LLNL-CONF-661841

LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

Large-Scale Deep Learning on the YFCC100M Dataset

K. Ni, K. Boakye, B. Van Essen, R. Pearce, D. Borth, B. Chen, E. Wang

October 3, 2014

Neural Information Processing Systems 2014
Quebec, Canada
December 8, 2014 through December 13, 2014

Disclaimer

Large-Scale Deep Learning on the YFCC100M Dataset

Anonymous Author(s)
Affiliation
Address
email

Abstract

We present a work-in-progress snapshot of learning with a 15 billion parameter deep learning network on HPC architectures applied to the largest publicly available natural image and video dataset released to-date. Recent advancements in unsupervised deep neural networks suggest that scaling up such networks in both model and training dataset size can yield significant improvements in the learning of concepts at the highest layers. We train our three-layer deep neural network on the Yahoo! Flickr Creative Commons 100M dataset. The dataset comprises approximately 99.2 million images and $800,000$ user-created videos from Yahoo's Flickr image and video sharing platform. Training of our network takes eight days on 98 GPU nodes at the High Performance Computing Center at Lawrence Livermore National Laboratory. Encouraging preliminary results and future research directions are presented and discussed.

1 Introduction

The field of deep learning via stacked neural networks has received renewed interest in the last decade [1, 2, 3]. Neural networks have been shown to perform well in a wide variety of tasks, including text analysis [4], speech recognition [5, 6, 7], various classification tasks [8, 9], and most notably unsupervised and supervised feature learning on natural imagery [1, 2, 3].

Deep neural networks applied to natural images have demonstrated state-of-the-art performance in supervised object recognition tasks [10, 1] as well as unsupervised neural networks [2, 3]. The classical approach to training neural networks for computer vision is via a large dataset of labeled data. However, sufficiently large and accurately labeled data is difficult and expensive to acquire. Motivated by this, [3] explored the application of deep neural networks in unsupervised deep learning and discovered that sufficiently large deep networks are capable of learning highly complex concept level features at the top level without labels.

Spurred by this advancement, [2] set out to construct very large networks on the order of 10^9 to 10^{10} parameters. A key advancement was the highly efficient multi-GPU architecture of their model. [2] employed both model and data parallelism and was able to process 10 million YouTube thumbnails in a few days processing time on a medium sized cluster. A notable result was the unsupervised learning of various faces, including those of humans and cats. Ultimately, improved feature learning at larger scales can improve downstream capabilities such as scene or object classification, additional unsupervised learning (*i.e.* via topic modeling [11] or natural language processing algorithms [12]).

In collaboration with the authors of [2], we have scaled a similar model and architecture to over 15 billion parameters on the Lawrence Livermore National Laboratory's (LLNL) Edge High Performance Computing (HPC) system. Our long-term goal is two-fold: (1) explore at-the-limit performance of massive networks ($>$ 10 billion parameters) and (2) train on and analyze datasets on the order of 100 million images.

As the number of network parameters grow, datasets need to be scaled accordingly to avoid overfitting the models. We take advantage of a brand-new dataset released jointly by Yahoo!, LLNL and the International Computer Science Institute (ICSI) called the Yahoo! Flickr Creative Commons 100M (YFCC100M) dataset10. The dataset is, to the authors' knowledge, the largest single publicly available image and video dataset ever published. In addition to the raw images and video, the YFCC100M also contains metadata for each entry including locations, camera types, keywords, titles, etc. Although beyond the scope of this paper, this rich associated meta-data potentially offers researchers additional avenues of semantic multi-modality learning to explore.

Working with datasets, models and computing architectures at the scales considered in this paper presents several daunting engineering challenges. For example, the significantly greater number of GPUs and compute nodes used in our system versus [2] creates communication issues in MPI. In addition, a typical model takes up over 40 GB of memory, making simple offline analysis tasks such as visualization challenging. Various network architectures were tested, balancing performance and computational constraints, before we arrived at our current model. Finally, as in [2], data throughput presents a bottleneck to model training. We present a novel pipeline approach to address this problem.

