Weak Supervision in the Age of Transfer Learning for NLP

Fei Fang, Zihan Xie
{feifang, xiezihan}@stanford.edu

Problem

Challenge: obtaining high-quality human-labeled training data.

Approach: overcome data scarcity in Natural Language Processing (NLP) through weakly-supervised data generation, where training data has programmatically-generated labels. Compensate for label fidelity with the scalability of data collection.

Machine learning task: sentiment classification in NLP.

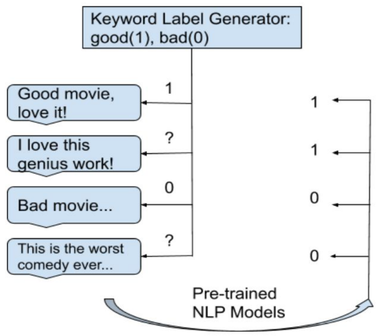
Questions to be answered:

1. Can weak supervision outperform models trained with hand-curated data?
2. With transfer learning, what is the effect of such weak supervision on pre-trained models?

Approach: Weak Label Generation

Different approaches to generating weak labels using **heuristic functions:**

➔ Baseline: a single rule-based labeling function using keyword detection



➔ Snorkel: multiple labeling functions, combined with an unsupervised generative model [2]

➔ Automatically assigns weights to each labeling function; the final label is a weighted average.

➔ Sources of labeling functions:

- Rule-based sentiment classifier (*not* trained on movie reviews)
- Keyword detection using multinomial event model
- Regex: "10 out of 10"

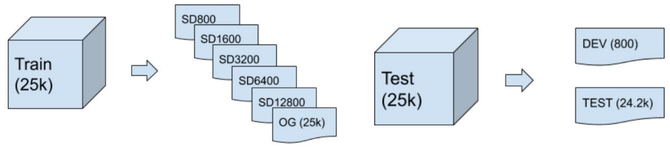
➔ Outputs: **probabilistic label** ("soft label")

- We filtered out soft labels that are close to 0.5.

Data Collection & Preprocessing

We used the IMDB movie review dataset [1].

- 25k training samples and 25k test samples, both splitted equally between 12.5k positive and negative movie reviews;
- sampled to OG (ideal scenario), SD (scarce data scenario), DEV, and TEST



Weak Label Data

We generated two weakly-labeled datasets, which contain texts taken from the original training set, but with programmatically generated labels:

- WD-N (Naive): generated with keyword-detecting labeler
  - Manually-maintained list of pos/neg keywords
- WD-S (Snorkel): generated by Snorkel [2]
  - Weighted average of several labeling functions

Dataset	WD-N	WD-S
Coverage	82.6%	94.3%
ROC-AUC*	0.75	0.88
Precision*	0.74	N/A**
Recall*	0.74	N/A**

Table 1: Statistics of weakly-generated labels compared with gold train labels

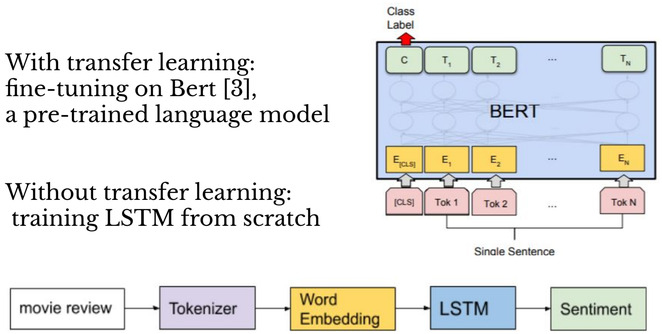
* Calculated using the gold labels in OG

** Precision and recall are not applicable to WD-S, since its labels are probabilistic.

Models & Experiments

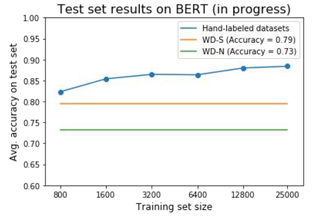
We performed two sets of training experiments on each of the following datasets described above.

1. With transfer learning: fine-tuning on Bert [3], a pre-trained language model
2. Without transfer learning: training LSTM from scratch



Results

Below are the results we have obtained on the test set:



Train set	Test AUC
WD-S	0.88
SD800	0.91
OG	0.95

Table 2: additional results on BERT

* AUC for BERT is under calculation, so plotted with accuracy

Discussion

Observations

(i) Without transfer learning, weak supervision helped achieve a 0.10 increase in AUC for LSTM models.

(ii) With transfer learning, the performance of BERT on SD800 surpassed the performance of LSTM on SD12800. In the meantime, weak supervision did not help achieve a further boost in performance.

Conclusions and hypotheses

(i) Weakly-supervised data generation would be useful when transfer learning is not available, e.g. when access to compute is limited.

(ii) BERT is an extremely effective learner that can train with limited data. Our setup of weak supervision did not help improve BERT performance in our experiments. Below are two of our hypotheses:

1. BERT is too effective at picking up signals from the training set, including noise
2. The size of our weakly-labeled training set might be too small to contain enough useful signals to BERT.

Future work: Validate hypotheses above, by collecting unlabeled movie reviews from other sources or generate synthetic movie review data, and creating a significantly larger (size ~250k) weakly-labeled dataset for BERT.

References

[1] "Learning Word Vectors for Sentiment Analysis." Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011).

[2] "Snorkel: Rapid Training Data Creation with Weak Supervision." Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, Christopher Re.

[3] "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova.