CIFT: Crowd-Informed Fine-Tuning to Improve Machine Learning Ability

John P. Lalor[1], Hao Wu[2], Hong Yu[1,3]

[1]University of Massachusetts, MA [2]Boston College, MA [3]Bedford VAMC and CHOIR, MA

Overview

For Machine Learning classification tasks, training data labels are often one-hot encodings of the correct class, where training maximizes categorical cross entropy for the correct class. The training goal is to get the probability of the correct class as close as possible to 1. This type of training does not take into account different difficulties of different examples. Some training examples may be easier or harder than others. We can estimate an example's difficulty with Item Response Theory (IRT) and use that as part of training. With enough human responses to a set of questions we can model parameters of specific examples such as difficulty and discriminating ability.

$$p_{ij}(\theta_j) = \frac{1}{1 + e^{-a_i(\theta_j - b_i)}}$$

IRT model of example parameters as a function of ability (θ)

$$P(U_j|\theta_j) = \prod_{i=1}^{n} p_{ij}(\theta_j)^{y_{ij}} q_{ij}(\theta_j)^{(1-y_{ij})}$$

Ability estimate for a set of examples

In this work we use crowd responses to estimate a distribution over classes. By treating the distribution as a proxy for difficulty, we can optimize according to estimated distribution

CIFT Algorithm

Algorithm 1 CIFT Algorithm

Input: NumEpochs e, X_{train}^{N}, X_{test}, $\text{CIFT}_{\text{train}}$, IRT_{test}, Loss Function l
for $i = 1$ **to** e **do**
  Train NSE with X_{train}^{N} with loss function CCE
end for
for $i = 1$ **to** e **do**
  Train NSE with $\text{CIFT}_{\text{train}}$ with loss function l
end for
calculate accuracy for X_{test}
calculate ability for IRT_{test}

Main idea: Fine-tune a pre-trained model with the crowd-generated probability data

References

Bowman, S. R. Angeli, G. Potts, C. and Manning, D. C. A large annotated corpus for learning natural language inference. EMNLP 2015
Lalor, J. P. Wu, H. and Yu, H. Building an evaluation scale using item response theory. EMNLP 2016
Marelli, M. Menini, S. Baroni, M. Bentivogli, L. Bernardi, R. and Zamparelli, R. 2014. A sick cure for the evaluation of compositional distributional semantic models. LREC 2014
Munkhdalai, T., and Yu, H. Neural semantic encoders. EACL 2017

Data

- Subset of Stanford Natural Language Inference (SNLI) dataset (Bowman et al. 2015, Lalor et al. 2016)
- 1000 human annotations from Amazon Mechanical Turk
  - Annotations used to build IRT tests to measure ability
- Here we use the AMT responses to estimate probability distributions over classes

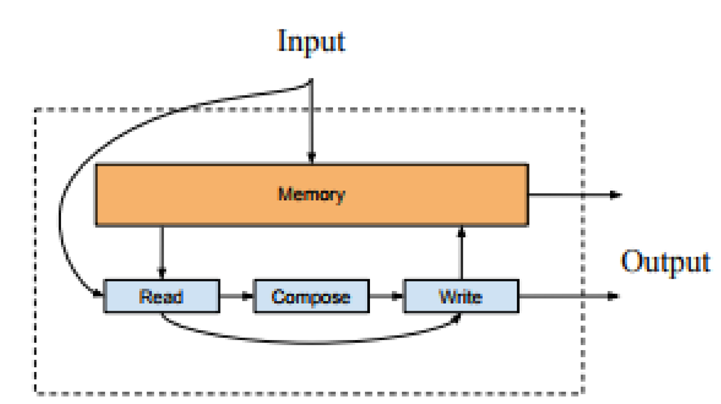
$$p(Y = y) = \frac{N_y}{N}$$

Premise	Hypothesis	Label	Difficulty
People were watching the tournament in the stadium	The people are sitting outside on the grass	Contradiction	0.51
Two girls on a bridge dancing with the city skyline in the background	The girls are sisters.	Neutral	-1.92
A little girl eating a sucker	A child eating candy	Entailment	-2.74
A young boy in a sweatshirt is doodling on a piece of paper	The class pictures are on display	Contradiction	0.78
A couple plays frisbee in a green field with trees in the background	A couple fixes dinner in their kitchen	Contradiction	-0.82
A girl in a newspaper hat with a bow is unwrapping an item	The girl is going to find out what is under the wrapping paper	Entailment	-2.69
A scene of snow and water	A snow and water scene at sunset	Neutral	-1.01

Table 1: Examples of sentence pairs from the IRT data sets, their corresponding label and difficulty as measured by IRT.

