

Impact of Training Dataset Size on Neural Answer Selection Models

Trond Linjordet and Krisztian Balog

University of Stavanger, Stavanger, Norway
 {trond.linjordet,krisztian.balog}@uis.no

Abstract. It is held as a truism that deep neural networks require large datasets to train effective models. However, large datasets, especially with high-quality labels, can be expensive to obtain. This study sets out to investigate (i) how large a dataset must be to train well-performing models, and (ii) what impact can be shown from fractional changes to the dataset size. A practical method to investigate these questions is to train a collection of deep neural answer selection models using fractional subsets of varying sizes of an initial dataset. We observe that dataset size has a conspicuous lack of effect on the training of some of these models, bringing the underlying algorithms into question.

1 Introduction

The impressive performance improvements brought by deep learning applied to certain domains—computer vision, audio speech-to-text, and natural language processing (NLP) [8, 13]—has motivated a great deal of interest to apply deep learning to other domains as well, including information retrieval (IR). However, the performance improvements from deep learning relative to conventional machine learning approaches have depended on increased computational power, larger datasets to learn from, and some developments on the algorithm and architecture level. Of these three factors, large datasets may represent the least tractable challenge faced by those who would apply deep learning to new domains. Quality training data, especially for supervised learning, requires intensive effort to prepare for the actual learning process.

A category of tasks at the intersection of the fields of IR and NLP, question answering (QA) means returning a correct answer sentence in response to a grammatically well-formed, natural language question. In the present work, a specific variant of the QA task is considered, namely answer selection, the task of matching single-sentence questions with single-sentence answers. The answer selection task is simply: given a question and a set of candidate answers, select the correct answer. This task has recently been investigated as a neural IR problem [9, 11].

This paper considers a practical approach to investigating the impact of training dataset size on the performance that can be achieved with various deep neural architectures for the task of answer selection. The approach taken by this paper can be summarized as follows: A pre-existing implementation of neural architectures for answer selection is investigated by truncating the training data to fractions of the original training dataset, to quantify the differences in performance by trained models given different amounts of training data from the same distribution.

One of the surprising experimental findings of this paper is that most models do not exhibit the expected behavior in terms of performance improvement in response to increased training dataset size.

2 Related Work

The impact of the size of training datasets has been investigated for convolutional neural networks (CNNs) trained on image data [2, 15]. In the latter work, it was observed that model performance improves roughly logarithmically as a function of increased training data. The idea of a logarithmic relationship between performance and dataset size was further corroborated empirically by Hestness, et al. [5].

An investigation of the generalization problem in deep neural networks, i.e., the discrepancy between the performance of a trained model on training data and test data, shows that the deep neural models have a representational capacity that enables “memorization” of training data: Zhang et al. [19] show the order-of-magnitude relationship between training dataset size (sample size), input data dimensionality, and the depth of a network with sufficient parameters to fully memorize the training dataset. They report a theorem with proof such that for any finite n -sized sample of d -dimensional inputs, there exists a two-layer ReLU neural network with $2n + d$ weights that can represent any function on the sample. As a corollary, this finding extends from this hypothetical shallow and wide network to a narrow and deep network where the relationship between sample size and number of parameters is conserved. This may not be how deep neural networks learn in practice [1], but the theorem indicates the challenge that finite datasets may present to generalization in deep learning models.

3 Approach

The approach presented in this paper is practical in that dataset size was manipulated and the effects were evaluated using a pre-existing implementation of multiple neural IR models with a single original dataset. Specifically, this paper presents work on the MatchZoo project¹ [3], where a number of deep neural architectures for text matching have been implemented. Here, answer selection is considered as a form of question answering, where the question text is matched with the text of the correct answer. The original dataset used for training, validation, and testing, was the canonical WikiQA dataset [18]. The performance of the implemented models on a given dataset was characterized in terms of Mean Average Precision (MAP) over the candidate answer rankings for each question in that dataset.

3.1 Data preparation

The training dataset was filtered to provide the models being trained with meaningfully labelled training data. The filtering rule was simply to omit any question and its associated set of candidate answer sentences if the set of candidate answer sentences did not include both true and false candidates.

¹ <https://github.com/faneshion/matchzoo>

	Training						
	10%	25%	50%	75%	100%	Valid.	Test
#Questions	78	209	414	639	857	122	237
#QA pairs	823	2256	4321	6537	8651	1126	2341

Table 1. Summary of datasets.