The rest of this paper is organized as follows. In Section 2 we give a brief overview of the YFCC100M dataset. The network architecture and computational framework being employed is described in Section 3. We present preliminary results and visualizations of our network in Section 4. Finally, we summarize and discuss future research directions in Section 5.

2 Overview of the YFCC100M Dataset

In late June 2014, Yahoo! released the Yahoo! Flickr Creative Commons dataset (YFCC100M). This dataset consists of 100 million Flickr user-uploaded images and videos (99,206,564 images and 793,436 videos) along with their corresponding metadata including title, description, camera type, tags, and geotags when available. All of the data is under Creative Commons licensing and is freely provided to scientists for the advancement of multimedia research [1]. In addition to the raw images, videos, and metadata, Yahoo! in collaboration with the ICSI and LLNL will be computing and providing standard computer vision and audio features using LLNL's supercomputing resources.

Wang et al. [13] have used YFCC100M data to build systems that associate images with more natural annotations like those found in user-generated captions. Others are interested in using the YFCC100M imagery and audio to geolocate where the photo or video was taken [14]. In fact, the 2014 MediaEval Placing Task is using YFCC100M as the source of benchmark data [15]. We are interested in using YFCC100M as our sandbox dataset for learning image features using massive unsupervised neural networks, repeating the experiment by [3] on an order of magnitude more data and neural network parameters. In particular, we want to see what other "grandmother neurons" [3] our network would automatically learn from YFCC100M.

The 99,206,564 images were created and posted by 578,268 different Flickr users. 76%, 20%, and 4% of the images have titles, auto-titles, or no titles, respectively. The average number of words per title is 3.08. 32% of the images have descriptions with an average of 22.52 words per description. Finally, 69% of the images have on average 7.07 tags per image. The top 60 tags are shown in Table 1. In Fig. 1 we show example images and associated meta-data for several YFCC100M images.

[1]Available at http://research.yahoo.com/Academic_Relations

Table 1: Top 60 Tags in YFCC100M Images

square	iphoneography	square format	instagram app	california	travel
nikon	usa	canon	london	japan	france
nature	art	music	europe	beach	united states
england	wedding	italy	new york	canada	city
vacation	germany	party	park	water	people
uk	spain	architecture	summer	festival	nyc
taiwan	paris	san francisco	australia	winter	sky
snow	concert	night	family	china	museum
food	street	live	washington	landscape	flower
sunset	photo	flowers	holiday	trip	photography

6700666461
002522a25a75c238be4a76cb7b
870 7327243@N05 Dougtone
2012-01-14 20:04:19.0
1326631741
Canon+PowerShot+A1200
Adirondack+Phantoms+vs.+Ma
nchester+Monarchs+-
+January+14%2C+2012
Adirondack+Phantoms+vs.+Ma
nchester+Monarchs+-
+Glens+Falls+Civic+Center+-
+Glens+Falls%2C+New+York+-
+January+14%2C+2
012
011412,adirondack,adirondack+
phantoms,ahl,american+hockey
+league,banner,dax,defencema
n,defenseman,faceoff,farm+tea
m,forward,glens+falls,glens+fall
s+civic+center,goalie,goaltender
,hockey,ice+hockey,los+angeles
+kings,manchester,manchester
+monarchs,mascot,minor+leagu
e,monarchs,netminder,new+yor
k,penalty,phantoms,philadelphi
a+flyers,prospect,puck,referee,r
ink,score,scoreboard,upstate,za
mboni -73.642852 43.308483 15
http://www.flickr.com/photos/
7327243@N05/6700666461/
http://farm8.staticflickr.com/71
52/6700666461_2a785d3baf.jp
g Attribution-ShareAlike License
http://creativecommons.org/lic
enses/by-sa/2.0/ 7152 8
2a785d3baf 0ac9dba756 jpg 0

6410782589
002a3e6e299e7eb5fc3a94732d
e341 14212025@N00 ascaro41
2011-11-27 10:20:57.0
1322393934 SONY+DSLR-A700
portofino+in+liberty
http://www.flickr.com/photos/
14212025@N00/6410782589/
http://farm7.staticflickr.com/60
60/6410782589_4a5f534b9e.jp
g Attribution-NonCommercial-
ShareAlike License htt
p://creativecommons.org/licens
es/by-nc-sa/2.0/ 6060 7
4a5f534b9e cc90f79424 jpg 0

