How much data is enough? Predicting accuracy on large datasets from smaller pilot data

Mark Johnson, Peter Anderson,
Mark Dras, Mark Steedman

Macquarie University
Sydney, Australia

July 13, 2018

Outline

Introduction

Empirical models of accuracy vs training data size

Accuracy extrapolation task

Conclusions and future work

ML as an engineering discipline

- A mature engineering discipline should be able to predict the cost of a project before it starts
- Collecting/producing training data is typically the most expensive part of an ML or NLP project
- We usually have only the vaguest idea of how accuracy is related to training data size and quality
  - More data produces better accuracy
  - Higher quality data (closer domain, less noise) produces better accuracy
  - But we usually have *no idea how much data or what quality of data is required to achieve a given performance goal*
- *Imagine if engineers designed bridges the way we build systems!*

See *statistical power analysis* for experimental design, e.g., Cohen (1992)

Goals of this research project

- Given desiderata (accuracy, speed, computational and data resource pricing, etc.) for an ML/NLP system, design for a system that meets these.
- Example: *design a semantic parser for a target application domain that achieves 95% accuracy across a given range of queries.*
  - What hardware/software should I use?
  - *How many labelled training examples do I need?*
- Idea: *Extrapolate performance from small pilot data to predict performance on much larger data*

What this paper contributes

- Studies different methods for predicting accuracy on a full dataset from results on a small pilot dataset
- We propose new *accuracy extrapolation task*, provide results for the 9 extrapolation methods on 8 text corpora
    - Uses the *fastText document classifier* and corpora (Joulin et al., 2016)
- Investigates *three extrapolation models* and *three item weighting functions* for predicting accuracy as a function of training data size
    - Easily inverted to estimate training size required to achieve a target accuracy
- Highlights the importance of *hyperparameter tuning* and *item weighting* in extrapolation

Outline

Overview

- *Extrapolation models* of how error e ($= 1 - accuracy$) depends on training data size n
  - *Power law:* $\hat{e}(n) = bn^c$
  - *Inverse square-root:* $\hat{e}(n) = a + bn^{-1/2}$
  - *Biased power law:* $\hat{e}(n) = a + bn^c$
- Extrapolation model estimated from multiple runs using *weighted least squares regression*
  - Model trained on *different-sized subsets of pilot data*
  - Same test set is used to evaluate each run
  - The evaluation of each model training/test run is a training data point for extrapolation model
- *Weighting functions* for least squares regression
  - *constant weight* (1)
  - *linear weight* (n)
  - *binomial weight* ($n/e(1-e)$)

See e.g., Haussler et al. (1996); Mukherjee et al. (2003); Figueroa et al. (2012); Beleites et al. (2013); Hajian-Tilaki (2014); Cho et al. (2015); Sun et al. (2017); Barone et al. (2017); Hestness et al. (2017)

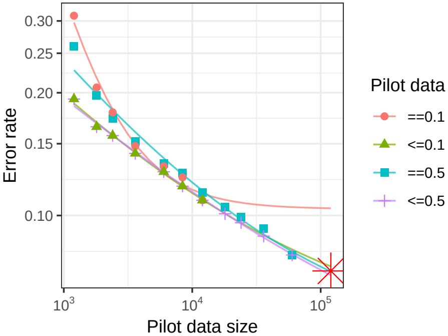
Outline

Accuracy extrapolation task

Corpus	Labels	Train (K)	Test (K)
Development			
ag_news	4	120	7.6
dbpedia	14	560	70
amazon_review_full	5	3,000	650
yelp_review_polarity	2	560	38
Evaluation			
amazon_review_polarity	2	3,600	400
sogou_news	5	450	60
yahoo_answers	10	1,400	60
yelp_review_full	5	650	50

- FastText document classifier & data
    - 4 development corpora
    - 4 evaluation corpora
    - Joulin et al. (2016)'s train/test division
- Pilot data is 0.5 or 0.1 of train data
- Goal: *use pilot data to predict test accuracy when trained on full train data*

