Topic 2: Deep Learning: The Best Thing Since Sliced Bread or Just Another Bottle of Snake Oil?

Assignment 1 Name: Harshita Jhavar
Topics in Algorithmic Data Analysis SS'17 Matriculation Number: 2566267
Submitted to: Dr. Jilles Vreeken eMail: s8hajhav@stud.uni-saarland.de
Due by: May 04, 2017 before 10.00 hours

The Modern Term: Deep Learning

Artificial Intelligence is a thriving field of research where we look forward for building intelligent, robust, practical applications to automate the routine labour at work or at home. Based on efficiency of humans and computers in problem solving, problems can be categorised into two types. One, which can be described in mathematical formulas that might be intellectually difficult for human beings but easy for computers. For ex: Calculating cube root of a large number in very less time. And two, the problems which are easy to solve intuitively by human beings but due its complexity, are difficult to describe in form of mathematical expressions to a computer. For ex: Speech Recognition, Object Identification etc. Deep Learning is all about solutions to these intuitive problems where we allow the computers to learn from experiences and to comprehend the world in terms of hierarchy of information and concepts, with each concept defined in terms of its relation to simpler concepts. In [1], the author defines Deep learning as *particular kind of machine learning that achieves great power and flexibility by learning to represent the world as a nested hierarchy of concepts, with each concept defined in relation to simpler concepts, and more abstract representations computed in terms of less abstract ones.* Thus, the hierarchy of concepts allows the computer to learn complicated concepts by building them out of simpler ones. However, if we draw a graph showing how these concepts are built on top of each other similar to human cognitive system functioning, the graph is deep, with many layers. For this reason, we call this approach as *Deep Learning.* In an attempt to simulate biological learning model of a human brain, scientists experiment their algorithms based on deep learning techniques and call them neural network models, inspired from neurons and whatever little knowledge we have about human brain.

Resounding Success of Deep Learning: a feather in hat or a myth

There are many different fundamental ideas to deep learning. In [3], *Convolutional Neural Network* is trained for classification of images on the subsets of ImageNet used in ILSVRC-2010 contest. They implemented this network with eight learned layers in which five were 2D convolutional and three were fully connected layers. Since, the dataset was very large, they used two GPUs which communicated mutually only at certain levels and thus, with cross GPU parallelization made training faster. However, the basis on which the connectivity between the two CPUs is chosen is purely experimental so as to obtain an acceptable fraction of the amount of computation. Design decisions made on experimental evidences cannot be generalised so easily as they purely come from the existing dataset and thus, may be biased with respect to that specific triaining set. Since, this neural network had around 60 million parameters, there was very high chance of over-fitting. To combat overfitting, they used 'Data Augmentation' by generating image translations and horizontal reflections, and also, changed the intensities of the RGB channels in the exisiting training images. Eventhough the size of the dataset increased by doing this, still this increased dataset was highly interdependent. 'Dropout' was another technique used to combat overfitting. The benefit of using convolutional neural network over standard

feedforward neural networks is that they have much lesser connections and parameters which makes it easier to train. Convolutional neural networks require a constant input dimensionality unlike the Recurrent Neural Network. Since, the Image Net consisted of variable resolution images, the images were downsampled to a resolution of 256X256 by rescaling and cropping the central patch of the image which sounds rather, arbitrary to downsample an image. There is a possibility that some images might have the target located in corners with target size being very small. By resampling and cropping blindly only central part of the image seems arbitrary and might lead to loss of important side patterns from the image. Also, ReLU non-linearity is used to model neuron's output in first four convolutional layer over tanh or sigmoid. However, if any large gradient flows through a ReLU neuron, then, this can cause the weights to update in a way that the neuron will never activate on any datapoint again. If ReLU ends up in this state, it is unlikely to recover, because the function gradient at 0 is also 0, so gradient descent learning will not alter the weights while the sigmoid and tanh neurons can suffer from similar problems as their values saturate, but there is always at least a small gradient allowing them to recover in the long term. Apart from this, the paper claims that the depth of the Convolutional neural network is really important to achieve the required results which clearly indicates that this process is very expensive in terms of resources like time, cost (requires GPUs for parallelization to make the training faster). Also, unsupervised pre-training on the dataset could have helped to increase size of network without an increase in labelled dataset as deeper the network, the more efficiently it will classify, provided the trade-off between resources and depth is optimized efficiently. Thus, Convolutional neural networks provide a way to specialize neural networks to work with data that has a clear grid-structured topology example images and to scale such models to very large size. To process one-dimensional sequential data, we turn now to another powerful specialization of the neural networks framework: Recurrent Neural Networks.