CIFT Training

- Model: Neural Semantic Encoders (Munkdalai & Yu 2017): memory-augmented neural network
- High performance on SNLI and other NLP benchmarks



- Train with full SNLI training set (500k examples)
- Fine-tune with CIFT data with one of two loss functions:
- Categorical Cross-Entropy (CCE): *Memorize* CIFT data
  - Higher performance for these examples leads to higher IRT ability estimates
- Mean Squared Error (MSE): Learn the crowd distribution over classes
  - Incorporates uncertainty in correct class for difficult examples
- Transfer learning baseline: fine-tune with SICK (Marelli et al. 2014)

$$L_i^{\text{CCE}} = -\log p(\hat{y}_i) \qquad L^{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}(\hat{p}(y_i) - p(y_i))^2$$

Manuscript available at jplalor.github.io

Results

Model	$\text{CIFT}_{\text{test}}$	Train	Test
Baseline		0.862	0.846
CIFT-CCE	4C	**0.874**	0.844
CIFT-CCE	4N	0.873	0.844
CIFT-CCE	5C	0.873	0.843
CIFT-CCE	5N	0.873	0.843
CIFT-CCE	5E	0.872	0.846
CIFT-MSE	4C	0.87	0.843
CIFT-MSE	4N	0.873	0.846
CIFT-MSE	5C	**0.874**	0.845
CIFT-MSE	5N	0.873	0.843
CIFT-MSE	5E	0.871	**0.849**

