NEURAL NETWORKS AS A TOOL FOR BIG DATA: THE STATE OF THE ART

Carlos Tarjano (tesserato@hotmail.com)

Valdecy Pereira (valdecy.pereira@gmail.com)

Abstract

This article aims to uncover the state of the art in the interface of big data and artificial neural networks fields, by means of a bibliographic review of scientific literature databases. In selecting which topics to be covered, and to what extent, in face of the prolific research in both areas, we favor a pragmatic approach guided, when applicable, by the concrete products or solutions that were enabled by the area of research reviewed. This approach is particularly fit for a fast growing, relatively new area, where state of the art in research is almost instantly applied, and barriers between industry and community led research, while still present, are very low.

key words: Neural Networks, Big Data, State of the art, Review

Resumo

Este artigo pretende apresentar o estado da arte na interface entre big data e redes de redes nervosas artificiais, por meio de uma revisão bibliográfica de bancos de dados da literatura científica. Ao selecionar os tópicos a serem abordados e em que medida, diante da abundância de pesquisas em ambas as áreas, favorecemos uma abordagem pragmática orientada, quando aplicável, pelos produtos ou soluções concretas que foram habilitadas pela área de pesquisa revisada. Esta abordagem é particularmente adequada para uma área de crescimento rápido, relativamente nova, onde o estado da arte em pesquisa é aplicado quase que instantaneamente, e as barreiras entre a indústria e pesquisas lideradas pela comunidade, conquanto ainda estejam presentes, são muito baixas.

palavras-chave: Redes Neurais, Big Data, Estado da Arte, Revisão Bibliográfica

Resumen

Este artículo tiene como objetivo descubrir el estado del arte en la interfaz de los grandes datos y los campos de redes neuronales artificiales, a través de una revisión bibliográfica de bases de datos de literatura científica. Al seleccionar los temas que deben cubrirse, y hasta qué punto, frente a la investigación prolífica en ambas áreas, favorecemos un enfoque pragmático guiado, cuando corresponda, por los productos concretos o soluciones que fueron habilitados por el área de investigación revisada. Este enfoque es particularmente adecuado para un área relativamente nueva y de rápido crecimiento, donde el estado del arte en pesquisa se aplica de forma casi instantánea, y las barreras entre la industria y la investigación liderada por la comunidad, aunque todavía están presentes, son muy bajas.

palabras clave: Redes neuronales, Big Data, estado de la técnica, revisión bibliográfica

1 Introduction

Artificial neural networks are the de facto standard for processing the massive volume of data introduced in the Big Data age. Familiarity with this area is, thus, of the foremost importance for those willing to extract meaningful information from the huge databases available today, fed with virtually every detail of our online activities.

They brought a new possible approach to problems that don't lend themselves well to conventional techniques due to their inherent flexibility and capacity of automatically extracting features from massive amounts of raw data. Benefiting from the growing availability of information, and advances in the theory underlying the field, alongside with improvements in computer processing power, they're already employed in a great variety of research fields, having potential to be useful in many others.

The widespread application of neural networks in many areas naturally led to a high degree of fragmentation in the literature and nomenclature used, leading to cases of redundant research efforts and somewhat hindering the access of new researchers interested in expanding the field or, more pragmatically, seeking solutions to problems in their own areas of expertise. To address this question, this work offers a brief introduction to the most common architectures e concepts found in the field.

It then proceeds by investigating the usefulness of neural networks, in its many architectural incarnations, as a tool not only to be used in the context of big data, but whose utility has been made possible by it. In line with the endeavor of presenting the state of the art from an applicated point of view, this work provides an overview of the major developments led by top technology companies such as Google, Facebook and Microsoft.

2 Overview of Neural Networks

Being a mostly empirical field, experimentation gave rise to a great variation in network types and topologies. We present below a non-exhaustive enumeration of the most prominent types of networks while trying to establish, at the same time, a graduation of complexity and maintaining, when possible, a chronological order.

