# Tweaking the AlexNet

Amit Sahu Ayushman Dash John Gamboa
Vitor Rey Thanitta Sutjaritvorakul

November 8, 2016

## Introduction

The AlexNet (Krizhevsky, Sutskever, and Hinton 2012) is a prominent CNN model that known to have produced the best results on the ImageNet Large Scale Visual Recognition Challenge in 2012, with an improvement of around 10% on the error rate in comparison to the second best model (see the LSVRC-2012 results for more information).

Figure 1 shows an overview of the model. The first convolutional layer with is composed by 96 filter maps of size $11 \times 11 \times 3$. The AlexNet uses a stride 4. This followed by a contrast normalization layer and a pooling layer of size $3 \times 3$, with a stride of $2 \times 2$.

The second convolutional layer is composed of 48 convolutional kernels of size $5 \times 5$, again followed by a contrast normalization layer and a pooling layer with the same parameters as the first one.

A third convolutional layer follows, composed by 384 filter maps of size $3 \times 3 \times 256$, followed by another one composed by 384 kernels of size $3 \times 3 \times 192$. Finally, one last convolutional layer consisting of 256 kernels of size $3 \times 3 \times 192$ completes the convolutional part of the network. In their implementation, these three layers were divided into two GPUs. For this reason, the fourth and the fifth layer had connections only to the part of the third layer that was present in the same GPU. This caused the learned filters to specialize (Krizhevsky, Sutskever, and Hinton 2012).

On top of the aforementioned structure, 2 fully connected layers with 4096 neurons each are followed by the final 1000-neurons output layer. The network is trained with backpropagation, and a dropout of 0.5 is used in the two last 4096 neurons fully connected layer. All layers use ReLU units.

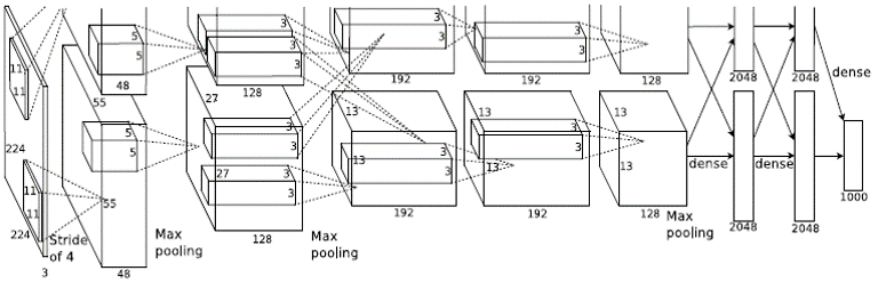


Figure 1: AlexNet original architecture.

# Tweaks

As a baseline, a model similar to the original AlexNet was trained. To facilitate visualiations, we use 100 filter maps in the first and second layers instead of the original 96 filter maps of the first layer and 48 of the second layer. Additionally, in the fourth layer, we use 384 kernels of size $3 \times 3 \times 384$, and in the fifth layer we use 256 kernels of size $3 \times 3 \times 384$. Finally, no separation into different GPUs was done. In what follows, we call this model the "original-AlexNet" (despite its misleading name).

We wanted to compare speed of convergence of the original AlexNet with that of an AlexNet using *tanh* units. In what follows, we call this model the "*tanh*-AlexNet".

Additionally, we trained a much smaller model to investigate how better the results of AlexNet would be compared to this cheapily trainable model. In what follows, we call it "small-AlexNet". This was composed of only 64 filters $11 \times 11 \times 3$ in the first layer, and 192 filters $5 \times 5$ in the second layer. The number of filters in the third, fourth and fifth layers were also reduced: 384, 256 and 256 respectively. Finally, the fully connected layers have 384 and 192 nodes, respectively.

Figure 2 shows the evolution of the loss of the original AlexNet.

Finally, we tried running other models (e.g., totally removing the second convolutional layer; removing the third, fourth and fifth layer; or removing one of the fully connected layers). The results produced by these models were not good, and are not reported here.