183829695
00253036c4ff8b97c254dbaef74
5b 13416711@N00 mlaaker
2006-07-06 20:59:40.0
1152244780
OLYMPUS+OPTICAL+CO.%2CLTD
+C830L%2CD340R P1010194
billy,emerald,europe,holiday,ire
land,isle,laaker,maday,micah,va
cation -7.717895 53.376774 6
http://www.flickr.com/photos/
13416711@N00/183829695/
http://f
arm1.staticflickr.com/48/18382
9695_831f78b819.jpg
Attribution-NonCommercial-
ShareAlike License
http://creativecommons.org/lic
enses/by-nc-sa/2.0/ 48 1
831f78b819 831f78b819 jpg 0

6382896907
002a13619cbf78e627887258bc
389 15462727@N07
Ingy+The+Wingy 1988-06-13
00:00:00.0 1321968247
Mill+Lane+Junction
Part+of+the+illuminated+track+
diagram+in+Mill+Lane+Junction
+signal+box.+Monday+13th+Ju
ne+1988%0A%0AMill+Lane+Jun
ction+signal+box+is+located+by
+the+Siding+11+line+alongside
+Mil
l+Lane+underbridge+in+Bradfor
d%2C+and+is+a+Railway+Signal
+Company+Limited+standard+d
esign+built+for+the+Lancashire
+and+Yorkshire+Railway+and+o
pened+on+7th+November+188
4+fitted+with+a+56+lever+Rail
way+Signal+Company+Limited+
frame%2C+replacing+an+earlier
+signal+box.+A+56+lever+frame
+%28the+origina
l+or+possibly+a+replacement+L
ancashire+and+Yorkshire+Railw
ay+Tappet%29+was+replaced+o
n+3rd+June+1973+by+a+British
+Railways+Eastern+Region+indi
vidual+function+switch+console
+in+connection+with+the+closu
re+of+Bradford+Exchange+stati
on+and+the+opening+of+the+r
eplacement+Bradford+Intercha
nge+station
%0A%0ARef+no+08625 -
1.748843 53.786191 16
http://www.flickr.com/photos/
15462727@N07/6382896907/
http://farm7.staticflickr.com/61
15/6382896907_6d0c84c8fe.jpg
Attribution-NoDerivs License
http://creativecommons.org/lic
enses/by-nd/2.0/ 6115 7
6d0c84c8fe 65bfeab023 jpg 0

7279325618
002a69f9f81df2335426cf1c163b
5 54383933@N03
monoprixgourmet_bis 2012-05-
25 18:53:51.0 1338126141
RICOH+GXR+S10
%E3%82%AB%E3%83%88%E3%
83%AA%E3%83%83%E3%82%AF
%E5%BC%98%E5%89%8D%E6%
95%99%E4%BC%9A%E6%95%99
%E4%BC%9A%E5%A0%82+%E5
%A4%A9%E5%9C%B0%E5%89%
B5%E9%80%A0%E3%81%8B%E3
%82%89%E3%82%AD%E3
%83%AA%E3%82%B9%E3%83%
88%E3%81%8C%E7%94%9F%E3
%81%BE%E3%82%8C%E3%80%
81%E5%B2%A9%E6%9C%A8%E
5%B1%B1%E3%81%AE%E5%9C
%B0%E3%81%AB%E6%AE%BF%
E5%A0%82%E3%81%95%E3%82
%8C%E3%82%8B%E3%81%BE%
E3%81%A7%E3%81%AE%E7%89
%A9%E8%AA%9E
aomori+prefecture,hirosaki,japa
n 140.471923 40.606281 16
http://www.flickr.com/pho
tos/54383933@N03/727932561
8/
http://farm8.staticflickr.com/70
95/7279325618_4ae0ae6102.jp
g Attribution License
http://creativecommons.org/lic
enses/by/2.0/ 7095 8
4ae0ae6102 f472e09676 jpg 0

Figure 1: Examples of YFCC Data, and the associated metadata. Photo credits to Yahoo! users “Dougtone”, “ascaro41”, “mlaaker”, “Ingy The Wingy”, “monoprixgourmet_bis”.