2.1 Architectures

2.1.1 Perceptron / Multilayer Perceptron

A Perceptron, one of the first steps towards the concrete use of neural networks, is a linear classifier model proposed in the late fifties by Frank Rosenblatt, capable of updating its weights to learn how to correctly classify linearly separable classes based on inputs and examples of desirable outputs. It consists of an activation function applied over the weighted sum of the inputs and a bias. Originally intended to be built as a machine for the Cornell Aeronautical Laboratory (Rosenblatt, 1957), its first implementation was in an IBM 704 computer (Bishop, 2006).

Multilayer Perceptrons are obtained by fully connecting layers of Perceptrons. Despite its apparent simplicity, this topology has been proven (Hornik, 1991) to potentially work as an arbitrary precision function approximator, given a sufficient number of hidden neurons, and arbitrary, non - polynomial (Leshno et al., 1993), activation functions. Together with backpropagation and gradient descent, an efficient way of training networks proposed by Paul Werbos (Werbos, 1974) in mid-seventies, they form the basis of the field. Below are two equivalent representations of a Perceptron; the left hand one illustrates its workings in more detail, while the other will be useful as a building block to the other architectures, presented in figure 1.

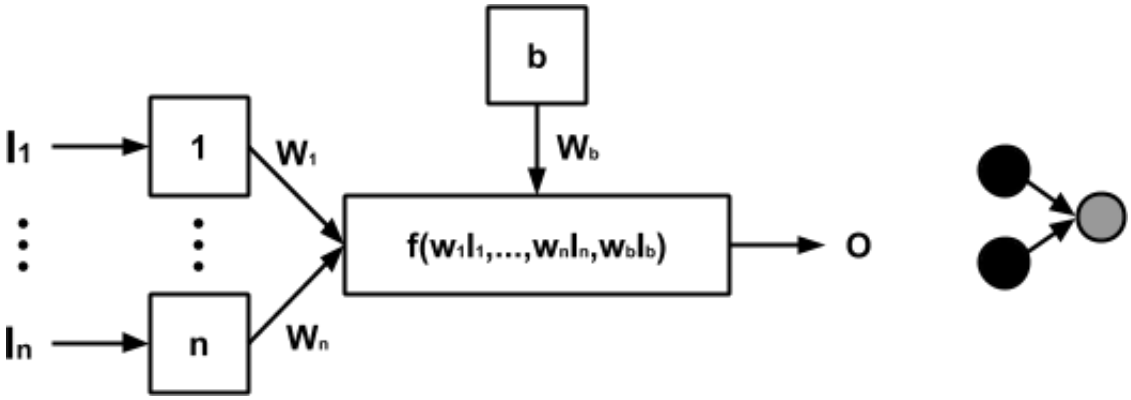

Figure 1 - Representation of the Perceptron
Source: Elaborated by the authors

2.1.2 Feed-Forward Neural Networks

The principal characteristic of this network, as the name implies, is the forward motion of the inputs, without loops or cycles: Each signal flows from one fully connected layer to the next, until they reach the output nodes (Ojha, Abraham, and Snášel, 2017). Common activation functions are sigmoidal, like the logistic $y = \frac{1}{1+e^{-x}}$ and the hyperbolic tangent $\tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$ (Goldberg, 2016). They are a class encompassing single and multi layer Perceptrons.

2.1.3 Deep Neural Networks

Deep neural networks are, in essence, feed-forward networks with many hidden layers, and thus, with a greater capacity of abstraction that, on the other hand, pose more difficulty to the training process. One of the main advantages of this type of topology is that the time-consuming task of extracting features from raw data is left to the networks themselves (Socher, 2014). Within each layer, more high-level representations can be created, based on the reasoning of anterior layers. One drawback of this technique, however, is that it's hard to infer exactly how this process is happening in the network.

2.1.4 Recurrent

In recurrent networks the neurons are partially fed with their own states from past iterations, generating an effect that is similar to adding links from neurons to posterior neurons (Veit, Wilber, and Belongie, 2016) in non-adjacent layers. This formulation, proposed by Elman (1990) with the aim of capturing features that could arise from serialization in data is quite powerful in some applications, as information tagging (W. Xu, Auli, and Clark, 2015) for example.

2.1.5 Convolutional

This type of (deep) feed-forward network is most commonly used in image applications, were they achieved state of the art performance in various tasks involving, for example, object and face recognition in images (Pang et al., 2017). Instead of receiving each pixel that forms an image as an input, a situation where fully connected layers can suffer rapidly from the curse of dimensionality, in this kind of network, the subsequent layers are sparsely connected with each other, conferring better trainability when compared with standard deep networks.

The convolutional layers can be interpreted as scanning the image (LeCun et al., 1998), one stride at a time (generally 1 or 2 pixels), and storing the result of each step in a point of the subsequent layer. This allows for the detection of the relevant feature in different places of the image. To provide better generalization, there are also layers that perform sub-sampling,

simply reducing the dimensions of the anterior layer by averaging portions of them (or taking their maximum). After an arbitrary number of convolutional and sub-sampling layers, the actual reasoning is performed by a (series of) traditional, fully connected layers.

2.1.6 Hopfield / Associative Memory Networks

Introduced by Hopfield (Hopfield, 1982) in 1982, this architecture, where each neuron is connected to all others, is among the most studied (Y. Wu et al., 2012), and one of its foremost applications is in the associative memory field, were, presented with an (imperfect) example of an object, the trained network can then retrieve its closest approximation.

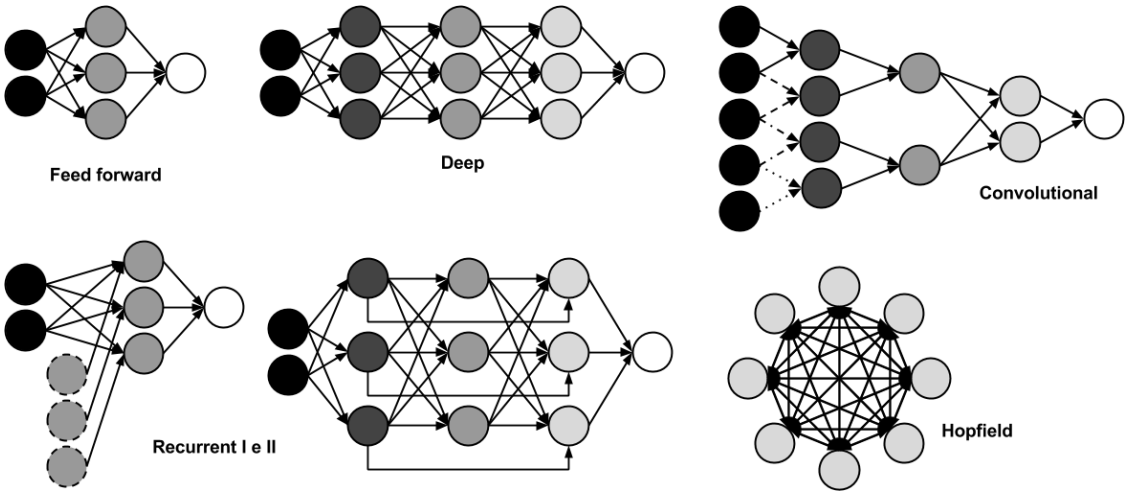

Figure 2 - Representation of different topologies
Source: Elaborated by the authors

2.2 Concepts

2.2.1 Support Vector Machines

While not strictly a neural network, this machine learning classification algorithm is closely related (Collobert and Bengio, 2004) to both the Perceptron and the multi-layer Perceptron. Support Vector Machines, as proposed by Vapnik (Boser, Guyon, and Vapnik, 1992) in 1982, are linear classifiers.

