From Particular to General: A Preliminary Case Study of Transfer Learning in Reading Comprehension

Rudolf Kadlec*, **Ondrej Bajgar*** **& Jan Kleindienst**
IBM Watson
V Parku 4, 140 00 Prague, Czech Republic
{rudolf_kadlec,obajgar,jankle}@cz.ibm.com

Abstract

In this paper we argue that transfer learning will be an important ingredient of general learning AI. We are especially interested in using data-rich domains to learn skills widely applicable in other domains. As a case study we explore transfer learning in reading comprehension. We train a neural-network-based model on two context-question-answer datasets, the Children's Book Test and its larger extension, the BookTest, and we monitor transfer to a subset of bAbI tasks. Our initial experiments show only limited transfer between these domains. However, the transferred system is still significantly better than a random baseline.

1 Introduction

Machine intelligence has had some notable successes, however often in narrow domains which are sometimes of little practical use to humans – for instance games like chess [2] or Go [22]. If aimed to build a general AI that would be able to efficiently assist humans in a wide range of settings, we would want it to have a much larger set of skills – among them would be an ability to understand human language, to perform common-sense reasoning and to be able to generalize its abilities to new situations like humans do.

If we want to achieve this goal through Machine Learning, we need data to learn from. One way to achieve wide applicability would be to provide training data for each specific task we would like the machine to perform. However it is unrealistic to obtain a sufficient amount of training data for some domains – it may for instance require expensive human annotation or all domains of application may be difficult to predict in advance – while the amount of training data in other domains is practically unlimited, (e.g. in language modelling or Cloze-style question answering).

The way to bridge this gap – and to achieve the aforementioned adaptability – is *transfer learning* [17] which would allow the system to acquire a set of skills on domains where data are abundant and then use these skills to succeed on a previously unseen domain. Despite how crucial generalization is for general AI, mainstream research keeps focusing on solving narrow tasks.

The main questions we are hence trying to address in this paper are

1. Whether we can realistically expect to be able to train models on natural-language tasks where data are abundant and transfer the learnt skills to tasks where in-domain training data are difficult to obtain. Specifically we would like to examine whether training on large-scale natural language datasets provides the model with generally applicable reasoning abilities.
2. Whether and how this ability to generalize improves with growing amount of data for the training task.

*These authors contributed equally to this work.

In this preliminary exploration we are certainly not aiming to answer these questions exhaustively; we are rather trying to show some initial results and stimulate further discussion.

As a case study we focus on the domain of text comprehension – a domain that would undoubtedly form an important sub-module of a general AI and that has lately attracted a lot of attention in the NLP community [7, 9, 11, 10, 3, 23, 6, 25, 26, 5, 4, 13, 21].

2 Case Study: Transfer Learning in Text Comprehension

To get more specific, let us now illustrate the ideas we've just outlined by small experimentation in the domain of text comprehension. Concretely, let us explore what transferable skills the successful Attention Sum Reader (AS Reader) [10] can learn from training on one of the recently popular text-comprehension tasks – the Children's Book Test (CBT) [9] – and its much larger extension – the BookTest (BT) [1] – thanks to which the model can give state-of-the-art performance on the CBT test data [1]. To evaluate the acquired skills we use the bAbI tasks [28] – artificial tasks each of which is designed to test a specific kind of reasoning.

While we use the original CBT and BookTest training data, we slightly modify the bAbI tasks.

2.1 Cloze Style bAbI Dataset

Since CBT and BookTest train the model for Cloze-style question answering, we modify the original bAbI dataset by reformulating the questions into Cloze-style. For example we translate a question "Where is John ?" to "John is in the XXXXX ."

Since our AS Reader architecture is designed to select a single word from the context document as an answer (the task of CBT and BookTest), we selected 10 bAbI tasks that fulfill this requirement out of the original 20. These tasks are: *1. single supporting fact, 2. two supporting facts, 3. three supporting facts, 4. two argument relations, 5. three argument relations, 11. basic coreference, 12. conjunction, 13. compound coreference, 14. time reasoning* and *16. basic induction*.

2.2 Experiments

Firstly we tested how our AS Reader architecture [10] can handle the tasks if trained directly on the bAbI training data for each task. Then we tested the degree of transfer from the CBT and BookTest data to the selected bAbI tasks.

In the first experiment we trained a separate instance of the AS Reader on the 10,000-example version of the bAbI training data for each of the 10 tasks we selected. On 7 of them the architecture was able to learn the task with accuracy at least 95% as is shown in Table 1. It should be noted that there are several machine learning models that perform better than the AS Reader in the 10k weakly supervised setting, e.g. [24, 29] [2], however they often need significant fine-tuning. On the other hand we trained plain AS Reader model without any modifications. Fine-tuning could further increase its performance on individual tasks however it goes directly against the idea of generality that is at the heart of this work. For comparison with state of the art we include results of DMN+ [29] in Table 1 which had the best average performance over the original 20 tasks.