3 Analysis with Large Scale Neural Networks

3.1 Network Architecture

For the large set of image data, we employed a three-layer, large-scale deep neural network with a reconstruction independent component analysis (RICA) cost function,

$$\begin{aligned} \min_{W,\alpha,b} \sum_i \left\| W^T(\alpha W x^{(i)}) + b - x^{(i)} \right\|_2^2 + \lambda\sqrt{(\alpha W x^{(i)})^2} \\ \text{subject to } \|W^{(k)}\|_2 = 1, \forall k \end{aligned} \tag{1}$$

As in [2], W is a weighting matrix, α is a scaling value and $x^{(i)}$ are the data points at the beginning of each layer. In addition, we introduce an offset, b, for increased model flexibility. The parameter λ controls the relative sparsity, and is set to 0.1 at the first two layers and 0.01 at the final layer. Unlike [2], we do not presently include a pooling layer, as we believe the scale of the network and training data allows a similar translational invariance to be automatically learned. A particular advantage conferred by the RICA construction in (1) is that the sparseness term $\lambda\sqrt{(\alpha W x^{(i)})^2}$ can be computed in-situ with the rest of the model parameters. This is in contrast to the conventional sparse autoencoder construction that requires a second pass through the data to compute a sparseness-specific gradient contribution.

Fig. 2 illustrates the structure of our network. The three layers are composed of two untied convolutional layers, and a third fully-connected layer. The first convolutional layer utilizes 5184 filters [2] of input size $16 \times 16 \times 3$ with stride 4 and output size $4 \times 4 \times 24$. The second layer takes 16 spatially contiguous [3] $4 \times 4 \times 24$ outputs of the first layer and connects them fully to a $4 \times 4 \times 24$ output. The stride length of the second layer is 4. The third layer is dense, and fully connects the $62 \times 62 \times 24$ outputs of the second layer to 4096 top-level neurons. The total number of parameters trained is 15 billion. After each layer, local contrast normalization (LCN) is applied prior to continuing onto the next layer. Though no pooling is applied, the window sizes at the next layer are large enough to incorporate spatial information from neighboring blocks.

Training data is arranged into 99,207 data blocks of 960 images. Each data block consists of 5 mini-batches, where each mini-batch contains 192 images. Due to the scale of the data, the proposed algorithm reduces training time by employing a pipeline technique where the next layer begins training before the previous layer has finished. Analogous to the example shown in Fig. 3, after a layer L has trained an initial set of data blocks (in our case, 1000), the next layer, $L+1$, starts training. To accomplish this, two instances of the layer L are run simultaneously: one which continues training and one that uses up-to-date parameters to forward propagate data from Block 0 to the layer $L+1$. The parameters of the forward-propagating layer L instance are periodically synchronized with the layer L instance that continued training. We observed that our model was not sensitive to the choice of synchronization frequency. As a rule of thumb, we wait to train layer $L+1$ until the objective of layer L stabilizes, which typically occurs after approximately one million images.

3.2 HPC Architecture

To train the neural network at scale, we used 98 nodes of the Edge HPC cluster at Lawrence Livermore National Laboratory. The Edge cluster consists of 206 nodes with 12 core Intel Xeon EP X5660 running at 2.8 GHz. Each node has 96 GB of DRAM and a Tesla M2050 (Fermi) NVIDIA GPU with 3 GB of GDDR5. The training algorithm is model parallel as described in [2], with the nodes and GPUs processing each mini-batch across the system and