Replacing the dot product in the original algorithm with a kernel function allows nonlinear classification by means of transforming the original nonlinear space into a linearly separable space. This first proposition already covered its applicability to Perceptrons, thus albeit not strictly a member of the neural network family, the use of the kernel function, commonly known as the kernel trick, is widely adopted in the field e.g. P. Zhu and Príncipe (2013), K.-K. Huang et al. (2017), Song et al. (2017)), and hence justifies the presentation of Support Vector Machines here.

2.2.2 Extreme Learning Machines

Rather than a topology, this is an analytical method for training feed-forward neural networks with one hidden layer, proposed in 2004 (G.-B. Huang, Zhu, and Siew, 2004), with the aim of speeding up the traditional back propagation training algorithm.

2.2.3 Generative Adversarial Networks

This technique/topology introduces the concept of one generative deep network creating content to be classified as artificially generated or not by another network (Goodfellow et al., 2014).

3 Neural Networks and Big Data: The state of the art

Neural Networks recently regained popularity in the exact measure as big data emerged. Particularly in the case of deep architectures, this relationship is almost symbiotic: while deep networks can make sense of large amounts of data that are intractable by other techniques, automatically extracting meaningful features from that data they, on the other hand, are very difficult to train, requiring large amounts of training samples to do their jobs in a satisfactory way.

While an investigation in research paper databases of the applications of neural networks in the big data field brings to light interesting works, we argue here that the observation of the research efforts of big companies like Google and Microsoft are just as insightful, if not more. The rationale is simple: for them, dealing in the most efficient and meaningful way with large databases of raw data is simply a matter of life or death, and it's possible to see that those organizations do so with the use of neural networks. It's also worth pointing out that most of those companies also incorporated this research in at last one of their first-class products, explicit stablishing the state of the art of the interface between big data and neural networks. the image below illustrates the volume of papers, from Google research website, in different areas.

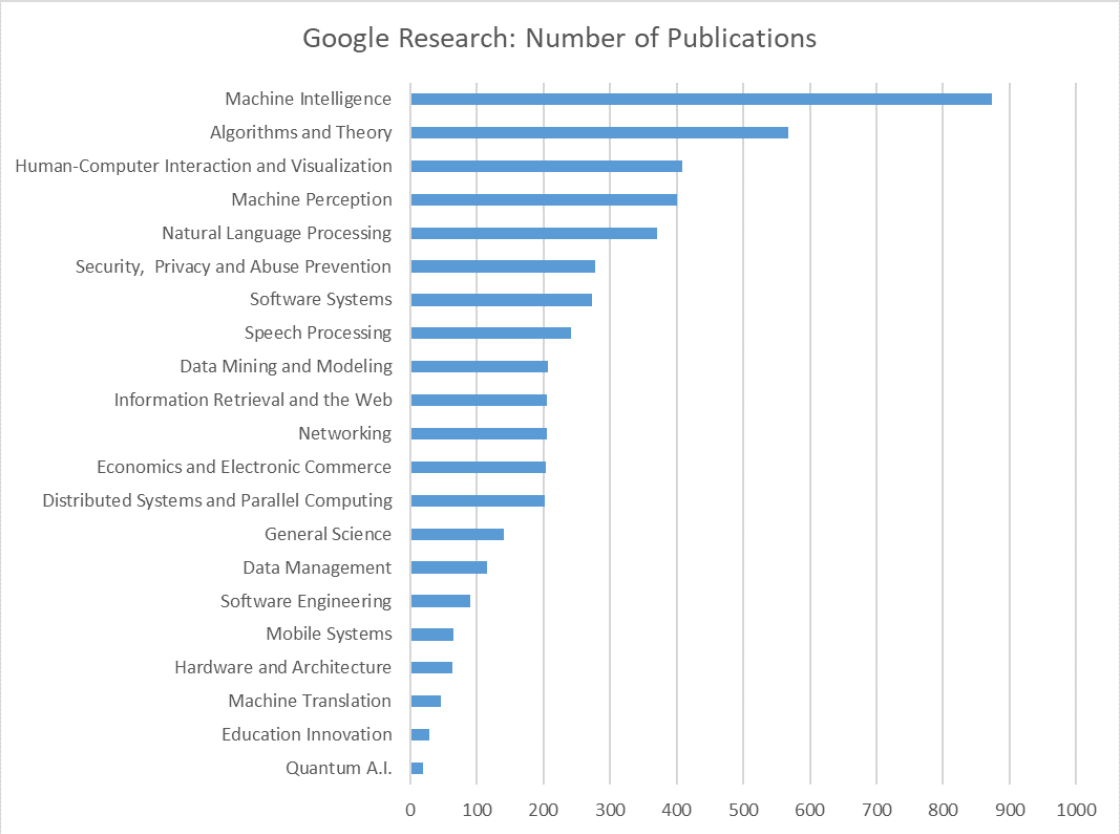

Figure 3 - Research at Google: publications by area
Source: Research at Google website

We notice that Machine Intelligence is by far the most researched area, and of the five top areas, four are directly related to machine learning and big data.

The state of research at Microsoft is less straightforward to assess, as such a summary is not readily available, but the enumeration of topics of interest can give some hints that the situation is not considerably different: Artificial Intelligence, Computer Vision, Human-Computer Interaction and Human Language Technologies are listed along with 14 other areas.

A similar pattern can be found in Facebook research efforts, where we find 5 related areas listed among eleven. Apple recently launched a blog with the solely purpose of publishing research in the artificial intelligence field. While Amazon's research efforts are less easy to pinpoint, and presumably more diffuse as we consider the company's broader area of actuation, it's also worth noting that Amazon's Echo is probably the most commercially successful artificial intelligence/big data based product to date, and some meaningful research endeavors ought to be pointing in that direction.

Considering that those are some of the top R&D investing companies in the word makes clear to see that the interface between big data and artificial intelligence is one of the biggest trends in research and application. In this context, we review below some specific publications, to analyze the specific technologies that constitute the state of the art.

3.1 Natural Language Processing

As most of the data stored today is in written form, be it digitalized or not, in a way, independent from the language, that is conceived to make sense to humans, research in this area is of the foremost importance, as it allows the emerging artificial intelligence technologies to tap into this enormous source. Many machine learning subareas are concerned with different stages of this problem, such as computer vision, that is utilized to transform raw scanned pages into strings of characters, to be further analyzed by other methods. One of the most active research areas concerning the use of neural networks, computer vision encompasses task such as image classification and character recognition.

To further make sense of strings, techniques from natural language processing are employed, an area that is actively researched in companies such as Google. Deng, Hinton, and Kingsbury (2013) offer an overview of the field, as of 2013, based on papers submitted to a special session at ICASSP-2013. A more recent short survey is presented by Goldberg (2016) with the aim of acquainting newcomers to the field, fomenting the migration of sparse liner models to dense neural network approaches.

To illustrate an application in this area, the work of See, Liu, and Manning (2017) offers a state of the art approach for abstractive text summarization, based on a long-short-term memory neural network. This architecture, improving the work of Nallapati et al. (2016), that employs sequence-to-sequence recurrent neural network, offers summaries of text that are not restricted to the terms actually present in the body of the text, instead learning to contextualize from training fed by other articles. To speed up the processing of large amounts of textual data, A. W. Yu, Lee, and Le (2017) offers insight about text skimming via a recurrent neural network, that learns which parts of the text it can safely ignore, while offering interpretation accuracy on par with conventional models.

Those automatic reasonings about data are only useful in the extent in which they can be used to inform human decisions, and in this context one can bring to light the work of Lu et al. (2017), that provides an interface for querying neural networks in natural language, via knowledge transfer from a maximum likelihood estimation trained generative neural dialog model, that tend to provide conservative answers, like "I don't know", to discriminative dialog models, that poses no such a difficult but, in the other hand, are incapable of being employed in natural language dialogs.

