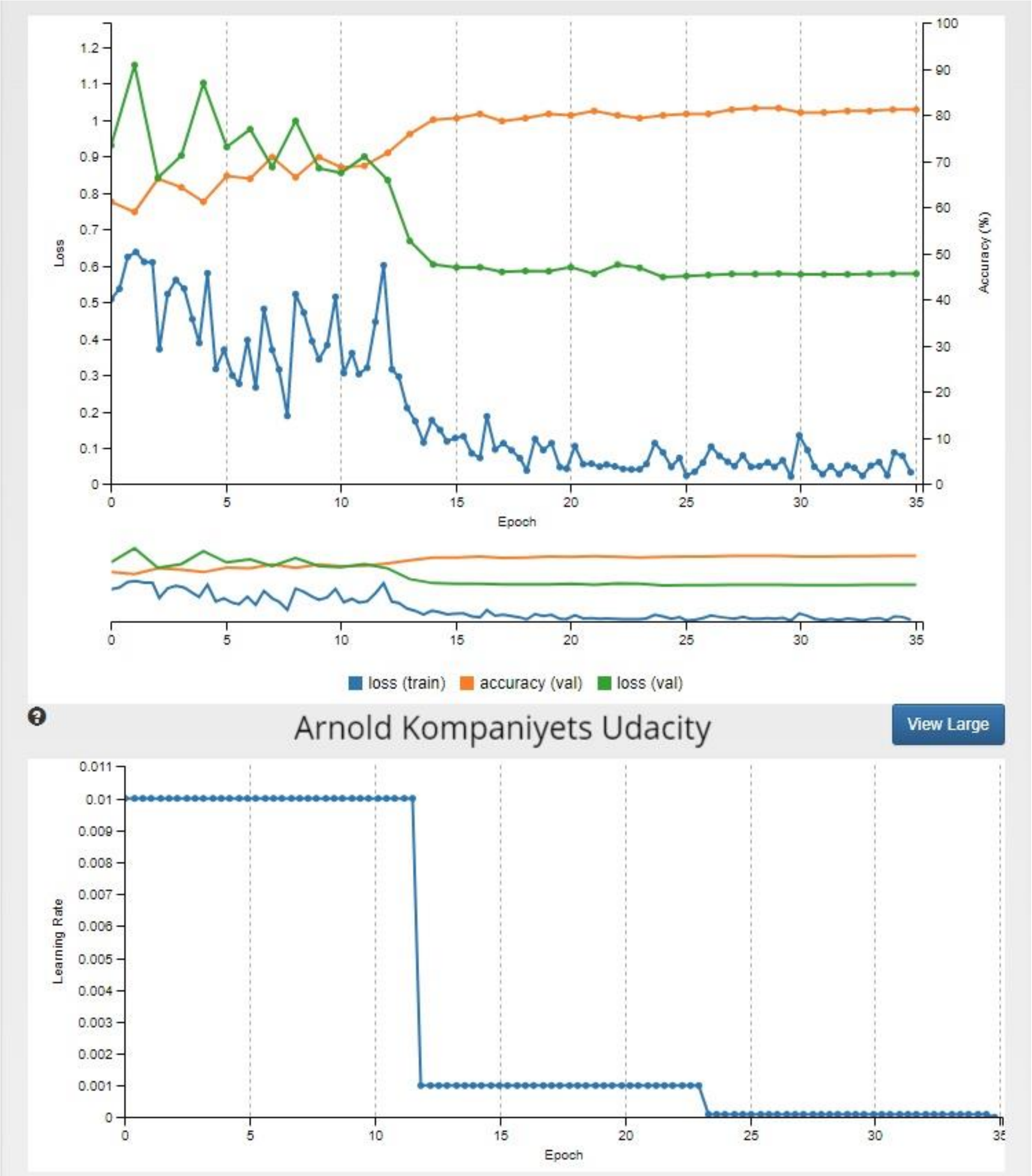
Results begin on next page:

<u>Data Set:</u> Color (25% used for validation)



Epoch →	10	40	80	115
Angry	25%	0%	50%	25%
Happy	25%	0%	25%	25%
Sad	75%	75%	25%	50%
Neutral	0%	25%	50%	0%
Total	31%	25%	38%	25%

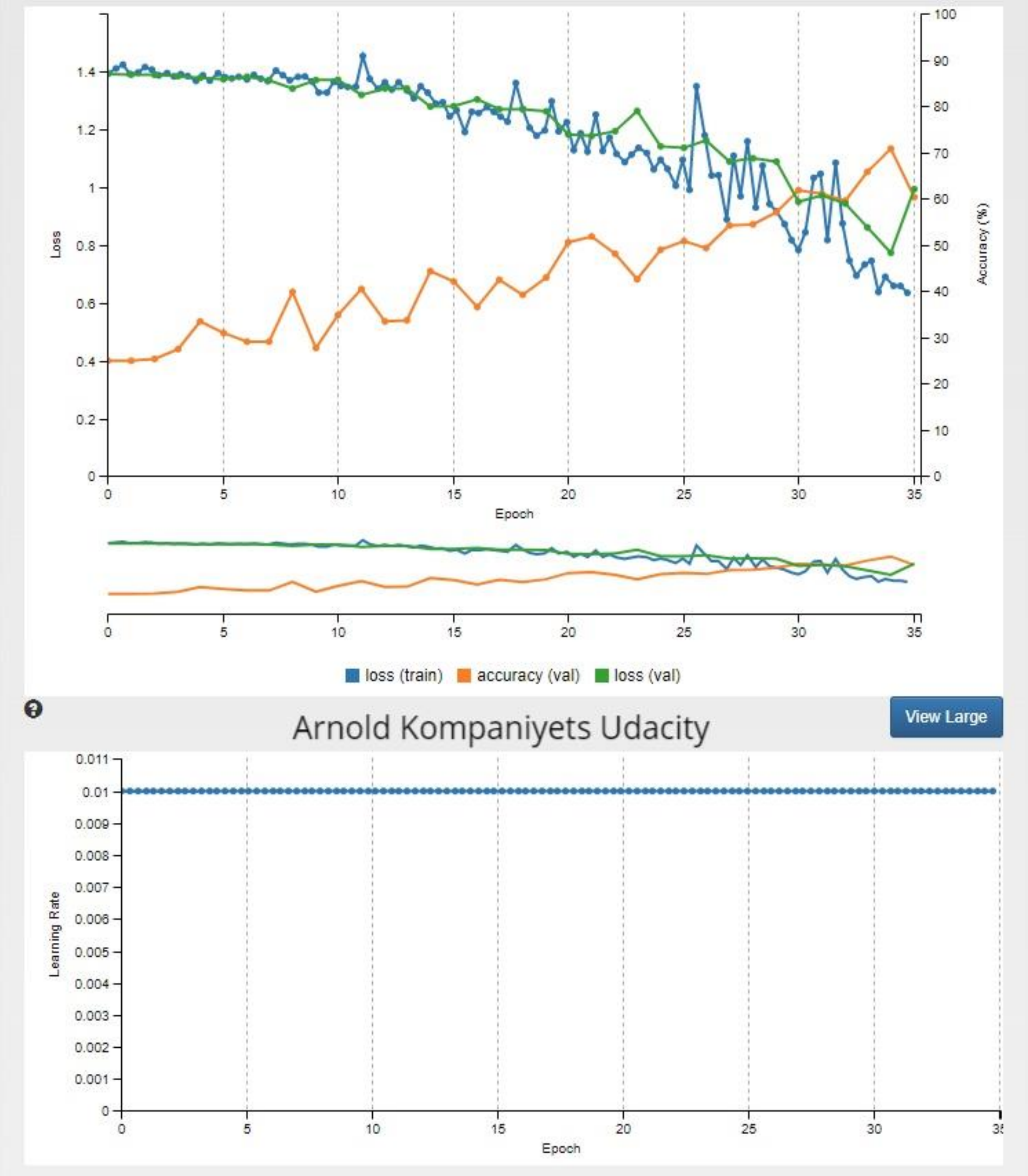
Data Set: Color (10% used for validation)