In [5], the blog post describes a simple architecture of Recurrent Neural Network and then, gives an interesting example of its application on generating samples of texts, programming code, latex scripts, based on its learning from the existing set of similar documents. While this is a great way of generating texts, this post reminded me of an assignment question from the class of Computational Linguistics in which we were asked to generate random text using n-grams. The traditional way for text synthesis involves storing the blocks of words and generating the next element in the sequence by sampling from the stored distribution given previous outputs. The advantage of using RNN is that there is now no requirement of fixing 'n' like that of n-gram model and better representations (in line with word2vec etc.) than literal words are used. However, for text generation, the cohesion during the long term cannot be maintained without introducing some structure outside of the details to which character comes next which seems to be a limitation of this. In blog post, the author writes that *If training vanilla neural nets is optimization over functions, training recurrent neural nets is optimization over programs.* This doesn't seem fully correct as training RNN is optimization over finite state automations mixed with multi layer perceptrons and not just, generally on 'program'. Also, if a text couldn't be expressed in form of a regular expression(finite state automata), then, gradient descent wouldn't be effective to get RNN to model a language. This might be the reason why the author writes about the mistake which RNN made with the latex model and she/he had to correct it manually. Also, probably the temporal sequences will be a slow step through the RNN states for events like punctuation etc. Additionally, linear regression is performed from hidden states to state properties. However, it seems difficult to predict if the current position is inside parenthesis or the length of the current sentence. RNN always have a problem with tasks which require to handle long-timescale working memory unless some external storage system is used. This is important to apply auto-correlation in letters or words to make this model more robust. One can also try using hard constraints by using hand-written rules of grammar or vocabulary or

logics i.e. writing explicit rules to the training network to enhance the model's output anyway. It is probably a good idea to make a heterogeneous model combining neural network and Logics as used in AlphaGo as discussed next. Why should a machine be made to learn something which is far more simpler to express in form of logics!

In [7], the authors talk about the development of a system which defeated the human European Go champion. The system is based on a model trained by a combination of Reinforcement Learning and Supervised Learning to train value network and policy network. The algorithm of this system works around Monte Carlo tree search programs as well. It combines Monte-Carlo tree search with deep neural networks that have been trained by supervised learning, from human expert games, and by reinforcement learning from games of self-play. Interestingly, this program is not purely a neural network based model, instead the Monte Carlo Tree search is used to estimate the value of each state in a search tree. It is a true fact that larger networks achieve better accuracy but are slower to evaluate during search. Thus, in order to construct a correct and current representation of the positions in the game, the board position image is passed through convolutional layers, thus, reducing the effective depth and breadth of the search tree, thereby evaluating positions from value network and sampling actions using a policy network. However, its representations are slower for both policy and value networks and thus, it takes more time to evaluate deep neural networks than linear representations.In prior, there used to be handcrafted patterns on which the evaluations used to be performed but in AlphaGo, they use rollouts and did position evaluation using value networks instead of handwritten rules. Infact, this concept could further be applied to other games as well in order to obtain, may be not precise but deterministic approximate answers to problems by working on developing hybrid models (Rule Based and Neural Network Based both) using techniques like Monte Carlo rollouts as used in AlphaGo. Had the scientists developed it using purely neural networks, this might have resulted in more slower and less ambitious model for AlphaGo which might have not led to this break-through.

In [6], the blog post talks about Google's Deep Dream Software which shows the learning of a neural network and shows the classification task which works by recognising patterns on the input images and would edit the image to look more like pattern. While this can also result in beautiful digital artwork, however, in [9], the post reads about how images of people with dark-skin pigment were labelled as 'gorillas' by the Google Algorithm. In [8], the post explicitly brings up the mistake with an example of how the convolutional neural network applied on images can go wrong. A machine which is trained with images of the animal- leopard and labelled as 'leopard' fails to assign correct label for a couch with a leopard print cover on it. Even if the convolutional neural networks does take the local features into consideration which results in transformational invariance, still, the object structure or its orientation is not known to the machine. Thus, we conclude that though convolutional neural networks occupy a significant position in current state of art of image classification still, there is lot of scope of work to be done in order to build a very accurate recognition algorithm which can identify every small elements of an image and label them correctly.