Hence if given appropriate training the AS Reader is capable of the reasoning needed to solve most of the selected bAbI tasks. Now when we know that the AS Reader is powerful enough to learn the target tasks we can turn to transfer from CBT and BookTest.

The last two columns of Table 1 summarize results of the transfer learning experiments. We see that both models tranined on CBT and BT achieve much lower accuracy than the model trained directly on bAbI tasks. This is clearly indicated by mean accuracy over the selected bAbI tasks (shown in the last row of Table 1). However, one positive result is that there is some transfer between the tasks since the AS Reader trained on either CBT or BT outperforms a random baseline[3] on bAbI. Another important observation is that, at least on two of the tasks, more training data help. The AS Reader

[2]There is also a handcrafted model based on prior human analysis of the tasks [12] that solves all the task almost perfectly without any learning on training data.

[3]The random baseline selects randomly uniformly between all unique words contained in the context document.

	Model:	Random	DMN+	AS Reader		
	Train dataset / Test dataset	not trained	bAbI 10k	bAbI 10k	CBT NE+CN	BookTest 14M	Δ CBT → BT
	CBT Common Nouns	NA	NA	NA	68.8	83.7†	+14.9
	CBT Named Entities	NA	NA	NA	71.9	78.4†	+6.5
1	Single supporting fact	7.8	100.0	100.0	34.7	37.3	+2.6
2	Two supporting facts	4.4	99.7	91.9	29.7	25.8	-3.9
3	Three supporting facts	3.4	98.9	86.0	25.2	22.2	-3.0
4	Two-argument relations	10.5	100.0	100.0	29.7	50.3	+20.6
5	Three-argument relations	4.4	99.5	99.8	58.6	67.6	+9.0
11	Basic coreference	6.2	100.0	100.0	31.4	33.0	+1.6
12	Conjunction	6.7	100.0	100.0	30.7	30.4	-0.3
13	Compound coreference	5.6	100.0	100.0	32.5	33.8	+1.3
14	Time reasoning	5.0	99.8	95.0	25.5	27.6	+2.1
16	Basic induction	7.5	54.7	50.3	0.0	0.0	0.0
	bAbI mean (10 tasks)	6.2	95.3	92.3	29.2	32.5	+3.3

Table 1: Performance of the AS Reader when trained on bAbI 10k, CBT and BT dataset and then evaluated on CBT and bAbI data. The Dynamic Memory Network (DMN+) is the state-of-the-art model in weakly supervised setting on the bAbI 10k dataset. Its results are taken from [29]. The last column shows the difference between models trained on the CBT and on the BookTest. Results marked with † are from [1].

trained on BT that is 60 times larger than CBT performs 3.3% better on average with the strongest improvement on tasks 4 and 5 where the accuracy increases by 20.6 and 9.0 percent absolute. On the other hand performance in two and three supporting facts decreases. This suggests that the BookTest probably does not require this kind of reasoning.

3 Conclusion

Our results show that even though the AS Reader trained on BookTest outperforms all published models [9, 23, 6, 25, 26, 5, 4] trained on the CBT it fails to solve any of the bAbI tasks with better than 65% accuracy. Therefore the transfer is still very limited. Although larger training dataset improved the average performance on bAbI tasks it seems unlikely that we would solve any bAbI task just by further scaling the BookTest training dataset. Training on more diverse collection of datasets seems more promising than just scaling a dataset of single type.

We should also investigate how the transfer is affected by the fact that CBT and BookTest training examples all contain exactly 20 sentences in each context while bAbI contexts have varying number of sentences.

In future work it would be interesting to monitor transfer to other, more natural datasets like [20, 27, 7, 18, 19, 16, 15, 8] besides monitoring only transfer to synthetic bAbI tasks, though we still consider the later a good diagnostic tool.

Our case study was performed in a purely supervised-learning setting. When we would like to extend it to a reinforcement learning setting the tasks accompanying *A Roadmap towards Machine Intelligence* [14] might replace bAbI tasks used in our case study. However, it remains unclear where to find an environment with real-world properties that would be analogous to BookTest for such experiments.

We think this brief paper should prompt the community to go beyond considering the various tasks being studied as entirely separate and to start thinking more about links between them, in particular the transfer of skills, for we see this as prerequisite for building a versatile general AI.

References

[1] Ondřej Bajgar, Rudolf Kadlec, and Jan Kleindienst. Embracing data abundance: BookTest Dataset for Reading Comprehension. *arXiv preprint arXiv:1610.00956*, 2016.

[2] Murray Campbell, A Joseph Hoane, and Feng-hsiung Hsu. Deep blue. *Artificial intelligence*, 134(1):57–83, 2002.

