Chapter 0

Prologue: Why the mathematics of deep learning?

To answer the question of why have a course on the *mathematics* of deep learning, it is first instructive to consider why have a course on deep learning at all. To many of you, the answer to this question may seem clear or the question itself rhetorical. Indeed, already at MSU there are several courses on deep learning or that address deep learning as part of their course content, including courses in CSE, ECE, and MTH. Nevertheless, let us consider this question for a moment.

Deep learning refers to a class of algorithms in machine learning and more generally artificial intelligence. One of the hallmarks of deep learning algorithms is that they compose a sequence of many simple functions, often alternating between linear or affine functions, point-wise nonlinear functions, and pooling operations. Figure 1 gives an illustration of the VGG16 network [1], which is a powerful and very popular convolutional neural network, that consists of 16 layers of linear/non-linear pairs of operations, as well as pooling operations (where the length and width of the image stacks shrink) every few layers. Note that all of the linear functions are learned from the given training data and the associated task, which for VGG16 was image classification on the the ImageNet data base [2] (more on this later). Thus the the input to the VGG16 network is an RGB image, and the output is a class label. The compositional structure illustrated in the VGG network, and used in all of deep learning (this is where the "deep" comes from), has been incredibly successful in machine learning and artificial intelligence tasks over the last decade.

Indeed, deep learning is now used in a multitude of different contexts, from computer vision to natural language processing to playing games to biology to physics and more. One of the most striking examples of the success of deep learning (and in this case, reinforcement learning), is the success of AlphaGo

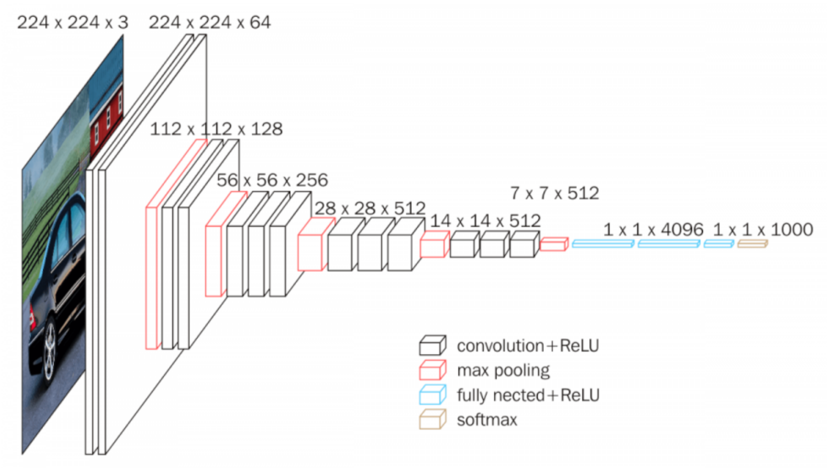
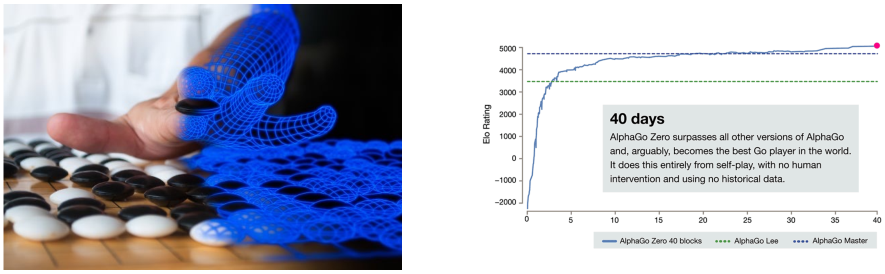


Figure 1: *The VGG16 network [1].*

Zero, which is an AI developed by Google DeepMind that has unparalleled ability at playing the game of Go (see Figure 2). Going back earlier in the decade, another key marker in the recent explosion in popularity of deep learning is the success of convolutional neural networks for image classification on the aforementioned ImageNet data base. In 2012, AlexNet [3], an eight layer convolutional neural network depicted in Figure 3, outperformed the next best classifier on the ImageNet data base by an astounding 10% on the top five prediction error rate. The existence of ImageNet [2] (see also Figure 4), and other databases such as the handwritten digit data base MNIST, which yield a "common task framework" have also been a key driver in the development of machine learning generally, and deep learning specifically.

But just how popular is deep learning? Indeed, the increasing popularity of deep learning has been rapid and breathtaking; see Figure 5 for the number of registrations and paper submissions to NeurIPS, the most popular machine learning conference. With this rapid increase in popularity, deep learning is being incorporated in a number of contexts with direct societal impact, such as self driving cars, medical diagnostics, insurance, and more.

However, deep learning is not without criticism. Indeed, despite its empirical successes, relatively speaking very little is known, precisely, on how and why it works so well. Indeed, the common task framework that has resulted in so much advancement has also resulted in the phenomena that many deep learning papers are the product of significant amounts of trial and error and less so on theoretically grounded process. Furthermore, even successful algorithms are hard to interpret. Thus, while advancement has been rapid in the last 10 years, one could argue that new, significant advancement will only



40 days

AlphaGo Zero surpasses all other versions of AlphaGo and, arguably, becomes the best Go player in the world. It does this entirely from self-play, with no human intervention and using no historical data.

