# 1 Introduction

Without going into the whole history of Neural Networks, and various ways of training them or how they correspond to various paradigms of learning, it is assumed that people are familiar with gradient based optimization and the approximation capabilities of general neural networks with sigmoidal family of activations.

# 2 Early Work

The first experiments took place in, what's known as *Deep* Networks, with Salakhutdinov's work and continued with Srivastava's Dropout paper. The breakthrough results were obtained with variants of Boltzmann's Machines and Convolutional Networks.

Although Jurgen Schmidhuber, who's himself made very significant contributions in the field of Connectioninst Architectures and Artificial Intelligence in general propounds an alternative version of development of the field. Which is sort of an antithesis to the ones propagated by the Hinton and co.

- Jürgen Schmidhuber. "Deep learning in neural networks: An overview". In: *Neural Networks* 61 (2015), pp. 85–117
- Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *Nature* 521.7553 (2015), p. 436

However Hinton and co. have had the major chunks of breakthroughs in recent years.

- Geoffrey E Hinton and Ruslan R Salakhutdinov. "Reducing the dimensionality of data with neural networks". In: *Science* 313.5786 (2006), pp. 504–507
- Ruslan Salakhutdinov and Hugo Larochelle. "Efficient learning of deep Boltzmann machines". In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. 2010, pp. 693–700
- Geoffrey E Hinton et al. "Improving neural networks by preventing co-adaptation of feature detectors". In: *arXiv preprint arXiv:1207.0580* (2012)
- Nitish Srivastava et al. "Dropout: A simple way to prevent neural networks from overfitting". In: *The Journal of Machine Learning Research* 15.1 (2014), pp. 1929–1958

The watershed moment was the ImageNet ILSVRC 12 challenge won by a Krizhevsky's *AlexNet*

- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in Neural Information Processing Systems*. 2012, pp. 1097–1105

Which used those techniques including ReLU

- Vinod Nair and Geoffrey E Hinton. "Rectified linear units improve restricted boltzmann machines". In: *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*. 2010, pp. 807–814

Combined with the use of two GPUs for distributing the large model and computational power possible with that, they were able to train those Convolutional Networks.

### 2.1 LSTMs and Gated Units

However, there was another (re) discovery that was going to be as profound, if not more so, than AlexNet. LSTM (Long Short Term Memory was a very (relatively) old heuristic to help Recurrent Networks learn longer sequences. RNNs were thought to be notoriously hard to train because of the dynamics involved in the nonlinearities of the activations and gradient based optimization was thought inefficient for it. LSTMs sought to solve that issue with a simple cell architecture.

- Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory". In: *Neural Computation* 9.8 (1997), pp. 1735–1780

- Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. "Learning to forget: Continual prediction with LSTM". in: (1999)

The *forget* cell was added in the latter work, which improved the sequence learning tasks as it learned to forgot context which was no longer relevant.

Since then other cell based *Gated* architectures have been proposed of which GRUs are the most pertinent which seek to decrease the computational and memory cost associated with LSTMs while having little effect on performance, though there is some variability in the results.

- Kyunghyun Cho et al. "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation". In: (2014)

- Junyoung Chung et al. "Empirical evaluation of gated recurrent neural networks on sequence modeling". In: *arXiv preprint arXiv:1412.3555* (2014)

## 3 CNNs

The way CNNs work is by building a hierarchy of features. They've been around for a long time but it just was not feasible to train them in the 90s. The capacity

and redundancy of features in the network is a matter of contention with several efforts to reduce the number of parameters while keeping the generalization ability intact. Consult following for a survey

- Yu Cheng et al. "A Survey of Model Compression and Acceleration for Deep Neural Networks". In: *arXiv preprint arXiv:1710.09282* (2017)

## 3.1 Understanding CNNs

The premise of the Convolutional Networks is the pyramidal architecture, which has been around in Vision literature for ages and Fukushima's Neocognitron. For the kind of features the CNNs learn Krizhevsky 2009, is a good source.

The following sources explore CNNs and the features they learn more in depth.

- Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps". In: *arXiv preprint arXiv:1312.6034* (2013)
- Matthew D Zeiler and Rob Fergus. "Visualizing and understanding convolutional networks". In: *European Conference on Computer Vision*. Springer. 2014, pp. 818–833
- Ross Girshick et al. "Rich feature hierarchies for accurate object detection and semantic segmentation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 580–587

## 3.2 Object Classification Networks

Object Classification aims to map an image to a label, generally one out of 5 possible labels as there can be multiple objects in an image so in effect the network learns saliency as well.

Krizhevsky 2012, being the first of the modern CNNs which excelled at that task, rest which improved upon that significantly and in a novel manner were:

- Zeiler and Fergus, "Visualizing and understanding convolutional networks"
- Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014)
- Christian Szegedy et al. "Going Deeper With Convolutions". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2015
- Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 770–778

### 3.3 Object Localization

The hierarchy of features and how they can be used was explored efficiently for multiscale object detection and localization in Girshick 2014,

They termed it Region CNN as it operated on Region Proposal Mechanism. They integrated it finally into an end to end model in Faster RCNN. These are now the state of the art in Object Localization

- Girshick et al., "Rich feature hierarchies for accurate object detection and semantic segmentation"
- Ross Girshick. "Fast R-CNN". in: *Computer Vision (ICCV), 2015 IEEE International Conference on*. IEEE. 2015, pp. 1440–1448
- Shaoqing Ren et al. "Faster R-CNN: Towards real-time object detection with region proposal networks". In: *Advances in Neural Information Processing Systems*. 2015, pp. 91–99

## 4 RNNs and Natural Language Processing

Recurrent Neural Networks have been around since the 60s. An alternative and very interesting history can be traced in:

- S. Grossberg. "Recurrent neural networks". In: *Scholarpedia* 8.2 (2013). revision #138057, p. 1888. DOI: 10.4249/scholarpedia.1888

Otherwise assuming one is familiar with what a RNN is and from the earlier references to LSTMs, the real advances in the architectures came with trying to solve the Language Generation (or Modelling) task with language being represented as a sequence.

An early attempt to model a sequence of characters, which can be replicated easily with modern hardware is

- Ilya Sutskever, James Martens, and Geoffrey E Hinton. "Generating text with recurrent neural networks". In: *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. 2011, pp. 1017–1024

This model treats language as a sequence of characters and gives fairly impressionable results. However, that was before the modern LSTM boom and they instead would be the second and higher order gradient methods (not to be confused with higher order connection networks)

But the problem compounded with word representations as the number of possible words is much more than the number of characters and mapping so many labels becomes difficult. To that work on Word Embeddings was developed as follows:

- Yoshua Bengio et al. "A neural probabilistic language model". In: *Journal of Machine Learning Research* 3.Feb (2003), pp. 1137–1155

- Holger Schwenk. "Continuous space language models". In: *Computer Speech & Language* 21.3 (2007), pp. 492–518

These were first couple of attempts to model the word representations in a distributed manner. An embedding of the words from a discrete space to a continuous (although with ReLU they are piecewise continuous but for approximation purposes that's sufficient) was imperative for any continuous model to learn Language Models built on them. Work to this end was carried forward into RNN Language Models

- Tomáš Mikolov et al. "Recurrent neural network based language model". In: *Eleventh Annual Conference of the International Speech Communication Association*. 2010

- Tomas Mikolov et al. "Efficient Estimation of Word Representations in Vector Space". In: *arXiv preprint arXiv:1301.3781* (2013)

- Tomas Mikolov et al. "Distributed representations of words and phrases and their compositionality". In: *Advances in Neural Information Processing Systems*. 2013, pp. 3111–3119

- Quoc Le and Tomas Mikolov. "Distributed representations of sentences and documents". In: *International Conference on Machine Learning*. 2014, pp. 1188–1196

First major real world Language Modelling task achieved with LSTMs was

- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. "Sequence to sequence learning with neural networks". In: *Advances in Neural Information Processing Systems*. 2014, pp. 3104–3112

They used Bidirectional LSTMs to automatically translate English to French and vice versa. This work was done at Google and it was one of the first steps which led to the complete overhaul of their translation system.

- *The Great AI Awakening*. `https://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html`. Accessed: 2018-01-28