Deep Expertise: the Need for Fine-tuning at Scale

Deyan Ginev

FAU Erlangen-Nuremberg

Recent News

This post contains some follow-up thoughts to our preprint introducing ”Scientific Statement Classification”, a supervised learning task over arXiv. We released

- a dataset of 10.5 million annotated paragraphs and
- a showcase of a BiLSTM encoder-decoder baseline (0.91 F1 score)

which may be helpful for building an intuition for the data and models discussed here.

There are two brief observations I would like to make today, with some numerical examples.

Specialized domains require model expertise

arXiv contains a specialized domain of discourse. It is largely a submission forum for scientific articles that end up as conference or journal proceedings. When approached by a human reader, there is a very clear regularity to document size, shape of the expositions, the subset of English vocabulary used[1]. There is certainly variation between the different STEM fields, and between the usual article and alternative types of monographs (books, theses, reports and supplements, among others). That adds complexity and richness to modeling scientific discourse, but it does not help bridge the gap to other genres. One could argue that a conference article is a lot closer to the text of a PhD thesis than it is to a great work of fiction, or newswire text.

It is very difficult to actually defend any of these claims rigorously, if one wants to do so with generality. We could pick a variety of metrics and measure, but each of those choices is informative only for models where these metrics are of direct relevance. So instead, I will make claims for specific data+model pairs we worked with.

Our dataset comprises 11.5 billion tokens, and a vocabulary of just over 1 million words (> 5 frequency), uncased. For our baselines, we decided to use GloVe as the grounding layer, and we chose a 300 dimensional embedding, trying to stay close to the original GloVe paper.

In order to check whether we need our own embeddings, we compare against two of the original 300d-uncased GloVe models. We test the three embedding sets by performing logistic regression on our task of classifying 10.4 million paragraphs into 13 statement categories.

name	vocabulary	logreg F1 score	source
glove.6B.300d	0.4m	0.63	Wikipedia+Gigaword
glove.42B.300d	1.9m	0.64	Common Crawl
glove.arxmliv.11B.300d	1.0m	0.77	arXMLiv 08.2018
glove.arxmliv.11B.600d	1.0m	0.79	-

---

[1] :There are foreign language works on arXiv as well, we excluded them from our data for this experiment.

Observation 1: Going from 6B to 42B tokens and 0.4m to 1.9m vocabulary only contributed +0.01 F1 score. Training our own GloVe embeddings over the 11B token arXiv data had a tenfold impact, +0.14 F1 score increment.

As an aside, a brief test with 600-dimensional embeddings showed an additional improvement, but for now we remain with the 300d models as it makes comparisons with other experiments more direct[2].

Given that a fine-tuned model on an expert domain shows a clear advantage to something as all-encompassing as Common Crawl, I have taken the liberty of referring to this upgraded delta as "expertise", somewhat informally. And of course, there isn't any aspect of it that deserves to be called "deep", aside from its utility as a grounding layer of deep-learning models. Thanks for clicking on the catchy title!

Data is king, if one can generalize

Our showcase baseline model improved another +0.14 over logistic regression, to a 0.91 F1 score. But does it really need 10 million samples to get there? We retrained that BiLSTM model with per-class maximum sample count caps, starting small and going up a magnitude at a time. We then test each model on the full 2.1 million unseen test samples.

class cap	total training set	validation accuracy	bilstm F1 score
1,000	12,987	0.741	0.65
10,000	121,240	0.828	0.83
100,000	1,068,368	0.882	0.86
n/a	8,354,000	0.903	0.91

And indeed, "data is king". Each magnitude of additional training data offers visible improvement to our BiLSTM model, +0.03 F1 score or more. We also see a great case for generalization. The BiLSTM trained on 0.1 million samples already outperforms a logistic regression trained on 8 million. Even more impressive, the BiLSTM trained on just a thousand samples per class outperforms a logistic regression using an off-the-shelf GloVe embedding and the full 8 million-strong training data.

Observation 2: Expertise plus generalization on small data can beat a less capable model at scale. Once all else is equal, data is king.

Discussion

Taken together, these observations suggest that language model "expertise" (commonly "fine-tuning"), ability to generalize and the scale of the underlying training data are independently valuable.

If we were to invert that, it means poor F1 scores may be due to any of:

- poor fit between pre-trained language model and domain of application
- failure to generalize to the specific posed task by a specific model (e.g. logistic regression on language classification)
- data starvation

Let me know if these findings resonate with your own experience, or if some of the claims seem shaky and/or failing to separate individual effects. The reasons for these measurements were to try and understand better both our domain-specific data, but also to get an empirical handle of the modeling toolkits we chose in our specific experimental setting.

---

[2] :600d is also somewhat heavy to use in practice (e.g. most 2019 laptops would struggle with the RAM allocation