Arnold Kompaniyets Udacity

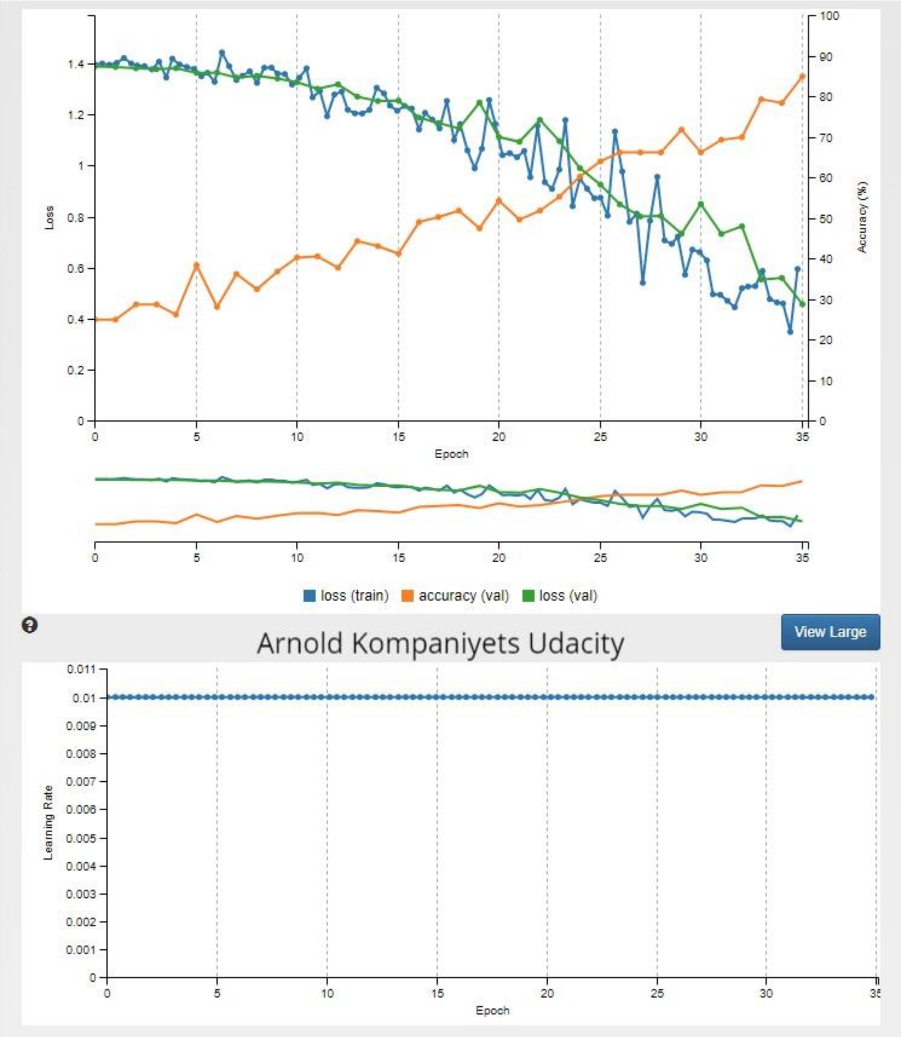
Epoch →	10	40	80	115
Angry	50%	25%	50%	50%
Happy	0%	0%	25%	25%
Sad	0%	50%	25%	25%
Neutral	0%	50%	50%	50%
Total	13%	31%	38%	38%

Data Set: Grayscale (25% used for validation)



Epoch →	10	40	80	115
Angry	0%	25%	0%	0%
Happy	75%	25%	50%	50%
Sad	0%	75%	75%	75%
Neutral	50%	50%	50%	50%
Total	31%	44%	44%	44%

<u>Data Set:</u> Grayscale (10% used for validation)