Table 1 summarizes the datasets used in the training of the various models. Note that the same validation and test sets were used throughout, while the training dataset used was systematically varied between the original (filtered) training set (100%), and various partial training sets truncated to 10%, 25%, 50%, and 75% of the original (filtered) training set. These partial training sets were made by randomly sampling (without replacement) on the questions in the original (filtered) training set. Each selected question was then included in the respective partial training set along with all corresponding candidate answers and their labels. The percentages thus represent the probability for each question to be included in each partial dataset. However, once the random sub-sampling was accomplished, these partial training sets were fixed. Each of the models was then trained five times independently on each dataset size.

3.2 Models

A number of models were able to train and perform nominally with the code provided by the MatchZoo project [3]. The models investigated in the present paper comprised:

- **Deep Structured Semantic Model (DSSM)** [7], which extends latent semantic analysis with deep architectures; a seminal work on neural IR.
- **Convolutional Deep Structured Semantic Model (CDSSM)** [14], which uses a convolutional neural network (CNN) to extend DSSM with contextual information at the word n-gram level.
- **Architecture-I (ARC-I)** [6], an extension of CDSSM whereby siamese CNNs learn to represent two sentences, deferring matching of sentence pairs to a final multi-layer perceptron (MLP).
- **Architecture-II (ARC-II)** [6], an alternative to ARC-I where sentences interact by 1D convolution before proceeding through a 2D CNN component which is purported to learn both the representation of the individual sentences, as well as the structure of their relationship. Again, matching of the representations is determined by a final MLP.
- **Multiple positional sentence representations (MV-LSTM)** [16], follows the aforementioned models by capturing local information on multiple levels of granularity within a sentence, using bidirectional long short-term memory networks (bi-LSTMs) to represent input sentences, modeling interactions with a similarity function (tensor layer), and aggregating interactions with k -Max Pooling before a final MLP to match the obtained representations.
- **Deep relevance matching model (DRMM)** [4], distinguishes relevance matching from semantic matching, using pre-trained neural embeddings of terms and build-

ing up fixed-length matching histograms from variable-length local interactions between each query term and document. Each query term matching histogram is passed through a matching MLP, and the overall score is aggregated with a query term gate—a softmax function over all terms in that query.

- **Attention-based neural matching model (aNMM)** [17], which follows a similar structure as ARC-II, except instead of position-shared weighting, aNMM has adopted a value-shared weighting scheme “to learn the importance of different levels of matching signals,” and incorporated a query term gate similar to that used in DRMM.
- **Combined local and distributed representations (DUET)** [10], which aims to combine local exact matching with embeddings of query-document pairs in semantic space. This relevance matching is enabled by both the local and distributed models, hence a “duet” of two parallel neural models. The final matching score is simply the sum of the two outputs.
- **MatchPyramid** [12], which uses a matching matrix layer to evaluate pairwise term similarity between two texts, followed by 2D convolutional and pooling layers, with a final matching MLP.
- **DRMM_TKS** [3], which is a variant of DRMM provided by the MatchZoo project, for matching short texts. The architecture is simply described by “Specifically, the matching histogram is replaced by a top-k max pooling layer and the remaining parts are fixed.”

Some of these models are motivated more by ad hoc search and document retrieval, whereas others were developed specifically for answer selection and the similar task of sentence completion. However, the commonality is that all the models are designed for text matching.

4 Experiment/Results

The following experimental results show the effect of varying training set size.

4.1 Final performance of trained models

Figure 1 presents the performance on the test dataset of the different models after training for 400 iterations on datasets of various sizes. These figures show that aside from the DSSM, CDSSM, and possibly MatchPyramid models, some improvement does appear to happen with greater training dataset sizes. However, by having an order of magnitude more training data (10% to 100%), only three models, CDSSM, ARC-II, and DRMM_TKS, achieve a relative improvement above 20%. Four more models, DSSM, MV-LSTM, aNMM, and DUET manage to achieve a relative improvement above 10%. For DRMM, performance even slightly decreases (by 1%). The relative improvements after having doubled (25% to 50%), tripled (25% to 75%), or quadrupled (25% to 100%) the training data size are similarly moderate for most models. Specifically, after doubling, only CDSSM and aNMM showed relative improvement above 10%, and with tripling and quadrupling, only DSSM, CDSSM, ARC-II, and aNMM showed relative improvement above 10%.