Speech processing can be interpreted as a generalization of the written case, as, in addition to the difficulties of the last, it has to deal with acoustic processing in order to translate raw sound into more useful representations. Products like Microsoft's Cortana, Google Home, Apples Siri and Amazon Echo all make use of advancements in this area, that are presented, in many cases, in papers, such those by Xiong et al. (2017) and B. Li et al. (2017). Both present state of the art research to be incorporated in the next iteration of their respective products. Also presenting a research approach addressing concrete production problems, Zhuang et al. (2017) offer some insight into how Apple deal with the problem of

porting trained natural language processing neural networks into multiple, diverse platforms, as the iPhone and Apple Watch. In an entry in its artificial intelligence blog, Apple also presents research in improving Siri's voice via a deep neural network.

4 Conclusion

To the best of our knowledge, there are no up to date guides on the state of the art of neural network applications in the big data context, from which (potential) researchers in the field could benefit. we hope to have presented in this work one such a guide, that could be used to gain an overview of the results achievable with the application of neural networks. One obvious of this work is to identify research opportunities. One can also use this review to gain a more general view of the field, allowing knowledge transfer to correlated areas.

We foresee a crescent increase of interest in this area in at last the near future, with more concrete applications, in many fields, coming to life in the same proportion as databases grow in size and complexity, and computer power becomes more available.

5 References

Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning*. springer.

Boser, Bernhard E, Isabelle M Guyon, and Vladimir N Vapnik. 1992. "A Training Algorithm for Optimal Margin Classifiers." In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 144–52. ACM.

Collobert, Ronan, and Samy Bengio. 2004. "Links Between Perceptrons, Mlps and Svms." In *Proceedings of the Twenty-First International Conference on Machine Learning*, 23. ACM.

Deng, Li, Geoffrey Hinton, and Brian Kingsbury. 2013. "New Types of Deep Neural Network Learning for Speech Recognition and Related Applications: An Overview." In *Acoustics, Speech and Signal Processing (Icassp), 2013 Ieee International Conference on*, 8599–8603. IEEE.

Elman, Jeffrey L. 1990. "Finding Structure in Time." *Cognitive Science* 14 (2). Wiley Online Library: 179–211.

Goldberg, Yoav. 2016. "A Primer on Neural Network Models for Natural Language Processing." *J. Artif. Intell. Res.(JAIR)* 57: 345–420.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. "Generative Adversarial Nets." In *Advances in Neural Information Processing Systems*, 2672–80.

Hopfield, John J. 1982. "Neural Networks and Physical Systems with Emergent Collective Computational Abilities." *Proceedings of the National Academy of Sciences* 79 (8). National Acad Sciences: 2554–8.

Hornik, Kurt. 1991. "Approximation Capabilities of Multilayer Feedforward Networks." *Neural Networks* 4 (2). Elsevier: 251–57.

Huang, Guang-Bin, Qin-Yu Zhu, and Chee-Kheong Siew. 2004. "Extreme Learning Machine: A New Learning Scheme of Feedforward Neural Networks." In *Neural Networks, 2004. Proceedings. 2004 Ieee International Joint Conference on*, 2:985–90. IEEE.

Huang, Ke-Kun, Dao-Qing Dai, Chuan-Xian Ren, and Zhao-Rong Lai. 2017. "Learning Kernel Extended Dictionary for Face Recognition." *IEEE Transactions on Neural Networks and Learning Systems* 28 (5). IEEE: 1082–94.

LeCun, Yann, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. "Gradient-Based Learning Applied to Document Recognition." *Proceedings of the IEEE* 86 (11). IEEE: 2278–2324.