Full training -> CIFT -> Full training outperforms baseline

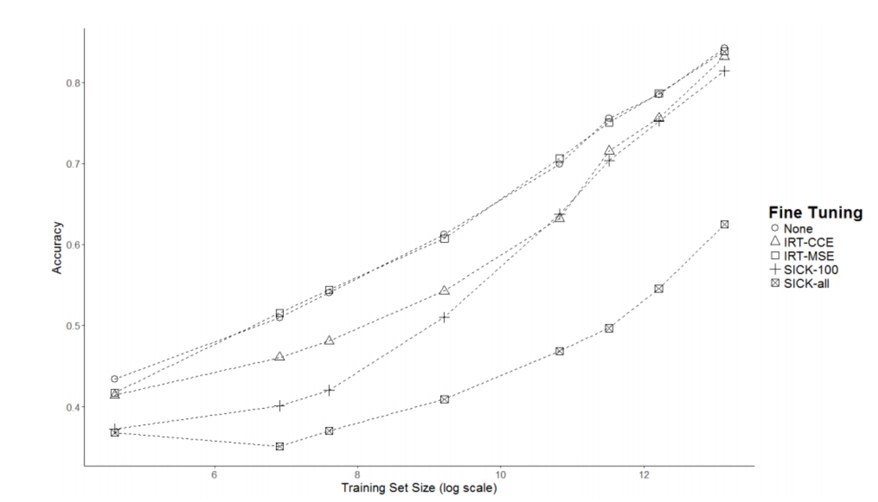
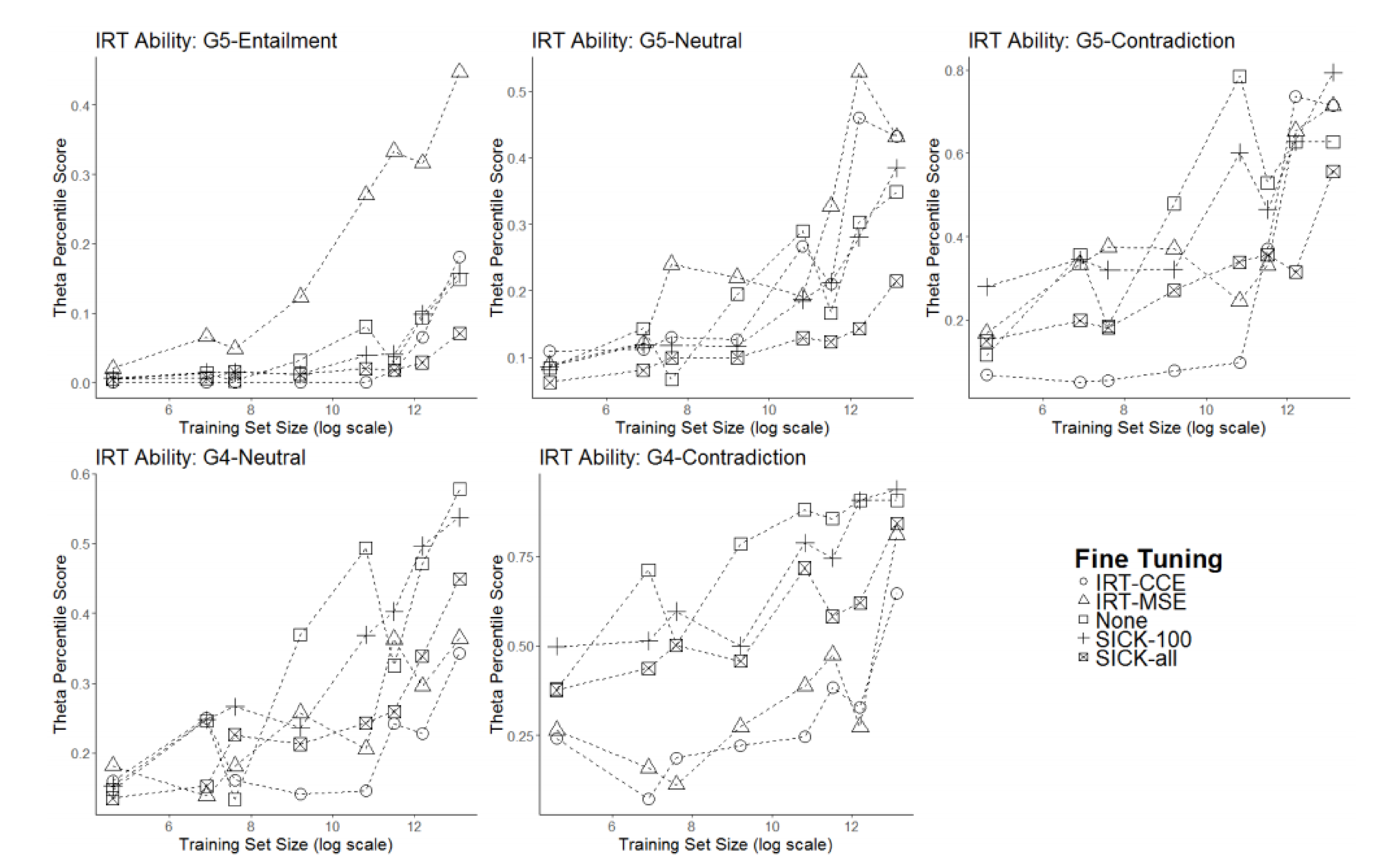


Figure 2: Accuracy scores for SNLI test set (10k examples)

CIFT: better generalization than standard transfer learning



Loss	5C	5N	5E	4N	4C
None	0.85	**0.893**	0.893	0.85	**0.893**
CIFT-CCE	**0.9**	**0.893**	0.857	0.65	0.679
CIFT-MSE	**0.9**	**0.893**	**0.929**	0.8	0.714
SICK-100	**0.9**	**0.893**	0.893	**0.9**	**0.893**
SICK-all	0.8	0.786	0.786	0.8	0.786

Table 2: Accuracy for the IRT subsets.

Ability estimates vary even when accuracies are the same (which examples you get right is important!)