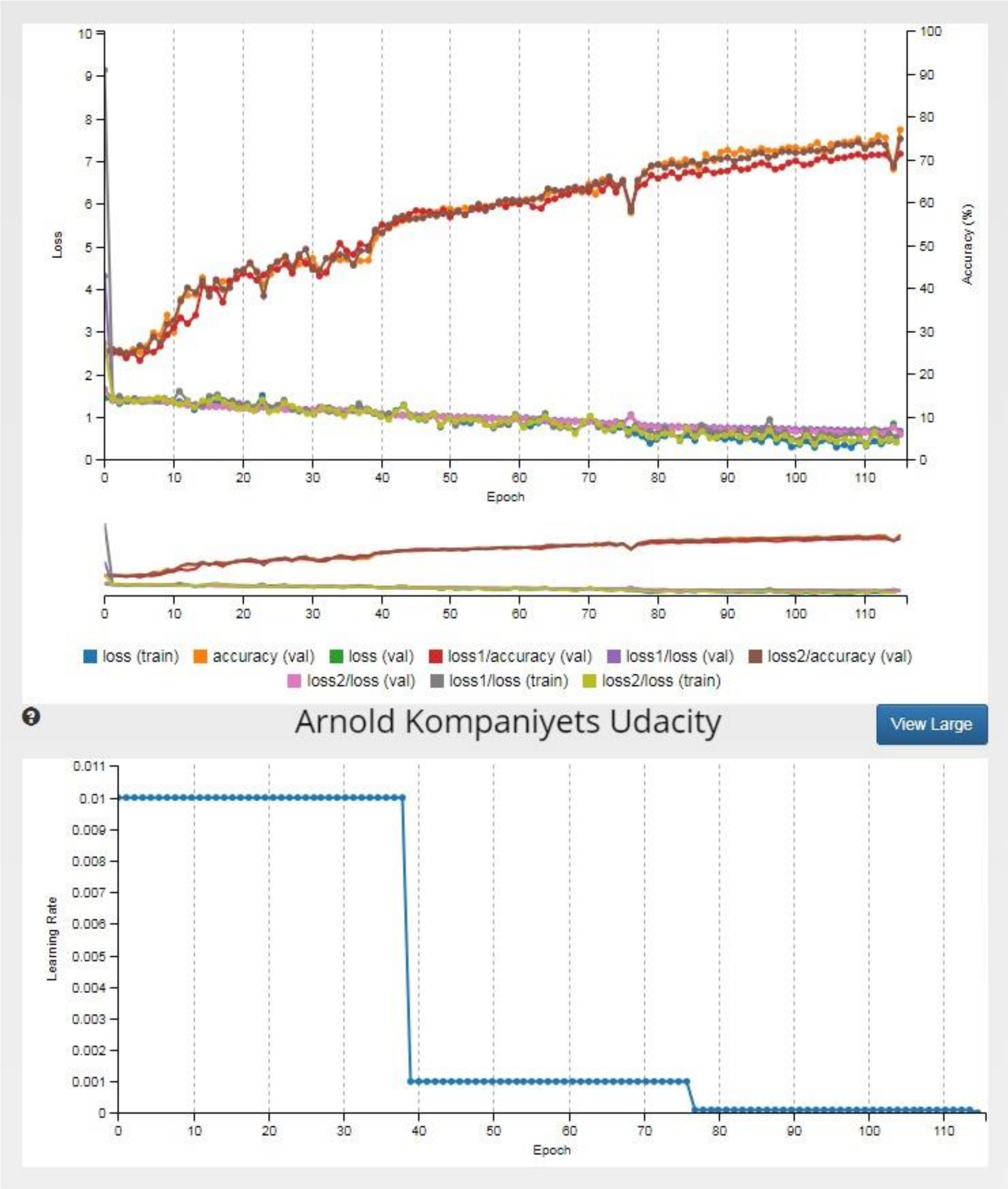


Epoch →	10	40	80	115
Angry	50%	25%	0%	25%
Happy	25%	50%	50%	50%
Sad	50%	50%	50%	50%
Neutral	25%	75%	100%	75%
Total	38%	50%	50%	50%

The custom AlexNet, although taking on average three times longer to train and classify (taking around 6 seconds to classify an image), did perform slightly worse overall. However, the patterns observed with the regular AlexNet classifications remained largely the same. The Grayscale data set did perform superiorly, with the correct/incorrect patterns of the 80th epoch almost identical to that of the 115th epoch. Also, using only 10% of the images for validation resulted in improved scores in both types of data sets.

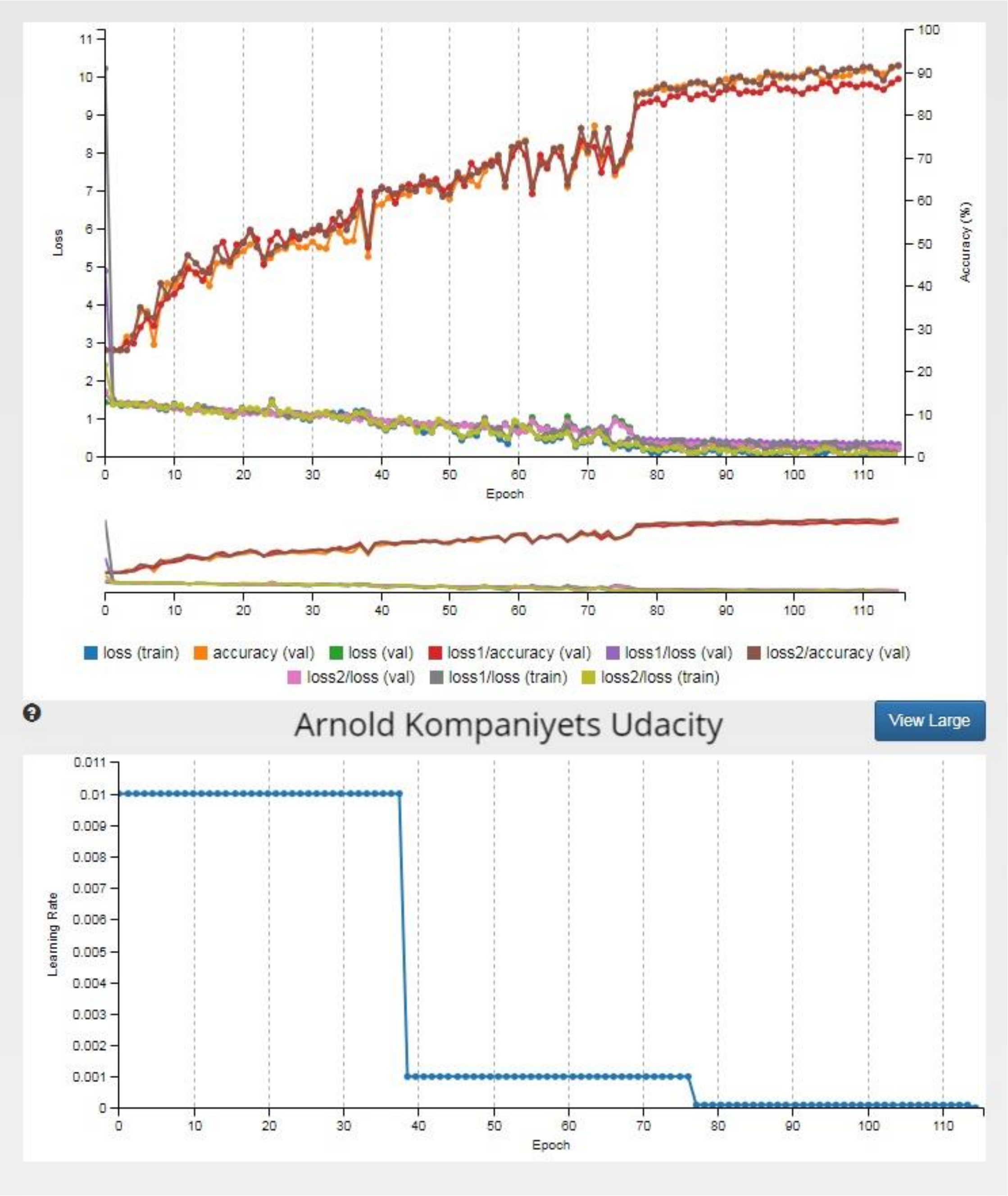
The third component of the experiment involved running all four data sets through GoogLeNet. A 0.01 starting learning rate was used first:

<u>Data Set:</u> Color (25% used for validation)


Loss
0
1
2
3
4
5
6
7
8
9
10
Accuracy (%)
0
10
20
30
40
50
60
70
80
90
100
Epoch
0
10
20
30
40
50
60
70
80
90
100
110
■ loss (train) ■ accuracy (val) ■ loss (val) ■ loss1/accuracy (val) ■ loss1/loss (val) ■ loss2/accuracy (val)
■ loss2/loss (val) ■ loss1/loss (train) ■ loss2/loss (train)
Arnold Kompaniyets Udacity
View Large
Learning Rate
0
0.001
0.002
0.003
0.004
0.005
0.006
0.007
0.008
0.009
0.01
0.011