[2] Arranged in a 72×72 grid
[3] Arranged in a 4×4 grid

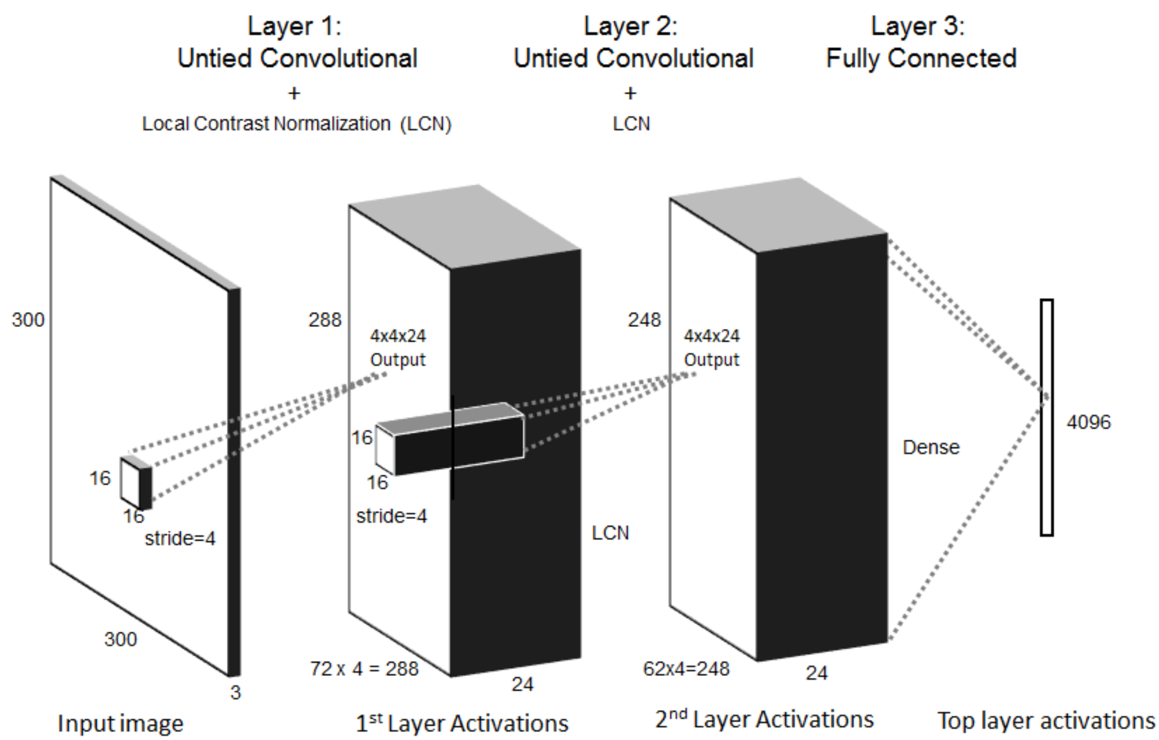


Figure 2: Network topology of large scale, trained network. Approximately 15 billion parameters

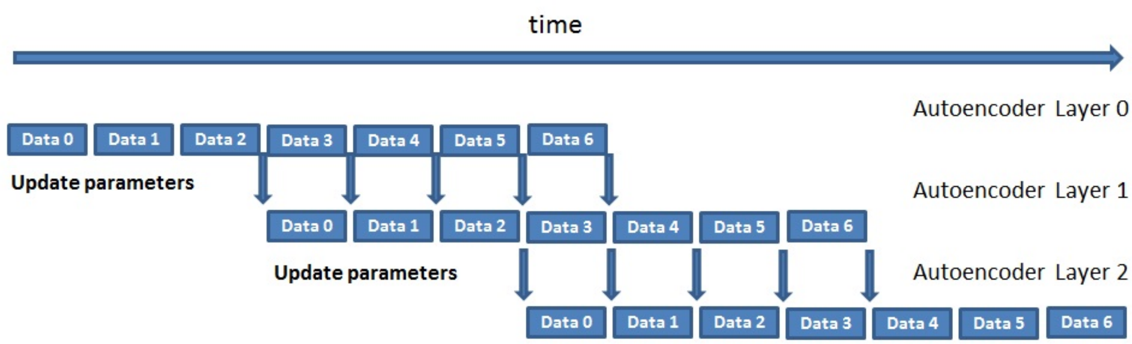


Figure 3: Pipeline for semi-parallel training of sparse autoencoders from a single data source

distributing the model across the GPUs. Communication was provided by MPI over Mellanox QDR Infiniband cards. The GPU accelerators were used with CUDA 5.5 and MPI-direct communication and the operating system was a 2.6.32 kernel RHEL 6 derivative.