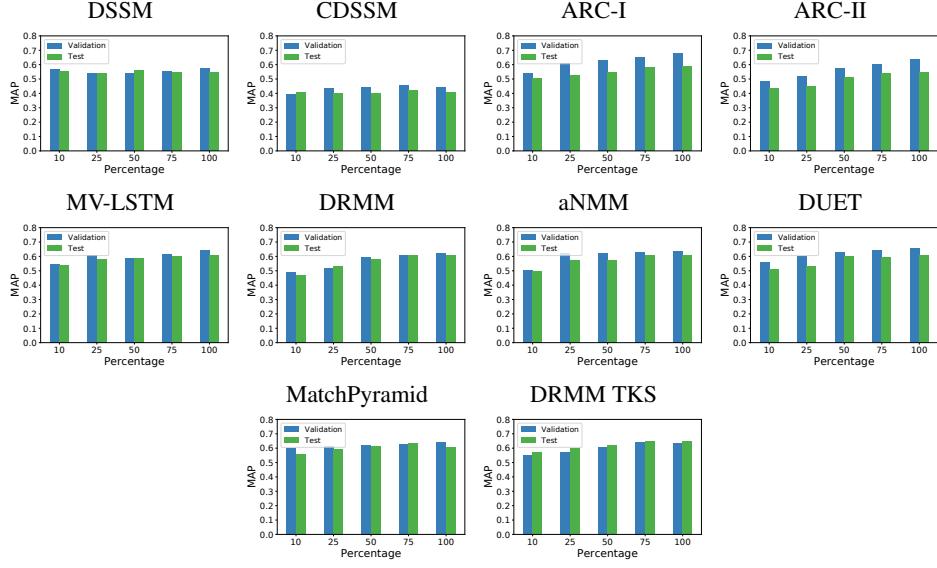


Fig. 1. Performance (as measured by mean average precision) on the validation (blue) and test (green) datasets with different training dataset sizes.

4.2 Model training histories

Figure 2 illustrates the relationship, for each model, between the size of the training dataset and performance improvements over the course of training. We can see that most models either reach a plateau or approximately monotonically increase on the training set (shown in blue curves in Fig. 2) within the recorded training history. There are, however, a few exceptions, namely DRMM and aNMM, which do not exhibit this desired behavior. Another outlier is DRMM TKS, which improves at a drastically slow rate. It is also worth pointing out that the models DSSM and MatchPyramid overfit very quickly. This may suggest a memorization effect.

Looking at the MAP scores on the validation set (shown in blue curves in Fig. 2), we see a discrepancy from expected behavior. The desired behavior would be that these follow the same monotonically increasing trend as the red lines, with the gap between the two lines decreasing as the amount of training data increases. Most of the models, however, do not behave like that. The validation lines plateau out quickly for most models, or even degrade (DRMM, aNMM).

5 Conclusions

We have briefly looked at the effects of dataset size on the neural IR task of answer selection for a number of deep architectures. The consequences of reducing the available training data logarithmically (10% versus 100%) are discernible, and indicate primarily a failure to generalize. This can be seen from the discrepancy between performance improvement on training data, compared to the modest improvements on validation

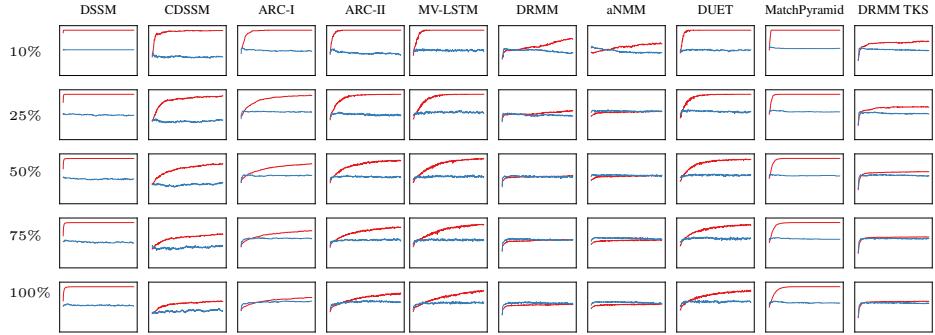


Fig. 2. Training histories for various models (columns) with varying training dataset size (rows). The red and blue lines correspond to performance for training and validation datasets, respectively. Performance is measured in terms of MAP, and indicated with respect to the y -axes, which range from 0 to 1. The x -axes indicate the number of training iterations (epochs), and range from 0 to 399. The x - and y -axes are identically scaled in each of the sub-plots.

data. Note that these findings are based on one particular implementation, and the inner workings of the implementation were not rigorously analyzed and verified, but were assumed to correctly enact the cited algorithms.

