Training Compact Models for Low Resource Entity Tagging using Pre-trained Language Models

Peter Izsak, Shira Guskin, Moshe Wasserblat
Intel AI Lab

Motivation

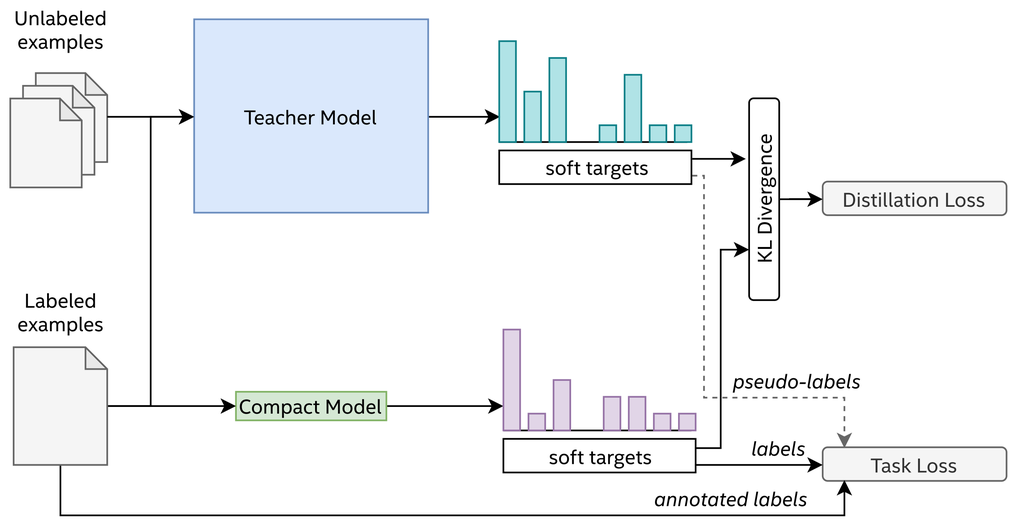
- Low-resource named entity recognition (NER) is a challenging Natural Language Processing (NLP) task [1] yet widely used for Information Extraction
- In most cases there is a lot of unlabeled data and almost no labeled
- Poor model performance can be fixed by:
  - ! Tagging more data
  - ! With expert DS/Linguist/Annotator
  - ! Sometimes cannot be fixed because of operational constraints
- Pre-trained contextual Language Models (LM) were shown to improve many NLP tasks
- Pre-trained LMs are also good for training models with scarcely labeled/low resource data
- Pre-trained LMs impose:
  - ! Heavy memory
  - ! Heavy compute
  - ! A challenge to deploy in production

Approach

1. Knowledge Distillation [2]
   - Teacher-Student training setup
   - Pre-trained LM teacher (BERT)
   - Compact student (~3M params.)
2. Pseudo-labeling [3]
   - Utilize abundance of unlabeled data
   - Large model generates *pseudo* entities for the compact model

Work-in-progress

Training Setup



$$L_{task} = \begin{cases} \text{CrossEntropy}(\hat{y}, y) & labeled\ example \\ \text{CrossEntropy}(\hat{y}, \hat{y}_{teacher}) & unlabeled\ example \end{cases}$$

$$L_{distillation} = \text{KL}(logits_{teacher} || logits_{compact})$$

$$Loss = \alpha \cdot L_{task} + \beta \cdot L_{distillation}, \quad \alpha + \beta = 1.0$$

Experimental Setup

<u>Models</u>

- Teacher – BERT [4] (110M/340M params.)
- Compact – LSTM-CNN [5] with Softmax/CRF classifier, GloVe embedding (3M params.)