The dataset was stored in a Lustre file system with a peak bandwidth of 10 GB/s. Each mini-batch was copied from Lustre into memory and then streamed into the GPU's memory. Each GPU is responsible for computing its section of the model parameters for the current mini-batch. Communication within the algorithm occurs when a layer's input (or output) field spans multiple GPUs. The communication is handled by a distributed array data structure (using MPI) within the training algorithm. Global communication is minimized by using untied local receptive fields, and allowing receptive fields to be trained independently.

4 Preliminary Results

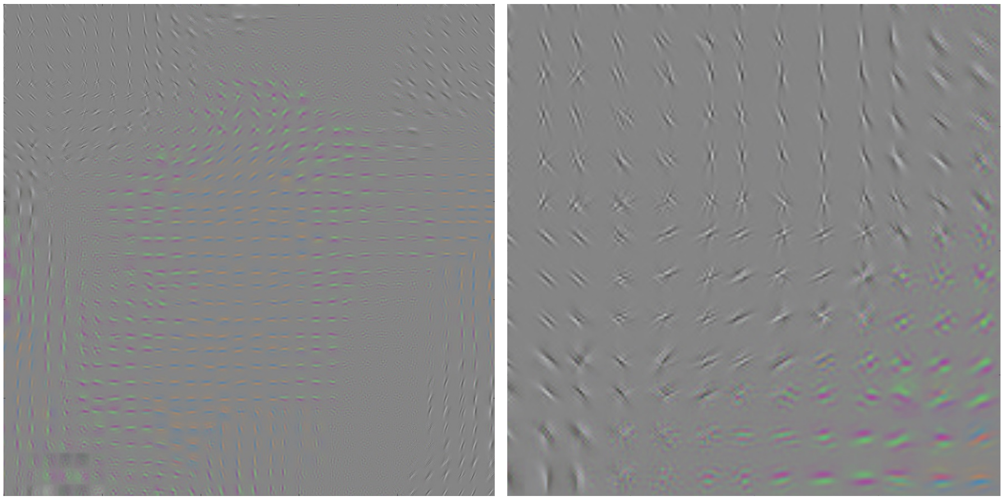
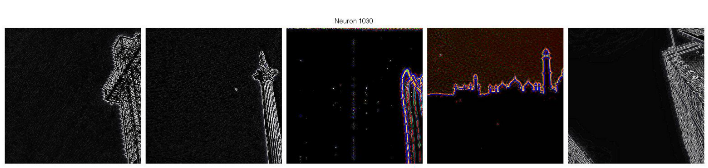


Figure 4: Visualization of a selection of typical first layer weights. The right figure is a zoomed-in crop of the left.

We trained the network using all images from the YFCC100M dataset. Images were preprocessed as in [2], and subsequently resized to 300 x 300 pixels by first centering, then scaling the smallest dimension to 300 pixels, and finally cropping. After training all three layers, we forward propagated 2 million images through the network in order to obtain activation values for visualization. Note that in this paper, the test set is significantly noisier than the benchmark Labeled Faces In the Wild [16] and ImageNet [17] datasets considered in previous works such as [3].

In Fig. 4 we visualize some typical first layer weights. As expected, they are trained to capture various types of edges and separate into color and texture focused neurons. In Fig. 5, we visualize some example neurons by showing the top 5 stimuli for each neuron. We observe that our network is capable of learning significant structure, identifying buildings, aircraft, text, cityscapes, and tower-like buildings, among many others. The network seems to cue in on distinctive textures such as the edges of text, sides of buildings and the sharp edge of airplanes against the smooth gradation of the sky. Moreover, the network seems to activate on large-scale structures within an image rather than local features.