(a) *Image from an article in Scientific American by Larry Greenemeier on AlphaGo Zero [4]; image credit Saran Poroong.*

(b) *Training time versus Elo rating for AlphaGo Zero.*

Figure 2: *After 40 days of training only against itself, AlphaGo Zero became arguably the best Go player in the world, and could beat its predecessors (older version of AlphaGo) nearly without fail.*

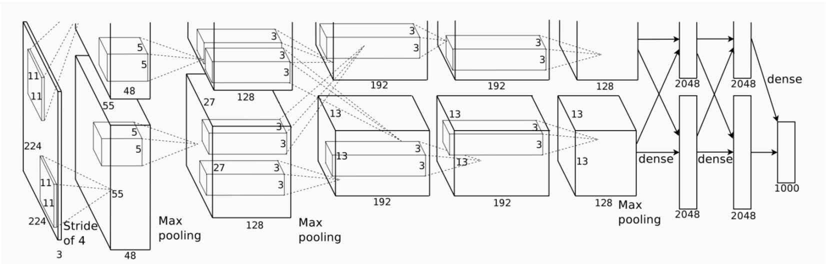


Figure 3: *The architecture of AlexNet [3].*

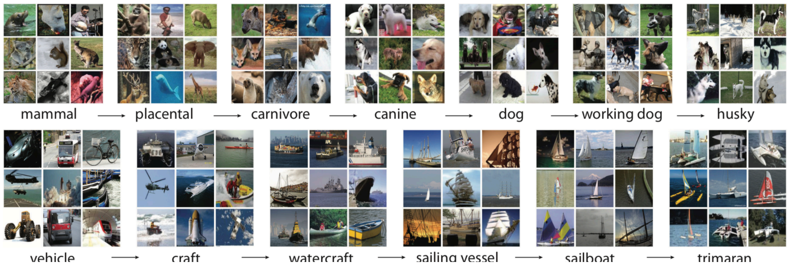
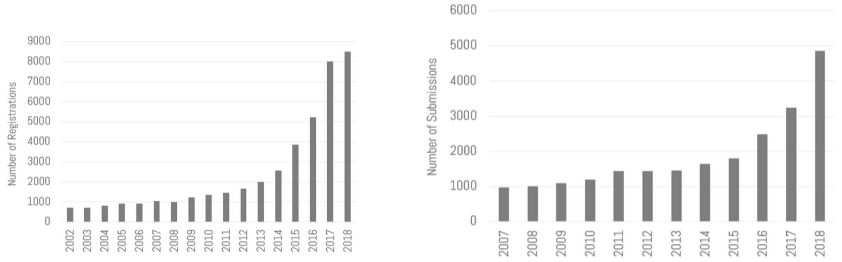


Figure 4: *A snapshot of a few images from the ImageNet data base [2], organized hierarchically by progressively more specific classes.*



(a) *NeurIPS conference registrations.* **(b)** *NeurIPS conference paper submissions.*

Figure 5: *NeurIPS conference registrations and paper submissions by year.*

come with increased understanding of the current state of the art. Furthermore, as deep learning finds it way more and more into the public realm, it will no longer be enough for machine learning algorithms and artificial intelligences to make a plausibly correct predictions or assessments, people will also increasingly want and need to know the reasons for such outcomes. Such considerations are also related to issues of fairness and bias in machine learning, which is only beginning to be understood but which will certainly increase in importance and study in the coming years.

The beginning point for further and more precise understanding, as with many scientific disciplines, is mathematics. The mathematical study of deep learning has progressed along many different paths, but many of these can be grouped into two primary avenues. On the one hand, there is the issue of training the networks. Because of their compositional and highly nonlinear structure, solving for the optimal weights of deep learning architectures results in a highly non-convex, high dimensional optimization problem. This is one avenue of mathematical study. On the other hand, one must first design the network before one can solve for the weights. The design of deep networks, and their resulting mathematical properties, is the second avenue of study. This course will focus primarily on the latter, leaving the mathematical details of optimization to another course. Within this context, we will consider supervised and unsupervised machine learning.

Bibliography

[1] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations*, 2015.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition*, 2009.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105. 2012.

[4] Larry Greenemeier. AI versus AI: Self-taught AlphaGo Zero vanquishes its predecessor. *Scientific American*, October 18, 2017.

[5] Pankaj Mehta, Marin Bukov, Ching-Hao Wang, Alexandre G.R. Day, Clint Richardson, Charles K. Fisher, and David J. Schwab. A high-bias, low-variance introduction to Machine Learning for physicists. arXiv:1803.08823, 2018.

[6] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer-Verlag New York, 2nd edition, 2009.

[7] J. Hartlap, P. Simon, and P. Schneider. Why your model parameter confidences might be too optimistic - unbiased estimation of the inverse covariance matrix. *Astronomy and Astrophysics*, 464(1):399–404, 2007.

[8] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Adaptive Computation and Machine Learning. The MIT Press, 2002.

[9] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[10] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference for Learning Representations*, San Diego, CA, USA, 2015.