[3] Danqi Chen, Jason Bolton, and Christopher D. Manning. A Thorough Examination of the CNN / Daily Mail Reading Comprehension Task. In *Association for Computational Linguistics (ACL)*, 2016.

[4] Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu, and Guoping Hu. Attention-over-Attention Neural Networks for Reading Comprehension. 2016.

[5] Yiming Cui, Ting Liu, Zhipeng Chen, Shijin Wang, and Guoping Hu. Consensus Attention-based Neural Networks for Chinese Reading Comprehension. 2016.

[6] Bhuwan Dhingra, Hanxiao Liu, William W. Cohen, and Ruslan Salakhutdinov. Gated-Attention Readers for Text Comprehension. 2016.

[7] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1684–1692, 2015.

[8] Daniel Hewlett, Alexandre Lacoste, Llion Jones, Illia Polosukhin, Andrew Fandrianto, Jay Han, Matthew Kelcey, and David Berthelot. WIKI READING : A Novel Large-scale Language Understanding Task over Wikipedia. *Acl 2016*, pages 1535–1545, 2016.

[9] Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. The goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301*, 2015.

[10] Rudolf Kadlec, Martin Schmid, Ondřej Bajgar, and Jan Kleindienst. Neural Text Understanding with Attention Sum Reader. *Proceedings of ACL*, 2016.

[11] Sosuke Kobayashi, Ran Tian, Naoaki Okazaki, and Kentaro Inui. Dynamic Entity Representation with Max-pooling Improves Machine Reading. *Proceedings of the North American Chapter of the Association for Computational Linguistics and Human Language Technologies (NAACL-HLT)*, 2016.

[12] Moontae Lee, Xiaodong He, Wen-tau Yih, Jianfeng Gao, Li Deng, and Paul Smolensky. Reasoning in Vector Space: An Exploratory Study of Question Answering. *ICLR*, 2016.

[13] Peng Li, Wei Li, Zhengyan He, Xuguang Wang, Ying Cao, Jie Zhou, and Wei Xu. Dataset and Neural Recurrent Sequence Labeling Model for Open-Domain Factoid Question Answering. 2016.

[14] Tomas Mikolov, Armand Joulin, and Marco Baroni. A Roadmap towards Machine Intelligence. pages 1–39, 2015.

[15] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A Corpus and Evaluation Framework for Deeper Understanding of Commonsense Stories. *Proceedings of NAACL*, 2016.

[16] Takeshi Onishi, Hai Wang, Mohit Bansal, Kevin Gimpel, and David McAllester. Who did What: A Large-Scale Person-Centered Cloze Dataset. *EMNLP*, 2016.

[17] Sinno Jialin Pan and Qiang Yang. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, oct 2010.

[18] Denis Paperno, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fern. The LAMBADA dataset : Word prediction requiring a broad discourse context. *Proceedings of ACL*, 2016.

[19] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ Questions for Machine Comprehension of Text. (ii), 2016.

[20] Matthew Richardson, Christopher J C Burges, and Erin Renshaw. MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text. *Empirical Methods in Natural Language Processing (EMNLP)*, pages 193–203, 2013.

[21] Yelong Shen, Po-Sen Huang, Jianfeng Gao, and Weizhu Chen. ReasoNet: Learning to Stop Reading in Machine Comprehension. 2016.

[22] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[23] Alessandro Sordoni, Phillip Bachman, and Yoshua Bengio. Iterative Alternating Neural Attention for Machine Reading. 2016.

[24] Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. End-To-End Memory Networks. pages 1–11, 2015.

[25] Adam Trischler, Zheng Ye, Xingdi Yuan, and Kaheer Suleman. Natural Language Comprehension with the EpiReader. 2016.

[26] Dirk Weissenborn. Separating Answers from Queries for Neural Reading Comprehension. 2016.

[27] Jason Weston, Antoine Bordes, Sumit Chopra, and Tomas Mikolov. Towards AI-Complete Question Answering: A Set of Prerequis ite Toy Tasks. 2014.

[28] Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart Van Merri, Armand Joulin, and Tomas Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks. 2016.

[29] Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic Memory Networks for Visual and Textual Question Answering. *ICML*, 2016.

A Method

Here we give a more detailed description of the method we used to arrive to our results. We highlight only facts particular to this experiment. A more detailed general description of training the AS Reader is given in [10].

The results given for AS Reader trained on bAbI are each for a single model with 64 hidden units in each direction of the GRU context encoder and embedding dimension 32 trained on the 10k training data provided with that particular task.

The results for AS Reader trained on the CBT and the BookTest are for a greedy ensemble consisting of 4 models whose predictions were simply averaged. The models and ensemble were all validated on the validation set corresponding to the training dataset. The performance on the bAbI tasks oscillated notably during training however the averaging does somewhat mitigate this to get more representative numbers.