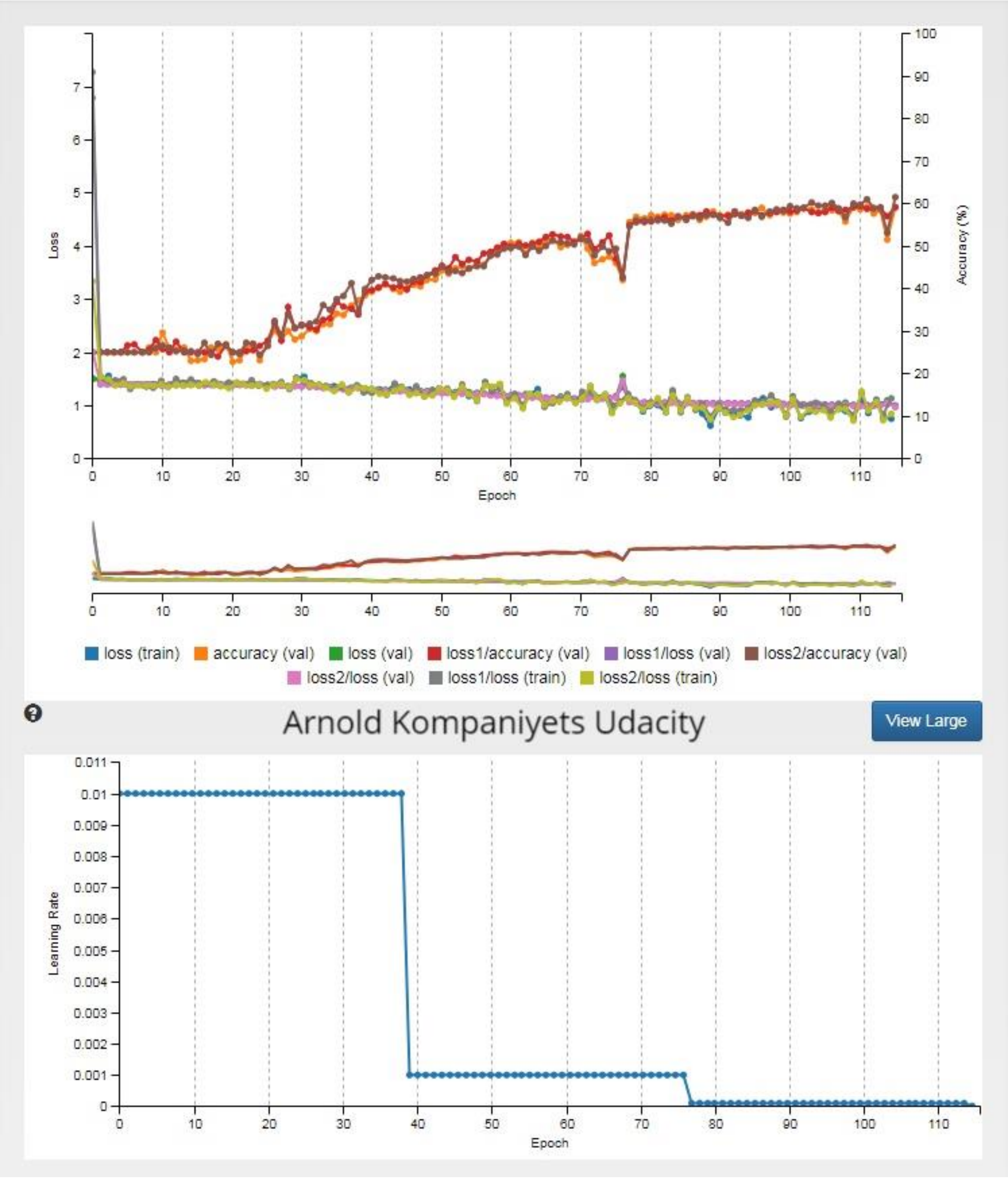
Epoch →	10	40	80	115
Angry	100%	0%	25%	25%
Happy	0%	0%	0%	0%
Sad	0%	50%	75%	50%
Neutral	0%	0%	0%	0%
Total	25%	13%	25%	19%

<u>Data Set:</u> Color (10% used for validation)



Epoch →	10	40	80	115
Angry	0%	0%	25%	25%
Happy	0%	0%	0%	0%
Sad	50%	75%	50%	50%
Neutral	0%	0%	25%	25%
Total	13%	19%	25%	25%

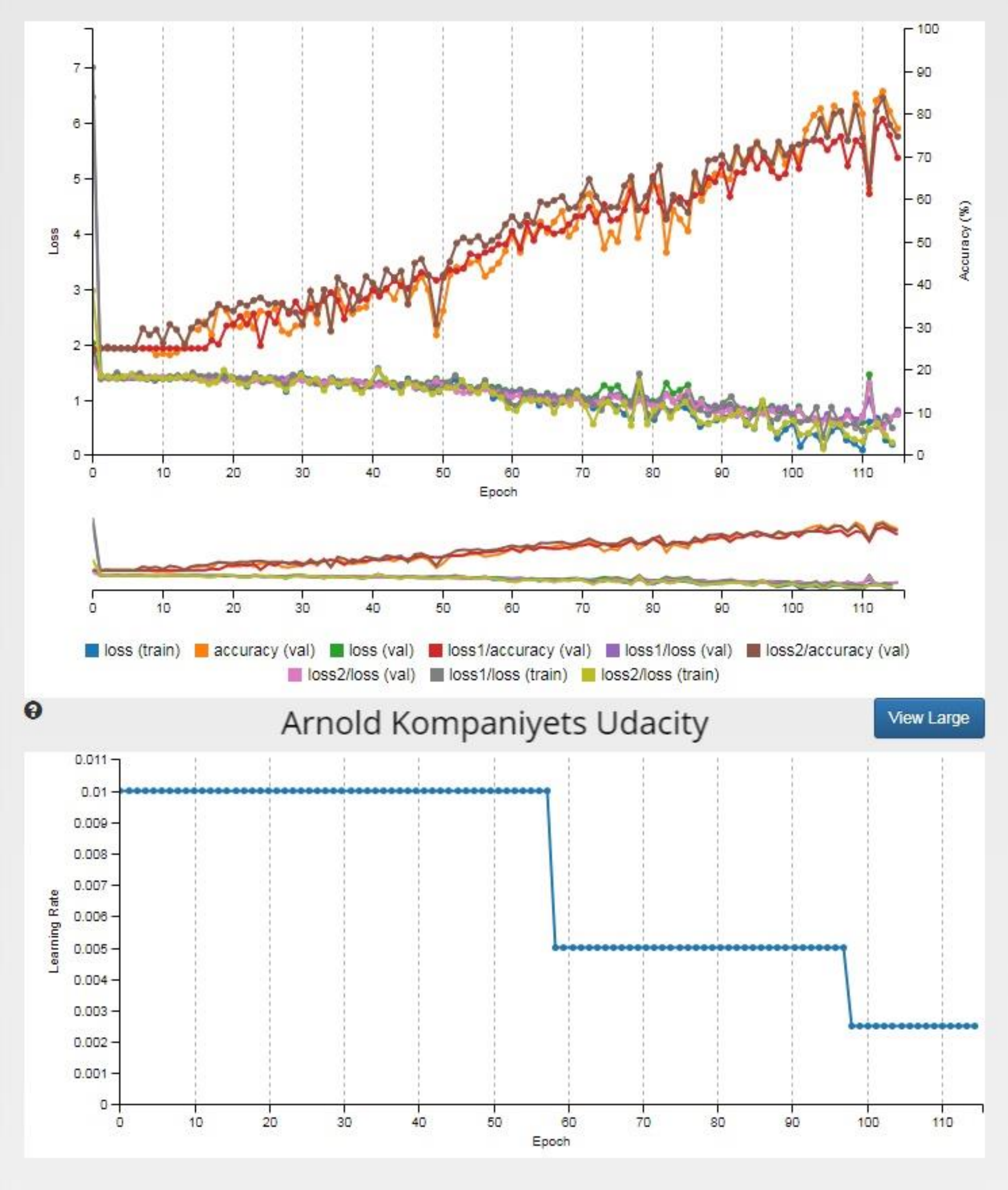
<u>Data Set:</u> Grayscale (25% used for validation)



Epoch →	10	40	80	115
Angry	0%	25%	0%	25%
Happy	100%	25%	50%	50%
Sad	50%	100%	75%	75%
Neutral	0%	75%	50%	50%
Total	38%	44%	44%	50%

Data Set: Grayscale (10% used for validation)

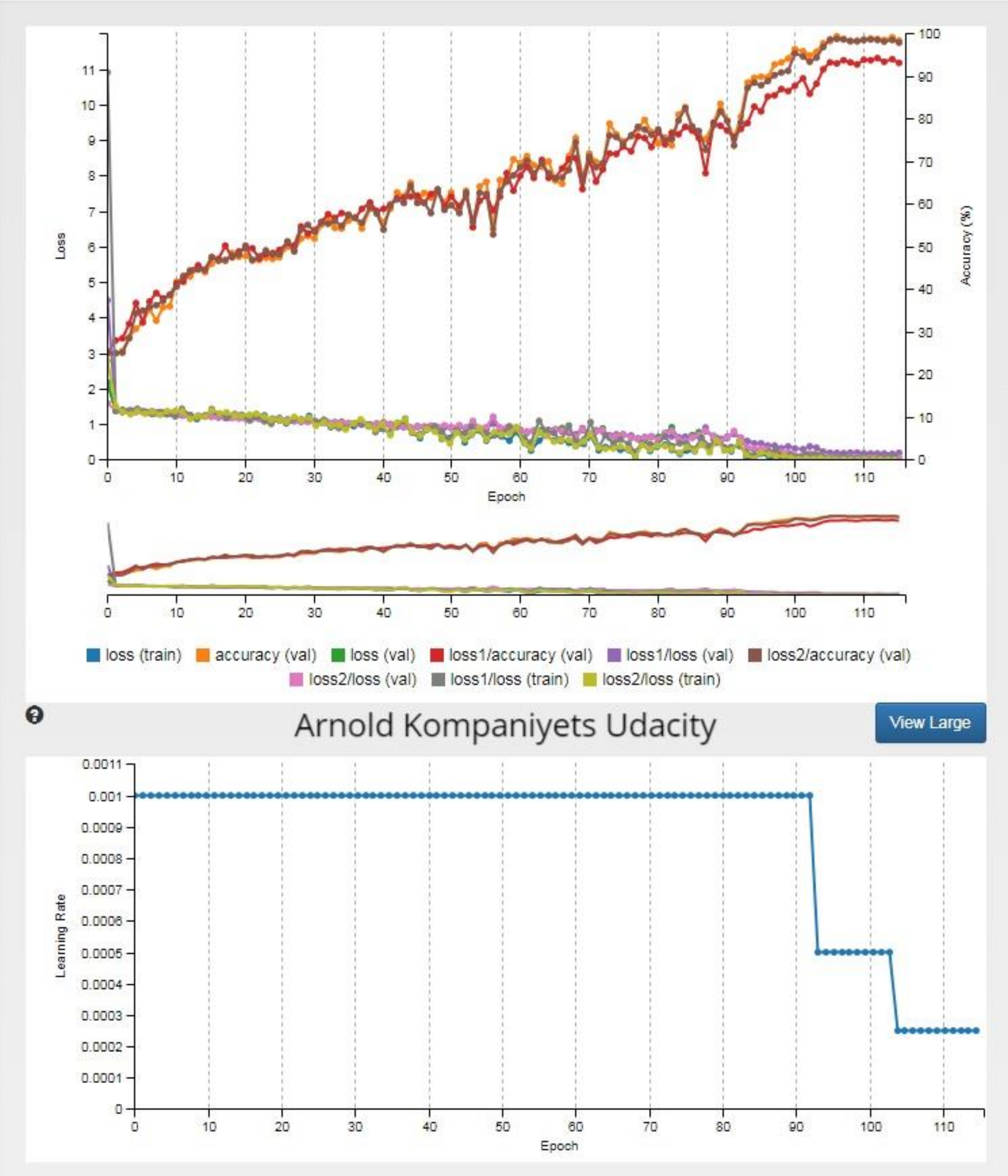
***training extended to 145 epochs in order to allow loss and accuracy values to plateau.**



Epoch →	10	40	80	115	145
Angry	0%	0%	0%	0%	25%
Happy	100%	75%	75%	50%	50%
Sad	0%	50%	75%	75%	75%
Neutral	0%	75%	75%	75%	100%
Total	25%	50%	56%	50%	63%