Our results, while encouraging, suggest that significant improvements can be achieved through improved network architecture and increased depth. As was demonstrated in [1], network architecture has a significant impact on the performance of deep networks. We believe that a significant contributor to our networks' performance is due to its large size being able to capture complex concepts. While the networks described in [3] were able to learn complex features in just three layers, our results suggest that extremely large datasets such as the YFCC100M can support deeper networks with improved high-level concept learning.



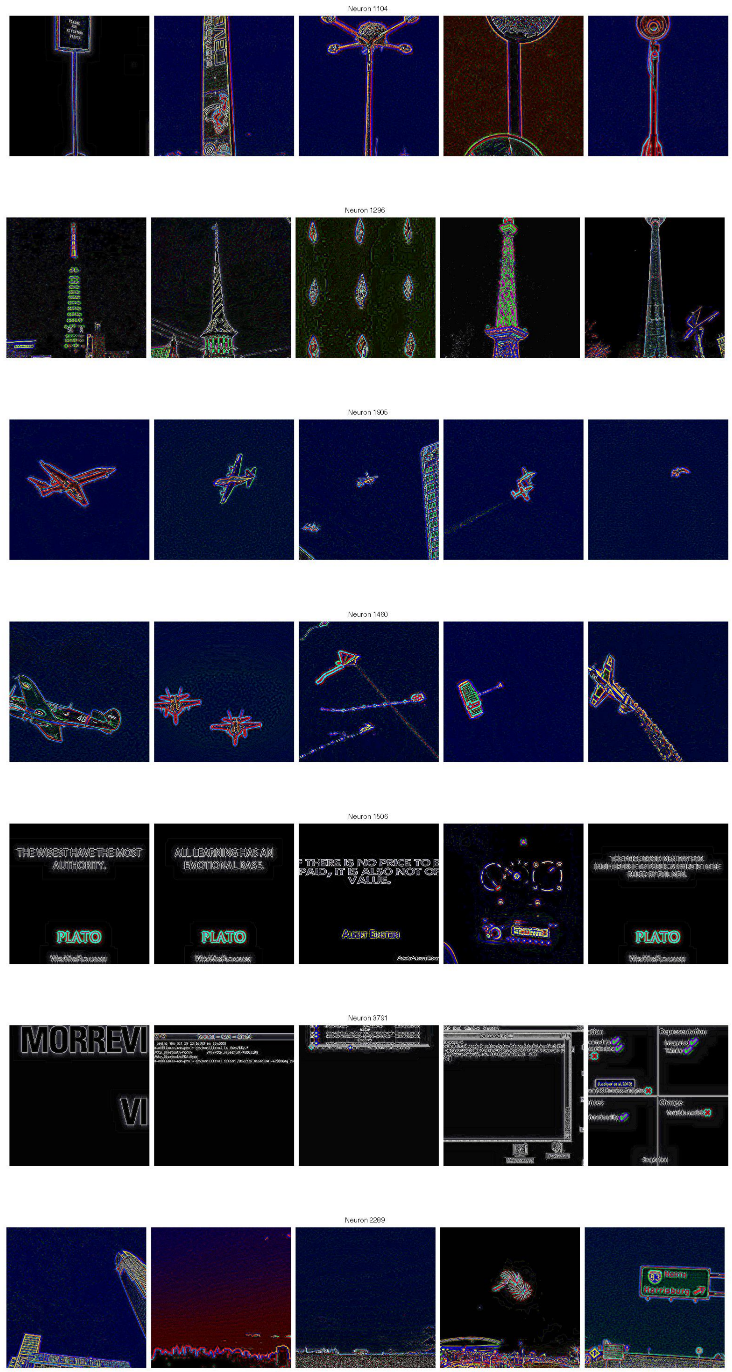


Figure 5: Top-5 stimuli of example layer 3 neurons. Images have been whitened.

5 Summary and Future Work

The results discussed in this paper present a snapshot of the work in progress at Lawrence Livermore National Laboratory in scaling up deep neural networks. Such networks offer enormous potential to researchers in both supervised and unsupervised computer vision tasks, from object recognition and classification to unsupervised feature extraction.

