Contents

1 Look for 2 other papers on the same topic **2**

2 Related papers **2**
2.1 Outrageously large neural networks 2

3 Write it all out here before putting into latex **3**

4 **3**

5 I have to write a report **3**

6 **4**

7 Criticism **4**

8 Commentry **4**

9 Article / Charts **4**

10 Problem or topic **4**

11 Contribution **4**

12 [New] Network / algorithm / technique **5**
12.1 How does it work? 5
12.2 Is it suited to the task? 5
12.3 Has it been well tested 5
12.4 Does it really work as claimed? 5
12.5 What are the limitations? 5

13 Which kind of network was chosen **5**
13.1 Why was it chosen? 5
13.2 Was it the right one? 5
13.3 Is it clearly described 5
13.3.1 Parameters 5
13.3.2 Settings 5

14 What strengths and/or weaknesses of the NN approach does it illustrate? **5**

15 Own questions / Additional relevant information **5**

16 Assignment 2 structure **5**
16.1 Title 5
16.2 Abstract 5
16.3 Introduction 6
16.4 Conclusion 6
16.5 References 6
16.5.1 The tutorial 6
16.6 reference your chosen papers 7
4 Page report on the topic of your choice.

1 Look for 2 other papers on the same topic

2 Related papers

2.1 Outrageously large neural networks

Under review as a conference paper at ICLR 2017

/home/shane/dump/home/shane/notes2018/projects/ir-assignment-2/1701.06538.pdf

Don't have to go into too much detail.

Language modelling.

The capacity of a neural network to absorb information is limited by its number of parameters.

Conditional computation, where parts of the network are active on a per-example basis, has been proposed in theory as a way of dramatically increasing model capacity without a proportional increase in computation. In practice, however, there are significant algorithmic and performance challenges. In this work, we address these challenges and finally realize the promise of conditional computation, achieving greater than 1000x improvements in model capacity with only minor losses in computational efficiency on modern GPU clusters.

We introduce a Sparsely-Gated Mixture-of-Experts layer (MoE), consisting of up to thousands of feed-forward sub-networks. A trainable gating network determines a sparse combination of these experts to use for each example. We apply the MoE to the tasks of language modeling and machine translation, where model capacity is critical for absorbing the vast quantities of knowledge available in the training corpora.

We present model architectures in which a MoE with up to 137 billion parameters is applied convolutionally between stacked LSTM layers. On

large language modeling and machine translation benchmarks, these models achieve significantly better results than state-of-the-art at lower computational cost

3 Write it all out here before putting into latex

4

Machine Learning for Systems and Systems for Machine Learning pdf | Hacker News

5 I have to write a report

Andrew wanted something from one of the journals he asked me to go looking through

Here is a different kind of The Great Convergence: when neural networks go after data structures, (hashes, etc....) and eventually database systems....

The Case for Learned Index Structures by Tim Kraska, Alex Beutel, Ed H. Chi, Jeffrey Dean, Neoklis Polyzotis Indexes are models: a B-Tree-Index can be seen as a model to map a key to the position of a record within a sorted array, a Hash-Index as a model to map a key to a position of a record within an unsorted array, and a BitMap-Index as a model to indicate if a data record exists or not. In this exploratory research paper, we start from this premise and posit that all existing index structures can be replaced with other types of models, including deep-learning models, which we term learned indexes. The key idea is that a model can learn the sort order or structure of lookup keys and use this signal to effectively predict the position or existence of records. We theoretically analyze under which conditions learned indexes outperform traditional index structures and describe the main challenges in designing learned index structures. Our initial results show, that by using neural nets we are able to outperform cache-optimized B-Trees by up to 70% in speed while saving an order-of-magnitude in memory over several real-world data sets. More importantly though, we believe that the idea of replacing core components of a data management system through learned models has far reaching implications for future systems designs and that this work just provides a glimpse of what might be possible.