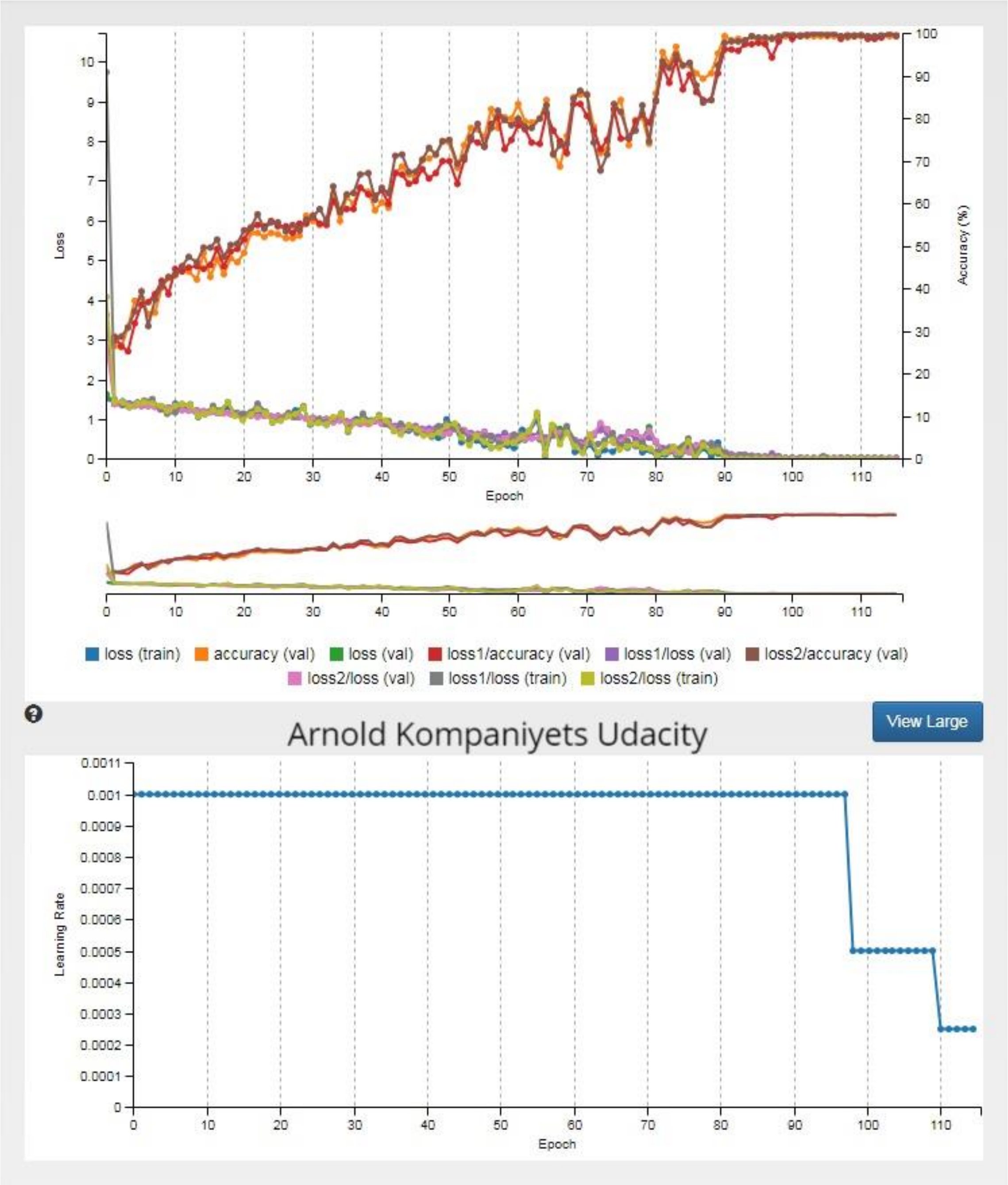
As with the AlexNet training sets, the four data sets were also run through GoogLeNet with a starting learning rate of 0.001:

Data Set: Color (25% used for validation)



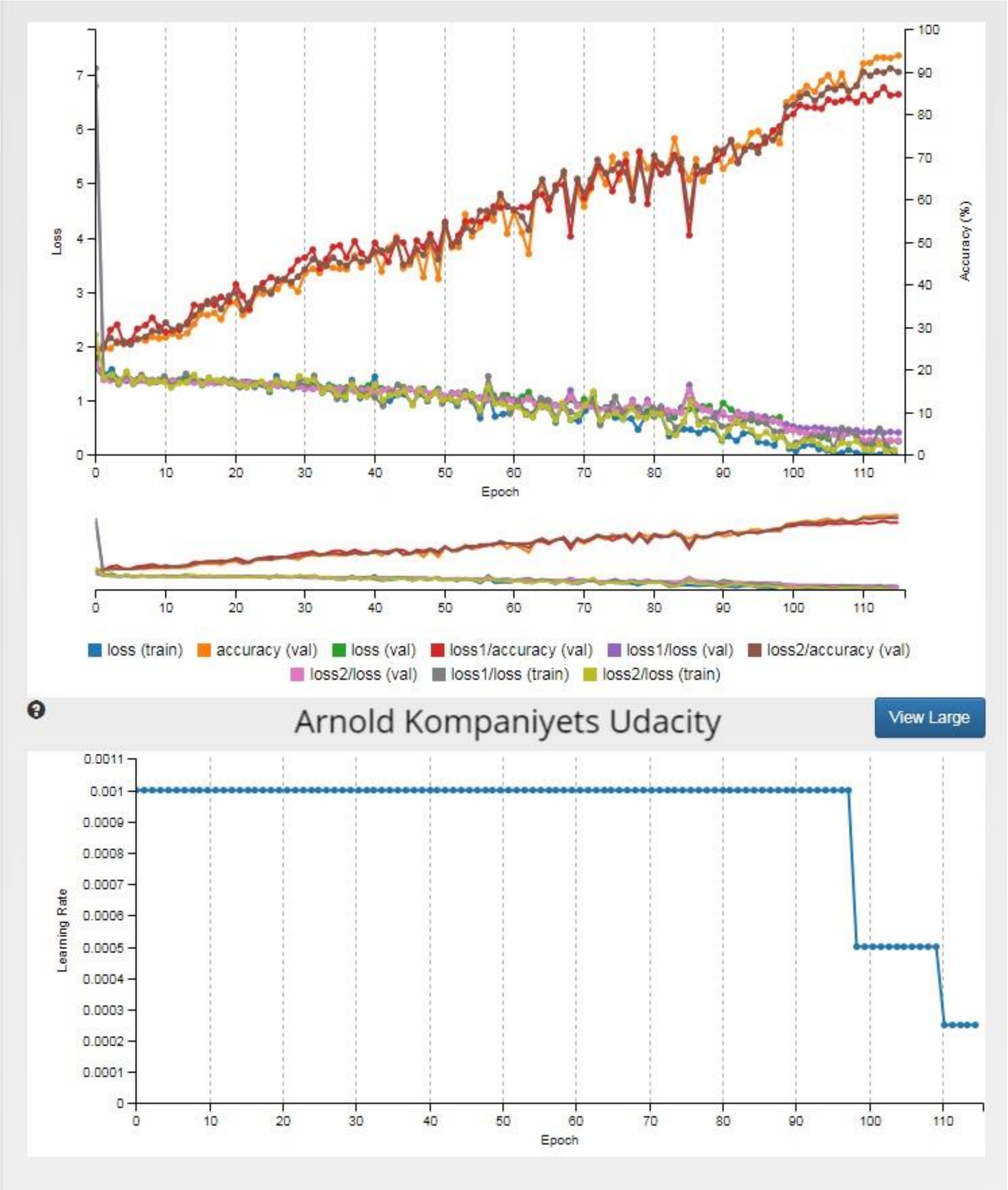
Epoch →	10	40	80	115
Angry	25%	25%	0%	25%
Happy	25%	25%	25%	25%
Sad	0%	75%	75%	75%
Neutral	0%	25%	25%	25%
Total	13%	38%	31%	38%

Data Set: Color (10% used for validation)


Loss
Accuracy (%)
Epoch
0
1
2
3
4
5
6
7
8
9
10
20
30
40
50
60
70
80
90
100
110
■ loss (train) ■ accuracy (val) ■ loss (val) ■ loss1/accuracy (val) ■ loss1/loss (val) ■ loss2/accuracy (val)
■ loss2/loss (val) ■ loss1/loss (train) ■ loss2/loss (train)
Arnold Kompaniyets Udacity
View Large
Learning Rate
0.0011
0.001
0.0009
0.0008
0.0007
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0
Epoch

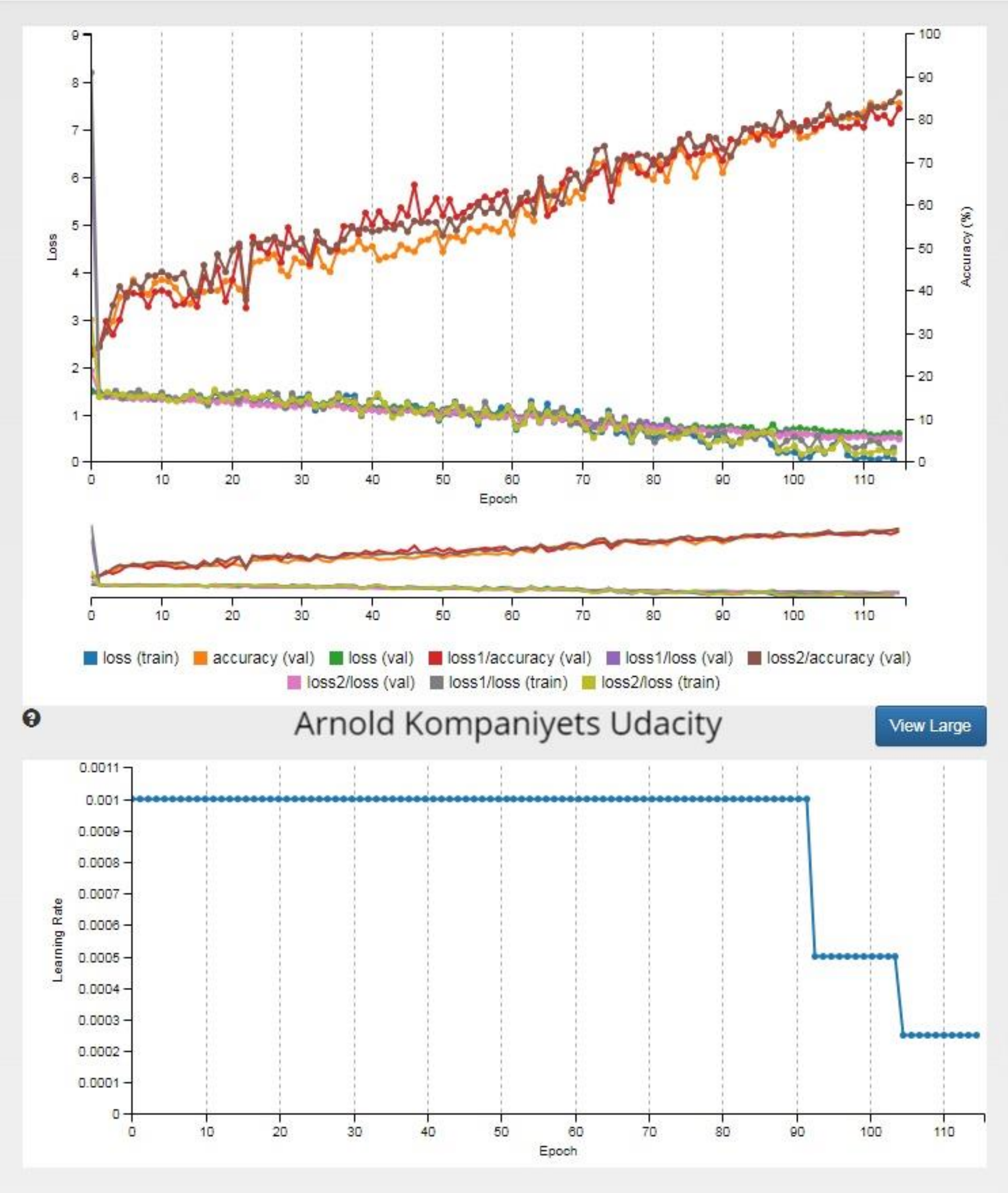
Epoch →	10	40	80	115
Angry	25%	0%	0%	0%
Happy	0%	0%	0%	0%
Sad	25%	50%	75%	50%
Neutral	50%	25%	25%	25%
Total	25%	19%	25%	19%

Data Set: Grayscale (25% used for validation)



Epoch →	10	40	80	115
Angry	0%	0%	75%	50%
Happy	0%	25%	50%	25%
Sad	100%	75%	75%	75%
Neutral	50%	50%	50%	50%
Total	38%	38%	63%	50%

<u>Data Set:</u> Grayscale (10% used for validation)



Epoch →	10	40	80	115
Angry	50%	25%	25%	50%
Happy	0%	50%	50%	50%
Sad	75%	75%	75%	75%
Neutral	0%	50%	75%	50%
Total	31%	50%	56%	56%

With GoogLeNet, the Color data sets performed very poorly, with even the fully trained networks unable to classify the test images better than chance alone. The Grayscale data sets performed significantly better, with the 10% validation sets showing slightly better accuracy. The GoogLeNet sets showed slightly greater variation from 80th to 115th epochs, although the overall percentages were similar. Specifically with the 0.001 starting learning rate sets, the trained networks performed consistently worse during the 115th epoch than during the 80th, illustrating potential overfitting. On average, an image with the trained GoogLeNet networks took 3-4 seconds to classify, with an entire network taking, on average, approximately 80 minutes to train.

Discussion:

All in all, this project was able to illustrate both the immense difficulty in correctly identifying human emotion, as well as the tremendous importance of the data set in neural network training. In forming the data set, as mentioned previously, difficulty arose in finding more than 100 different faces for each emotion using publically-available images. Therefore, the data sets were artificially expanded by using rotation and color inversion. It was thought that by doing this, the neural networks would not focus on deriving emotion from simply the orientation of a face, nor would they learn to derive emotion from an image's particular color pattern. With many rotations and color patterns present for each emotion's data set, it was hoped that the neural networks would instead choose to identify different patterns amongst each image class, specifically the angles of the eyebrows, mouth, jaw, etc., along with the shapes of the eyes, forehead, and face as a whole. Unfortunately, this did not appear to be the case.

It must be remembered that all neural networks are doing during their training is, in essence, simply solving a very complex equation. They are tasked with finding exactly which series of weights and variables will lead to (when multiplied together with the pixel RGB values of an image) proper classification of each provided and labeled image. It can certainly be hoped that the neural network will choose to pursue patterns of identification that make most sense to the human observer, but this will not always be the case. A neural network simply "sees" a series of number matrices, ranging from 0 to 255, and if it can discover a superior way to solve its equation, that specific method will be chosen. Of course, it cannot be stated with certainty the patterns that each of the neural networks above chose to focus on, but it can be definitively stated that more than just the emotional composition of the face was taken into account. Out of all the trained networks, the higher success rate achieved was 63%. This is an acceptable score, but nowhere near the desired score for successful human-robot communication. More subtle emotional expressions aside, it is very difficult for a human observer to misjudge an angry or elated face. However, especially in classifying an angry face, virtually all of the neural networks would struggle tremendously. This leads to the thinking that as much as it was tried to be avoided, the neural networks most likely trained themselves to see certain color patterns and head shapes as being characteristic of a specific emotion, which is obviously not the case for a human observer.

The factor that illustrates the above point very well is the fact that the Grayscale data sets virtually always outperformed the Color data sets. By compressing the RGB color layers to one layer, the neural networks could no longer rely as much on classifying training data according to specific color patterns. This led to much improved overall classification scores, but did not seem to eliminate the problem entirely.

Continuing, it must be acknowledged that having a neural network even find the correct patterns to classify emotions are a difficult task. It may seem obvious to a human observer, but to a neural network, all of the images provided would look very similar – like a human face. If asked to differentiate between a human face, a giraffe face, and a dragonfly face, I am confident that any of the implemented neural networks would find near 100% success rates rather quickly, as there are mountains of differences to find between those three classes. With emotions, though, the case is much more complex and requires extracting more minute data. When given the entire face, all of the little clues that point to a specific emotion may be too numerous for a single neural network to properly learn.

In further regard to the data set itself, it became quite clear after seeing the derived results, 3,200 images were not sufficient to achieve desired results (if doing so is possible with whole face neural network training). Even lowering the percentage of images used for the validation set from 25% to 10% improved the performance of multiple neural networks, leading to the conclusion that more images should improve performance. More specifically, after observing the results above, it seems that a superior data set would consist of only upright, grayscale faces. Having 800 different faces, all upright and expressing the desired emotion, would most likely allow the neural networks to learn to recognize the desired facial patterns better than the datasets used. It can be speculated that using this approach would lead to grossly incorrect classifications if a test image is rotated to a non-upright position, but this would need to be tested further.

With nearly all of the trained networks requiring around 115 epochs (sometimes more) to achieve desired success rates, inference times were understandably long – in the order of seconds. The resulting equations after 115 epochs would have contained exponentially greater numbers of weights and variables than the supplied data set's 5 epochs, which led to the greater inference times. In order to achieve more believable human-robot communication, the inference time for classifying the emotion of a face would need to be much faster (closer to the 5 milliseconds for the supplied data set), since a conversation could potentially change drastically in 5 seconds, and a communicating robot would need to be able to keep up. Unfortunately, accuracy would remain even more important than inference time, since misclassifying a given person's emotion would degrade the quality of a human-robot interaction more severely than a delayed reply. Nonetheless, the ultimate goal would remain to improve both factors. The inference time could be significantly longer than 5 milliseconds to achieve a fluid pace of conversation, but it would certainly need to be less than a second.

Lastly, it was rather interesting to observe that the appearance of a trained neural network (as illustrated by the graph of loss and accuracy rates) almost never coincided with the overall success rates of the neural networks. Certain neural networks would achieve incredibly low loss rates (in the thousandths) and have training data classification at nearly 100%, but would be unable to achieve even a 25% success rate on the test images. The opposite would be true for certain other networks, which could only achieve an accuracy of around 70% on the training data, but would then be able to score over 50% on the test data. Starting with different learning rates didn't seem to affect the final results either. Decreasing the starting learning rate by a factor of 10 resulted in a flatter initial accuracy curve, but the slower learning rate did not appear to help the neural networks derive the more intricate and complex facial features with greater success. All this further shows the high importance that the dataset itself plays in successful neural network classification.

Conclusion/ Future Work:

Overall, although quite a lot of work went into this particular project, the final results left much to be desired. Nonetheless, the results did provide an invaluable learning experience in regards to which changes and alterations would be needed to make a successful emotion-classification interface. A few different changes could be done to the format implemented in this experiment, such as changing the dataset to only upright faces, setting them all to grayscale, and totaling at least 1,000 images for each class. However, a better approach would most likely involve breaking down the face identification process into a more complex program, involving multiple neural networks.

As discussed above, neural networks, in their training process, are oblivious to the desired patterns of identification; instead, they need to be guided towards the correct training method. Therefore, it should be better to divide the process into discrete steps. To start, a detection network would need to be used to identify the human face(s) in a given field of view. Detection neural networks have been illustrated in the past to successfully identify human faces out of a given image, so this step should proceed without difficulty. Afterwards, the aforementioned detection box (or boxes) would undergo the subsequent emotion identification. First, just two neural networks could be trained, one to look for specific patterns in the top of the face (eyes, eyebrows, forehead) and one to look at the bottom of the face (mouth, cheeks, jaw). Each of these networks would be trained to classify how their respective part of the face would appear with specific emotions and then provide a certainty percentage. The program would then use combined percentages to attain a final emotion (or the predominant emotion features present on the face).

If that scheme succeeds, then more work can be done to further subdivide areas of the face and train neural networks to look at specific features of the face: eyes, mouth, eyebrows, forehead, etc. Although it is not concretely known how the human brain identifies emotion visually, it can be speculated that the visual cortex, functionally, looks for specific facial features and connects all the information together (although to our conscious minds this process would appear instantaneous and fully-connected). Using this approach would hopefully provide both faster results (since the neural networks would be training to recognize much simpler shapes and should attain desired accuracy in fewer epochs) and result in significantly fewer grossly incorrect classifications. The logic of the latter statement would be that if the emotion identification program has the results of 6 or more different neural networks to build from, even if one or two of the features are misclassified, it would not detract from the final classification; if anything, it could lead to later subsequent development of properly recognizing hints of other emotions presents amidst the dominantly expressed one.

As the field of robotics continues to improve, robots will become integrated into virtually every component of everyday life. Having computer systems capable of successfully interpreting data and making intelligent decisions will certainly provide tremendous benefits for humanity. However, through all this, it is imperative to remember that we, as people, are not aiming to transform the world around us into an impersonal, cold environment best suited for all those robots (if not careful, though, this may very well be what happens). Instead, the interrelation of technological advancements, most definitely including robotics, with the human world must be factored into. So, while specific advancements in robotics could have the intention of improving the comfort and productivity of people around them, unless said robots are able to do it in a fashion that allows people to feel comfortable and at-ease with the robots, general acceptance and excitement for robotics may not occur.

Currently, various computer system exist that can take in various basic voice commands and respond in a few pre-determined ways. There is even extensive work being done in building robotic systems which attempt to appear very human-like, especially regarding the face. However, it has, interestingly enough, been found that the human eye is extremely keen on recognizing even minute “in-humanities”, so to speak, meaning that visual imitation may not be the way to go. Instead, the key may be in achieving believable communication and interaction. Even if the responses are coming from a non-descript box, a positive reaction should be achieved if a person can feel truly understood and properly listened to. While multiple factors are involved in this, emotional understanding and response are absolutely instrumental in achieving believable human-robot communication. A robot can be trained to output a specific sentence given specific verbal inputs, but the person on the receiving end would very likely not feel the authenticity of the output, unless it is done so with consideration to emotional context. Neural networks can definitely play a part in achieving this level of emotional understanding, with the hope that in the not-so-distant future, people will be able to interact with robots not through blunt commands, but through normal, everyday language, as well as receive meaningful responses.