Deep Learning, Data Mining and Knowledge Discovery

Deep Learning and Data Mining, both aim at deep analysis of the raw data and retrieval of knowledge from it in a more explicit manner. There are currently many works which are heading in integrating these fields. In [4], the author suggests a Deep Learning based Data Mining model which integrates the supervised and unsupervised learning(Restricted Boltzman Machine and autoencoders) while also considering dynamic user inputs as well. With this integration, there have to be no hand-written engineering for identifying correlations or associations in data.

We can now also deal with unlabelled data. Raw data can be both numeric or Textual or both. There is a lot of work in the field of Information Retrieval from textual data like Relationship Extraction, Knowledge Graphs which have models based on neural networks. In [2], the author classifies the problems in Data Mining into four major categories: clustering, classification, association pattern mining, and outlier analysis. Neural network is giving new direction to improve the conventional data mining models but a tradeoff between its limitations like overfitting in case of large number of parameters, cost of resources (GPUs, CPUs, Time) for training large dataset on many-layered architecture is necessary. Also, a good trade-off between rule based system and Neural network could be another approach as well. The exponential explosion of available data, the rise of the graphics processing unit, the invention of advanced algorithms dealing with dimensionality reduction or feature detection, has brought deep learning into a position of solving complex problems which have troubled the community since ages. However, the black box problem which is the inability to know how an ANN reached its prediction and overfitting are a usual concern with this approach. Since progress in computer hardware is continuing, one might reasonably expect that the advances will arise from more powerful data storage and processing ability. Thus, deep learning is definitely not just another bottle of snake oil but is definitely for the better since the sliced bread.

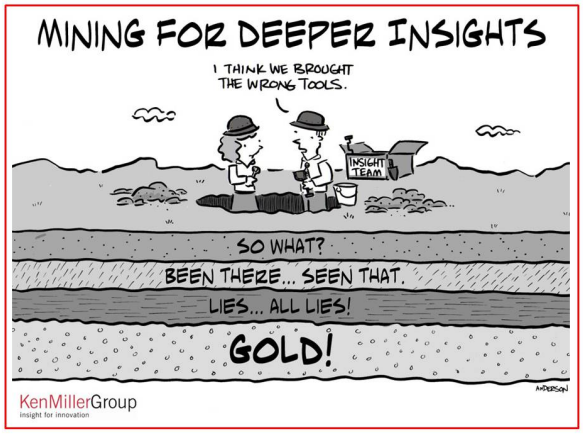

Figure 1: *(www.kenmillergroup.com)*

References

[1] Ian Goodfellow, Yoshua Bengio and Aaron Courville (2016) *Deep Learning* MIT Press

[2] Charu C. Aggarwal (2015) *Data Mining : The TextBook* Springer International Publishing

[3] Krizhevsky, A., Sutskever, I. and Hinton, G.E. *ImageNet Classification with Deep Convolutional Neural Networks.* In NIPS '12, pages 1097-1105, 2012.

[4] Y. Ma, Y. Tan, C. Zhang, Y. Mao *A data mining model of knowledge discovery based on the deep learning* In 2015 IEEE 10th Conference on Industrial Electronics and Applications

[5] Karpathy, A. *The Unreasonable Effectiveness of Recurrent Neural Networks.* , 2015.

[6] Titcomb, J. *Google unleashes machine dreaming software on the public, nightmarish images flood the internet.* , 2015.

[7] D. Silver, A. Huang, C.J. Maddison, A. Guez, L. Sifre, G.V.D. Driessche,J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe,J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel and D. Hassabis *Mastering the game of Go with deep neural networks and tree search.* Nature, 529(7587):484-489, 2016.

[8] Khurshudov, A. *Suddenly, a leopard print sofa appears.* , 2015.

[9] Curtis, S. *Google Photos labels black people as 'gorillas'.* , 2015.