Leshno, Moshe, Vladimir Ya Lin, Allan Pinkus, and Shimon Schocken. 1993. "Multilayer Feedforward Networks with a Nonpolynomial Activation Function Can Approximate Any Function." *Neural Networks* 6 (6). Elsevier: 861–67.

Li, Bo, Tara Sainath, Arun Narayanan, Joe Caroselli, Michiel Bacchiani, Ananya Misra, Izhak Shafran, et al. 2017. "Acoustic Modeling for Google Home." *INTERSPEECH-2017*.

Lu, Jiasen, Anitha Kannan, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2017. "Best of Both Worlds: Transferring Knowledge from Discriminative Learning to a Generative Visual Dialog Model." In *Advances in Neural Information Processing Systems*, 313–23.

Nallapati, Ramesh, Bowen Zhou, Caglar Gulcehre, Bing Xiang, and others. 2016. "Abstractive Text Summarization Using Sequence-to-Sequence Rnns and Beyond." *arXiv Preprint arXiv:1602.06023*.

Ojha, Varun Kumar, Ajith Abraham, and Václav Snášel. 2017. “Metaheuristic Design of Feedforward Neural Networks: A Review of Two Decades of Research.” *Engineering Applications of Artificial Intelligence* 60. Elsevier: 97–116.

Pang, Yanwei, Manli Sun, Xiaoheng Jiang, and Xuelong Li. 2017. “Convolution in Convolution for Network in Network.” *IEEE Transactions on Neural Networks and Learning Systems*. IEEE.

Rosenblatt, Frank. 1957. *The Perceptron, a Perceiving and Recognizing Automaton Project Para*. Cornell Aeronautical Laboratory.

See, Abigail, Peter J Liu, and Christopher D Manning. 2017. “Get to the Point: Summarization with Pointer-Generator Networks.” *arXiv Preprint arXiv:1704.04368*.

Socher, Richard. 2014. “Recursive Deep Learning for Natural Language Processing and Computer Vision.” Citeseer.

Song, Qing, Xu Zhao, Haijin Fan, and Danwei Wang. 2017. “Robust Recurrent Kernel Online Learning.” *IEEE Transactions on Neural Networks and Learning Systems* 28 (5). IEEE: 1068–81.

Veit, Andreas, Michael J Wilber, and Serge Belongie. 2016. “Residual Networks Behave Like Ensembles of Relatively Shallow Networks.” In *Advances in Neural Information Processing Systems*, 550–58.

Werbos, Paul John. 1974. “Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences.” *Doctoral Dissertation, Applied Mathematics, Harvard University, MA*.

Wu, Yue, Jianqing Hu, Wei Wu, Yong Zhou, and KL Du. 2012. “Storage Capacity of the Hopfield Network Associative Memory.” In *Intelligent Computation Technology and Automation (Icicta), 2012 Fifth International Conference on*, 330–36. IEEE.

Xiong, Wayne, Jasha Droppo, Xuedong Huang, Frank Seide, Mike Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig. 2017. “The Microsoft 2016 Conversational Speech Recognition System.” In *Acoustics, Speech and Signal Processing (Icassp), 2017 Ieee International Conference on*, 5255–9. IEEE.

Xu, Wenduan, Michael Auli, and Stephen Clark. 2015. “CCG Supertagging with a Recurrent Neural Network.” In *ACL (2)*, 250–55.

Yu, Adams Wei, Hongrae Lee, and Quoc V Le. 2017. “Learning to Skim Text.” *arXiv Preprint arXiv:1704.06877*.

Zhu, Pingping, and José C Príncipe. 2013. “Kernel Recurrent System Trained by Real-Time Recurrent Learning Algorithm.” In *ICASSP*, 3572–6.

Zhuang, Xiaodan, Arnab Ghoshal, Antti-Veikko Rosti, Matthias Paulik, and Daben Liu. 2017. “Improving Dnn Bluetooth Narrowband Acoustic Models by Cross-Bandwidth and Cross-Lingual Initialization.” *Proc. Interspeech 2017*, 2148–52.