Extrapolation on ag_news corpus



- Extrapolation with *biased power-law model* ($\hat{e}(n) = a + bn^c$) and *binomial weights* ($^n/_{e(1-e)}$)
- Extrapolation from 0.5 training data is generally good
- Extrapolation from 0.1 training data is poor unless *hyperparameters are optimised at each subset of pilot data*

Relative residuals ($\hat{e}/e - 1$) on dev corpora

RMS relative residuals on test corpora

Pilot data	amazon review polarity	sogou news	yahoo answers	yelp review full	**Overall**
$= 0.1$	0.1016	0.2752	0.0519	0.0496	0.1510
≤ 0.1	**0.0209**	**0.1900**	**0.0264**	**0.0406**	**0.0986**
$= 0.5$	0.0338	0.0438	0.0254	0.0160	0.0315
≤ 0.5	**0.0049**	**0.0390**	**0.0053**	**0.0046**	**0.0200**

- Based on dev corpora results, use:
  - biased power law model ($\hat{e}(n) = a + bn^c$)
  - binomial item weights ($^{n}/_{e(1-e)}$)
- Evaluate extrapolations with RMS of *relative residuals* ($\hat{e}/e - 1$)
- Larger pilot data $\Rightarrow$ smaller extrapolation error
- Optimise hyperparameters at each pilot subset $\Rightarrow$ smaller extrapolation error

Outline

Conclusions and future work

- The field need methods for predicting how much training data a system needs to achieve a target performance
- We introduced an *extrapolation task* for predicting a classifier's accuracy on a large dataset from a small pilot dataset
- Highlight the importance of *hyperparameter tuning* and *item weighting*
- Future work: *extrapolation methods that don't require expensive hyperparameter optimisation*

We are recruiting PhD students and Postdocs!

Centre for Research in AI and Language (CRAIL)

Macquarie University

Parsing, Dialog, Deep Unsupervised Learning, Language in Context

Vision and Language, Language for Robot Control

- We are recruiting top **PhD Students** and **Postdoc Researchers**
  - With *generous pay* and *top-up scholarships to $41K tax-free*
- Send CV and sample papers to **Mark.Johnson@MQ.edu.au**

References

Barone, A. V. M., Haddow, B., Germann, U., and Sennrich, R. (2017). Regularization techniques for fine-tuning in neural machine translation. *CoRR*, abs/1707.09920.

Beleites, C., Neugebauer, U., Bocklitz, T., Krafft, C., and Popp, J. (2013). Sample size planning for classification models. *Analytica chimica acta*, 760:25–33.

Cho, J., Lee, K., Shin, E., Choy, G., and Do, S. (2015). How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? *arXiv:1511.06348*.

Cohen, J. (1992). A power primer. *Psychological bulletin*, 112(1):155.

Figueroa, R. L., Zeng-Treitler, Q., Kandula, S., and Ngo, L. H. (2012). Predicting sample size required for classification performance. *BMC medical informatics and decision making*, 12(1):8.

Hajian-Tilaki, K. (2014). Sample size estimation in diagnostic test studies of biomedical informatics. *Journal of biomedical informatics*, 48:193–204.

Haussler, D., Kearns, M., Seung, H. S., and Tishby, N. (1996). Rigorous learning curve bounds from statistical mechanics. *Machine Learning*, 25(2).

Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M. M. A., Yang, Y., and Zhou, Y. (2017). Deep learning scaling is predictable, empirically. *arXiv:1712.00409*.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv:1607.01759*.

Mukherjee, S., Tamayo, P., Rogers, S., Rifkin, R., Engle, A., Campbell, C., Golub, T. R., and Mesirov, J. P. (2003). Estimating dataset size requirements for classifying DNA microarray data. *Journal of computational biology*, 10(2):119–142.

Sun, C., Shrivastava, A., Singh, S., and Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. *arXiv:1707.02968*.