<u>Low-resource Dataset Simulation</u>

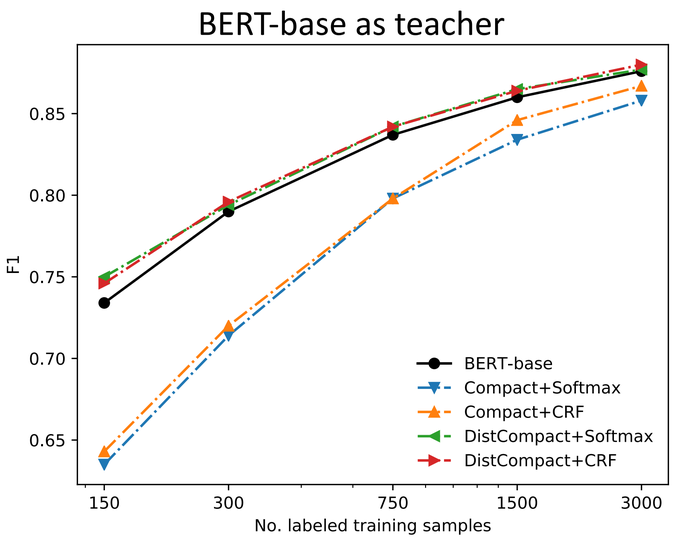
- CoNLL 2003 [6] – English (PER, ORG, DATE, MISC)
- Randomly sample labeled training sets
- Train set sizes: 150/300/750/1500/3000
- Use remaining training data as unlabeled examples

<u>Training Steps</u>

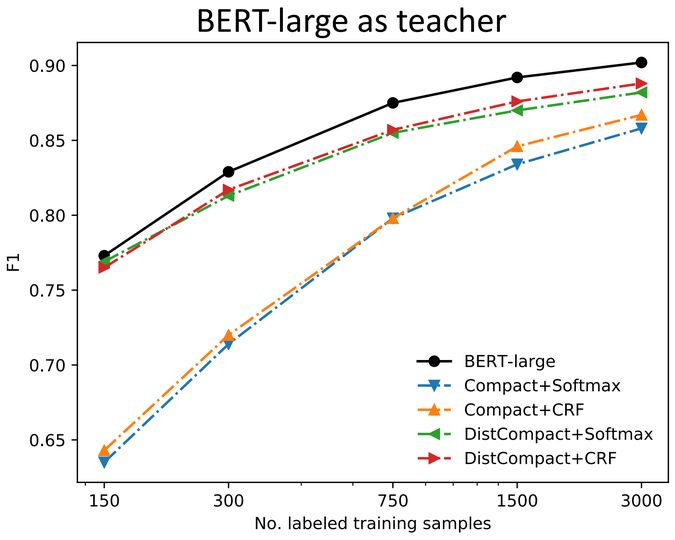
1. Fine-tune BERT on labeled data
2. Train compact model using distillation + pseudo-labeling loss

Compact Model Accuracy

BERT-base as teacher



BERT-large as teacher



Inference Speed vs. Teacher Model

CPU backend*

	BERT-base		BERT-large	
classifier	CRF	Softmax	CRF	Softmax
Batch=1	3.3x	4.3x	8.1x	10.6x
32	28.6x	33.7x	85.2x	100.4x
64	40.0x	45.2x	109.5x	123.8x
128	49.9x	55.6x	123.6x	137.8x

GPU backend*

	BERT-base		BERT-large	
classifier	CRF	Softmax	CRF	Softmax
Batch=1	0.8x	2.6x	1.5x	5.0x
32	1.3x	5.1x	3.5x	13.1x
64	2.3x	7.2x	6.3x	19.6x
128	3.9x	9.7x	10.8x	27.0x

*Dual Intel® Xeon® Processor E5-2699A v4 @ 2.40GHz; 251GB RAM; GPU: Titan Xp 12GB; OS: Ubuntu 16.04.1 (4.15.0-50- generic); Tensorflow 1.12 and PyTorch 1.0

Summary

- Compact models have major performance advantage vs. pre-trained LMs in low-resource scenarios
- Teacher models are qualitatively better with more labeled data
- We are exploring:
  - Additional NLP tasks
  - Additional Transformer based pre-trained LM
  - Non-recurrent compact models for fast inference
- Code available in NLP Architect library

NLP ARCHITECT

Paper

peter.izsak@intel.com

[1] B. Zhang, X. Pan, T. Wang, A. Vaswani, H. Ji, K. Knight, and D. Marcu. Name tagging for low-resource incident languages based on expectation-driven learning. In *HLT-NAACL*, 2016.
[2] G. E. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531, 2015.
[3] D.-H. Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. 2013.
[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. ArXiv, abs/1810.04805, 2018.
[5] X. Ma and E. H. Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. ArXiv, abs/1603.01354, 2016.
[6] E. F. T. K. Sang and F. D. Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. ArXiv, cs.CL/0306050, 2003.

The 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing @ NeurIPS 2019