These findings show that when choosing algorithms and strategies in regard to data volume, there are factors which must be considered beyond the reported benchmarks of fully trained models. The actual performance of the models during different stages of training, relative to different scales of training data, must be considered to discover any unexpected trends.

Furthermore, performance on validation sets is clearly a very important basis for comparison, to gain an intuition about how fast models generalize from different volumes of training data, and with different numbers of training epochs.

Future work may consist of a deeper investigation into the reproducibility of answer selection state-of-the-art results, as well as into quantifying the relationship between training dataset size and the impact of diverse neural models on generalizability.

Bibliography

- [1] D. Arpit, S. Jastrz̄bski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio, and S. Lacoste-Julien. A closer look at memorization in deep networks. In *Proc. of ICML '17*, pages 233–242, 2017.
- [2] J. Cho, K. Lee, E. Shin, G. Choy, and S. Do. How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? *ArXiv e-prints*, 2015.
- [3] Y. Fan, L. Pang, J. Hou, J. Guo, Y. Lan, and X. Cheng. MatchZoo: A Toolkit for Deep Text Matching. *ArXiv e-prints*, July 2017.
- [4] J. Guo, Y. Fan, Q. Ai, and W. B. Croft. A deep relevance matching model for ad-hoc retrieval. In *Proc. of CIKM '16*, pages 55–64, 2016.
- [5] J. Hestness, S. Narang, N. Ardalani, G. Diamos, H. Jun, H. Kianinejad, M. M. A. Patwary, Y. Yang, and Y. Zhou. Deep Learning Scaling is Predictable, Empirically. *ArXiv e-prints*, Dec. 2017.
- [6] B. Hu, Z. Lu, H. Li, and Q. Chen. Convolutional neural network architectures for matching natural language sentences. In *Proc. of NIPS '14*, pages 2042–2050, 2014.
- [7] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck. Learning deep structured semantic models for web search using clickthrough data. In *Proc. of CIKM '13*, pages 2333–2338, 2013.
- [8] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436, may 2015.

- [9] B. Mitra and N. Craswell. Neural Models for Information Retrieval. *ArXiv e-prints*, May 2017.
- [10] B. Mitra, F. Diaz, and N. Craswell. Learning to match using local and distributed representations of text for web search. In *Proc. of WWW '17*, pages 1291–1299, 2017.
- [11] K. D. Onal, Y. Zhang, I. S. Altingovde, M. M. Rahman, P. Karagoz, A. Braylan, B. Dang, H.-L. Chang, H. Kim, Q. McNamara, A. Angert, E. Banner, V. Khetan, T. McDonnell, A. T. Nguyen, D. Xu, B. C. Wallace, M. de Rijke, and M. Lease. Neural information retrieval: at the end of the early years. *Information Retrieval Journal*, Nov 2017.
- [12] L. Pang, Y. Lan, J. Guo, J. Xu, S. Wan, and X. Cheng. Text matching as image recognition. In *Proc. of AAAI '16*, pages 2793–2799, 2016.
- [13] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85 – 117, 2015.
- [14] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil. Learning semantic representations using convolutional neural networks for web search. In *Proc. of WWW '14*, pages 373–374, 2014.
- [15] C. Sun, A. Shrivastava, S. Singh, and A. Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proc. of ICCV '17*, pages 843–852, 2017.
- [16] S. Wan, Y. Lan, J. Guo, J. Xu, L. Pang, and X. Cheng. A deep architecture for semantic matching with multiple positional sentence representations. In *Proc. of AAAI '16*, pages 2835–2841, 2016.
- [17] L. Yang, Q. Ai, J. Guo, and W. B. Croft. anmm: Ranking short answer texts with attention-based neural matching model. In *Proc. of CIKM '16*, pages 287–296, 2016.
- [18] Y. Yang, S. W.-t. Yih, and C. Meek. Wikiqa: A challenge dataset for open-domain question answering. In *Proc. of EMNLP '15*, pages 2013–2018, 2015.
- [19] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In *Proc. of ICLR '17*, pages 1–2, 2017.