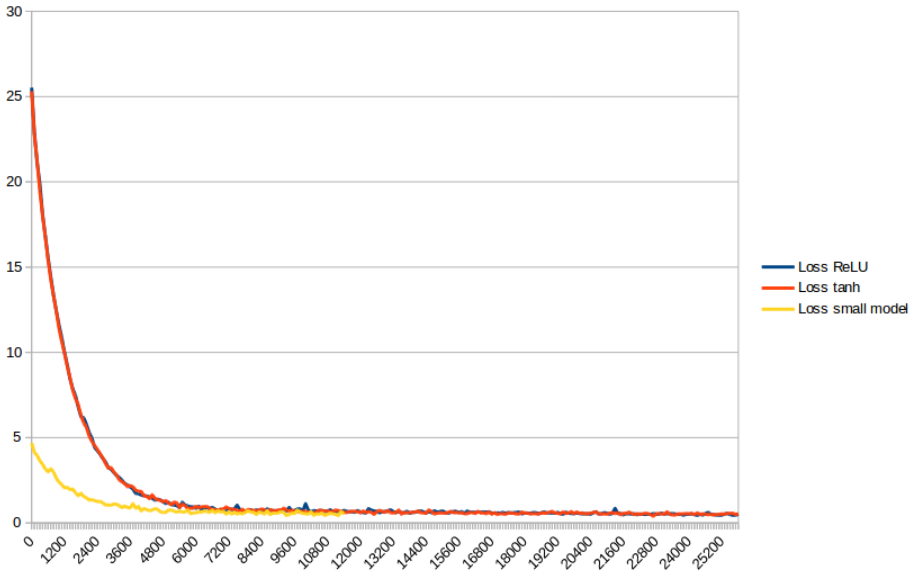


Figure 2: Losses were recorded after each hundred iterations.

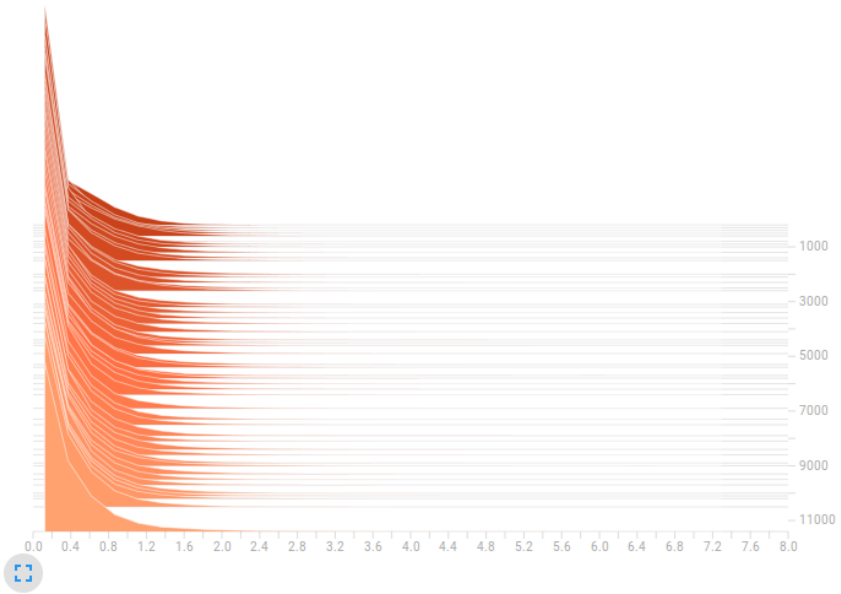


Figure 3: Histogram of small-AlexNet second fully connected layer.

## Discussion

Notice that the small-AlexNet was trained only until the iteration number 11400. Still, as can be seen in Figure 2, all models converged to a similar loss.

When running the trained models in a small test set of 10000 images, we found the results below. For time constraints, we ran the model for an undetermined number of iterations without having used any validation method.

| Model | Loss | Iterations |
|---|---|---|
| original-AlexNet | 0.852 | 27000 |
| tanh-AlexNet | 0.891 | 25800 |
| small-AlexNet | 0.883 | 11400 |

From the table above, it seems that the original-AlexNet has overfitted a little: we had expected it to have produced the best results. Still, it is clear that neither the tanh-AlexNet nor the small-AlexNet are too bad in comparison, having produced similar (in this case, better) results than the original version.

There seems to be a significant amount of ReLU units (for the two networks composed by ReLU units) that are dying during the training process. This can be seen in Figure 3, showing a histogram of the activations of all units through the training epochs of the second fully connected layer of the small-AlexNet. Similar activations were also found in the original-AlexNet, and througout all layers composing the model.

## References

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton. 2012. "Imagenet Classification with Deep Convolutional Neural Networks." In *Advances in Neural Information Processing Systems*, 1097–1105.