~2X space improvement over Bloom Filter at same false positive rate

6

research questions
    main contribution of each paper you select

7 Criticism

https://dawn.cs.stanford.edu/2018/01/11/index-baselines/

8 Commentry

https://news.ycombinator.com/item?id=12815231

9 Article / Charts

https://arstechnica.com/information-technology/2016/10/google-ai-neural-network-crypto

10 Problem or topic

11 Contribution

What is new?

12 [New] Network / algorithm / technique

12.1 How does it work?

12.2 Is it suited to the task?

12.3 Has it been well tested

12.4 Does it really work as claimed?

12.5 What are the limitations?

13 Which kind of network was chosen

13.1 Why was it chosen?

13.2 Was it the right one?

13.3 Is it clearly described

13.3.1 Parameters

13.3.2 Settings

14 What strengths and/or weaknesses of the NN approach does it illustrate?

15 Own questions / Additional relevant information

`https://news.ycombinator.com/item?id=15894896`

16 Assignment 2 structure

16.1 Title

Report on the convergences of neural networks with information retrieval

16.2 Abstract

This report goes over what I've learned about some of the convergences of neural networks and information retrieval. I've tried to take an unbiased approach. After reading 'The Case for Learned Indexes', I wanted to make a case for Classical Data Structures too so I researched other opinions. I

will quickly go over some of the applications of deep learning to IR. Then I'll go into depth with "The case for learned index structures", where I'll acknowledge some of the benefits of learned indexes. Finally, I will review some examples of standard data structures and show that they still have a place.

16.3 Introduction

Machine learning plays a role in many aspects of modern IR systems, and deep learning is applied in all of them, such as in semantic matching, learning to rank, modelling user behaviour and learning to index.

Neural networks are going after the data structures now, (B-trees, hashes, etc.). Deep learning has its tendrils all over modern IR systems.

It is interesting to see what key insights into IR problems the new technologies are able to give us and to clarify what is the best tool for the job.

16.4 Conclusion

standard data structures and show that they still have a place.

16.5 References

16.5.1 The tutorial

Gives a clear overview of current tried-and-trusted neural methods in IR and how they benefit IR research. It covers key architectures, as well as the most promising future directions.

Key architectures

1. promising future directions
   (a) [?] NIPS `https://arxiv.org/pdf/1706.03762.pdfw`
      i. commentary `https://news.ycombinator.com/item?id=15938082` Neural Networks for Information Retrieval `https://arxiv.org/pdf/1707.04242.pdf` `https://github.com/nn4ir/nn4ir.github.io` Slides `https://github.com/nn4ir/nn4ir.github.io/tree/master/sigir2017/slides`
      Semantic matching: methods for supervised, semi- and unsupervised learning for semantic matching.
      Learning to Rank with Neural Networks: Feature-based models for representation learning, ranking objectives and loss

functions and training a neural ranker under different levels of supervision.
Modeling user behavior with Neural Networks: Probabilistic graphical models, Neural click models, and modeling biases using neural network will be described.
Generating Models: The ideas on machine reading task, question answering, conversational IR, and dialogue systems will be covered.

ii. Lexical vs Semantic matching
   A. Traditional IR models estimate relevance based on lexical matches of query terms in document.
   B. Representation learning based models garner evidence of relevance from all document terms based on semantic matches with query. Both lexical and semantic matching are important and can be modelled with neural networks.

2. Learning to rank shoelace : `https://github.com/rjagerman/shoelace` [Jagerman et al., 2017]

   (a) quickrank QuickRank : `http://quickrank.isti.cnr.it` [Capannini et al., 2016]
   Capannini, G., Lucchese, C., Nardini, F. M., Orlando, S., Perego, R., and Tonellotto, N. Quality versus efficiency in document scoring with learning-to-rank models. Information Processing & Management 2016. `https://www.sciencedirect.com/science/article/abs/pii/S0306457316301248`

16.6 reference your chosen papers

/home/shane/dump/home/shane/notes2018/projects/ir-assignment-2/acmart-master/acmart.bbl

In your report you should outline the research questions and main contribution of each paper you select. You should discuss how the papers you chose are related to each other. Finally, you should formulate two new research questions in the area and discuss how you would address these.