Learning to Rank for **Plausible Plausiblity**

Zhongyang Li[1,2] Tongfei Chen[1] Benjamin Van Durme[1]

[1] Johns Hopkins University; [2] Harbin Institute of Technology

Summary

- The categorical nature of RTE / NLI leads to common use of cross-entropy loss for training, even when some data are designed for *plausibility* prediction, not *entailment*
- This loss is intuitively odd: it drives models to assign score 0.0 or 1.0, in constrast to the proposed *margin-based* loss
- Leads to better calibrated predictions to human intuitions
- State-of-the-art result on the Choice of Plausible Alternatives (COPA) task

The Plausibility Task

Given a premise, which hypothesis is *preferred*?

p	*I just stopped where I was*	
h_E	***I stopped in my tracks***	✓
h_N	***I stopped running right were I was***	✗
h_N	***I stopped running right were I was***	✓
h_C	***I continued on my way***	✗

E = Entailed; N = Neutral; C = Contradictory

In the plausiblity task, the correct label depends on the alternatives!

Cross-entropy loss

$$P(h_i|p) = \frac{\exp F(p, h_i)}{\sum_i \exp F(p, h_i)}$$

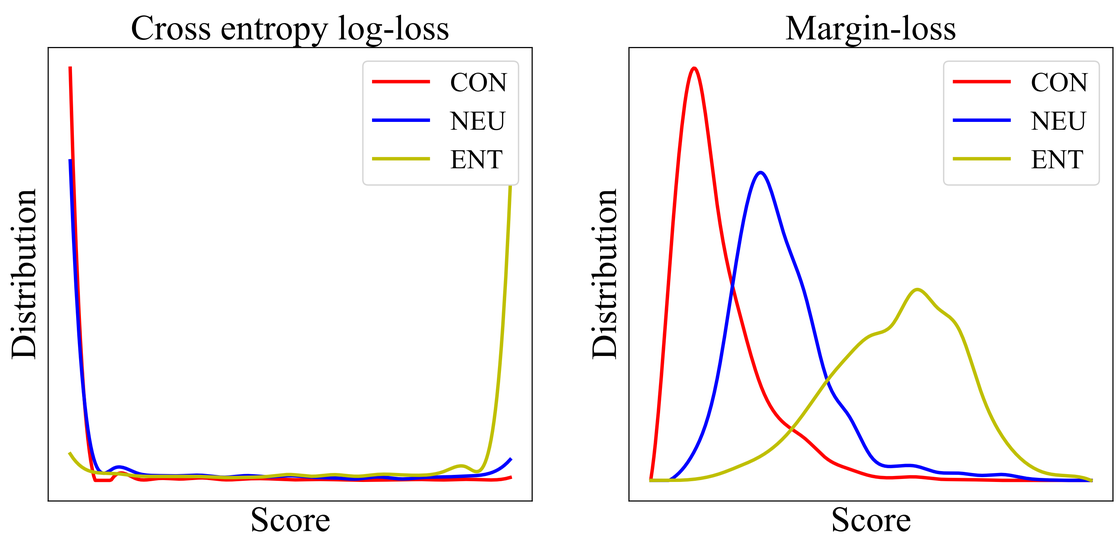
As a *classification* problem – maximizing the posterior probability normalized over all hypothesis alternatives

Margin-based loss

$$L = \frac{1}{N} \sum_{h \succ h'} \max\{0, \xi - F(p, h) + F(p, h')\}$$

As a *learning-to-rank* problem – more plausible hypothesis should *rank* higher than other less plausible hypotheses

Plausible Plausibility



Margin-based loss gives much more **plausible** scores, leading to a more **plausible** *plausibility* formulation!

Example Predictions

Recast MultiNLI

p	*I just stopped where I was*	Log	Margin
h_1	*I stopped in my tracks*	0.919	0.568
h_2	*I stopped running right were I was*	0.081	0.358
h_3	*I continued on my way*	1.71×10^{-8}	0.074

JOCI

p	*Cheerleaders performs in a lift stunt.*	Log	Margin
h_1	*The stunt is a feat.*	0.508	0.304
h_2	*The stunt is no fluke.*	0.486	0.279
h_3	*The stunt is dangerous.*	2.72×10^{-4}	0.166
h_4	*The stunt is remarkable.*	4.13×10^{-3}	0.153
h_5	*The stunt backfires.*	2.36×10^{-4}	0.107

Results on COPA

Model adapted from the original BERT sentence pair model.

Method	**Accuracy**
PMI (Jabeen+, 2014)	58.8
PMI_EX (Gordon+, 2011)	65.4
CS (Luo+, 2016)	70.2
CS_MWP (Sasaki+, 2017)	71.2
BERT (cross-entropy)	73.4
BERT (margin)	**75.4**

Related Discussions

J. Opitz & A. Frank (2018): Addressing the Winograd Schema Challenge as a sequence ranking task. *1st Int'l Workshop on Language Congition and Computational Models.*

Paper