To date, we see highly encouraging results from training our large 15 billion parameter three-layer neural network on the YFCC100M dataset in an unsupervised manner. The results suggest that the network is capable of learning highly complex concepts such as cityscapes, aircraft, buildings, and text, all without labels or other guidance. That this structure is visible upon examination is made all the more remarkable due to the noisiness of our test set (taken at random from the YFCC100M dataset itself).

Future work on our networks will focus on two main thrusts: (1) improve the high-level concept learning by increasing the depth of our network, and (2) scaling our network's width in the middle layers. On the first thrust, we aim for improved high-level summarization and scene understanding. Challenges on this front include careful tuning of parameters to combat the "vanishing gradient" problem and design of the connectivity structure of the higher-level layers to maximize learning. On the second thrust, our challenges are primarily engineering focused. Memory and message passing constraints become a serious concern, even on the large HPC systems fielded by LLNL. As we move beyond our current large neural network, we plan to explore the use of memory hierarchies for staging intermediate/input data to minimize the amount of node-to-node communication, enabling the training and analysis of even larger networks.

6 Acknowledgments

We would like to thank Adam Coates, Brody Huval and Andrew Ng for providing their COTS HPC Deep Learning software and helpful advice. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

References

[1] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 2012.

[2] A. Coates, B. Huval, T. Wang, D. J. Wu, A. Y. Ng, and B. Catanzaro. Deep learning with cots hpc. In *International Conference on Machine Learning*, 2013.

[3] Q. V. Le, M. Ranzato, R. Monga, M. Devin, K. Chen, G. S. Corrado, J. Dean, and A. Y. Ng. Building high-level features using large scale unsupervised learning. In *International Conference on Machine Learning*, 2012.

[4] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*, 2013.

[5] O. Abdel-Hamid, L. Deng, and D. Yu. Exploring convolutional neural network structures and optimization techniques for speech recognition. In *Interspeech 2013*, 2013.

[6] G. Hinton, L. Deng, D. Yu, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, G. Dahl, and B. Kingsbury. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, 29(6):82–97, November 2012.

[7] H. Bourlard and N. Morgan. *Connectionist Speech Recognition: A Hybrid Approach*. Kluwer Academic Publishers, 1993.

[8] D. Claudiu Ciresan, U. Meier, L. M. Gambardella, and J. Schmidhuber. Convolutional neural network committees for handwritten character classification. In *International Conference on Document Analysis and Recognition*, 2011.

[9] D. Reby, S. Lek, I. Dimopoulos, J. Joachim, J. Lauga, and S. Aulagnier. Artificial neural networks as a classification method in the behavioural sciences. *Behavioural Processes*, 40:35–43, 1997.

[10] R Uetz and S. Behnke. Large-scale object recognition with cuda-accelerated hierarchical neural networks. In *IEEE International Conference on Intelligent Computing and Intelligent Systems*, 2009.

[11] L. Cao and L. Fei-Fei. Spatially coherent latent topic model for concurrent object segmentation and classification. In *Proceedings of International Conference on Computer vision*, 2007.

[12] R. Socher and M. Ganjoo and C. D. Manning and A. Y. Ng. Zero Shot Learning Through Cross-Modal Transfer. In *Advances in Neural Information Processing Systems 26*. 2013.

[13] J. K. Wang, F. Yan, A. Aker, and R. Gaizauskas. A poodle or a dog? Evaluating automatic image annotation using human descriptions at different levels of granularity. In *Proceedings of the Workshop on Vision and Language*, 2014.

[14] James Hays and Alexei A. Efros. im2gps: estimating geographic information from a single image. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[15] J. Choi, B. Thomee, G. Friedland, L. Cao, K. Ni, D. Borth, B. Elizalde, L. Gottlieb, C. Carrano, R. Pearce, D. Poland. The Placing Task: A Large Scale Geo-Estimation Challenge for Social-Media Videos and Images. *3rd ACM Multimedia Workshop On GeoTagging and Its Applications in Multimedia*. URL `http://www.multimediaeval.org/mediaeval2014/placing2014/`.

[16] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments.

[17] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-fei. Imagenet: A large-scale hierarchical image database. In *In CVPR*, 2009.