Learning to Rank for Plausible Plausibility

Zhongyang Li[†‡] **Tongfei Chen**[‡] **Benjamin Van Durme**[‡]
[†] Harbin Institute of Technology
[‡] Johns Hopkins University
zyli@ir.hit.edu.cn, {tongfei,vandurme}@cs.jhu.edu

Abstract

Researchers illustrate improvements in contextual encoding strategies via resultant performance on a battery of shared NLU tasks. Many of these tasks are of a categorical prediction variety: given a conditioning context (e.g., an NLI *premise*), provide a label based on an associated prompt (e.g., an NLI *hypothesis*). The categorical nature of these tasks has led to common use of a cross entropy log-loss objective during training. We suggest this loss is intuitively wrong when applied to *plausibility* tasks, where the prompt by design is neither categorically *entailed* nor *contradictory* given the context. Log-loss naturally drives models to assigning scores near 0.0 or 1.0, in contrast to our proposed use of a margin-based loss. Following a discussion of our intuition, we describe a confirmation study based on an extreme, synthetically curated task derived from MNLI. We find that a margin-based loss leads to a more plausible model of plausibility. Finally, we illustrate improvements on the Choice Of Plausible Alternative (COPA) task through this change in loss.

1 Introduction

Contextualized encoders such as GPT (Radford et al., 2018) and BERT (Devlin et al., 2018) have led to improvements on various structurally similar Natural Language Understanding (NLU) tasks such as variants of Natural Language Inference (NLI). Such tasks model the conditional interpretation of a sentence (e.g., an NLI *hypothesis*) based on some other context (usually some other sentence, e.g., an NLI *premise*). The structural similarity of these tasks has meant a structurally similar modeling approach: (1) concatenate the conditioning context (premise) to a sentence to be interpreted, (2) *read* this pair using a contextualized encoder, then (3) employ the resultant representation to support classification under the label set of the task. NLI datasets employ a categorical label scheme (*Entailment*, *Neutral*, *Contradiction*) which has led to the use of a cross-entropy log-loss objective at training time: learn to maximize the probability of the correct label, and thereby minimize the probability of the competing labels.

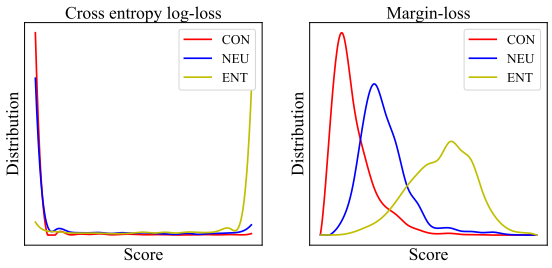


Figure 1: Dev set score distribution on a COPA-style task derived from MNLI (choose the most likely hypothesis), after training with cross-entropy log-loss and margin loss. Margin-loss leads to a more intuitively *plausible* encoding of *Neutral* statements.

We suggest this approach is intuitively problematic when applied to a task such as COPA (Choice Of Plausible Alternative) by Roemmele et al. (2011), where one is provided with a premise and two or more alternatives, and the model must select the most sensible hypothesis, with respect to the premise *and the other options*. As compared to NLI datasets, COPA was designed to have premises that are neither strictly true nor false in context: a procedure that maximizes the probability of the correct item at training time, thereby minimizing the probability of the alternative(s), will seemingly learn to *misread* future examples.

We argue that COPA style tasks should intuitively be approached as *learning to rank* problems (Burges et al., 2005; Cao et al., 2007), where an encoder on competing items is trained to assign *relatively* higher or lower scores to candidates, rather than maximizing/minimizing probabilities. In the following we investigate three datasets, be-

ginning with a constructed COPA-style variant of MNLI, designed to be adversarial. Results on this dataset support our intuition (see Figure 1). We then construct a second synthetic dataset based on JOCI, which employed a finer label set than NLI, and we find a margin-based approach strictly outperforms log-loss in this case. Finally, we demonstrate state of the art on COPA, showing that a BERT-based model trained with margin-loss significantly outperforms a log-loss alternative.

2 Background

A series of efforts have considered COPA: by causality estimation through pointwise mutual information (Gordon et al., 2011) or data-driven methods (Luo et al., 2016; Sasaki et al., 2017), or through a pre-trained language model (Radford et al., 2018, GPT).[1]

Under the Johns Hopkins Ordinal Commonsense Inference (JOCI) dataset (Zhang et al., 2017), instead of selecting which hypothesis is the most plausible, a model is expected to directly assign ordinal 5-level Likert scale judgments (from *impossible* to *very likely*). If taking an ordinal interpretation of NLI, this can be viewed as a 5-way variant of the 3-way labels used in SNLI (Bowman et al., 2015) and MNLI (Conneau et al., 2017).

In this paper, we recast MNLI and JOCI as COPA-style plausibility tasks by sampling and constructing (p, h_i, h_j) triples from these two datasets. Each premise-hypothesis pair (p, h) is labeled with different levels of plausibility $y_{(p,h)}$.[2]

3 Models

In models based on GPT and BERT for plausibility or NLI, similar architectures have been employed. The premise p and hypothesis h are concatenated into a SEP token separated sequence, along with a special sentinel CLS token inserted.

$$\begin{aligned} \text{BERT}: &\quad [\text{CLS}\ ;\ p\ ;\ \text{SEP}\ ;\ h\ ;\ \text{SEP}] \\ \text{GPT}: &\quad [\text{BOS}\ ;\ p\ ;\ \text{EOS}\ ;\ h\ ;\ \text{CLS}] \end{aligned}$$

The concatenated string is passed into the BERT or GPT encoder. One takes the encoded vector of the CLS state as the feature vector extracted from the (p, h) pair. Given the feature vector, a dense layer is stacked upon to get the final score $F(p, h)$.

[1] As reported in https://blog.openai.com/language-unsupervised/.

[2] For MNLI, *entailment* > *neutral* > *contradiction*; for JOCI, *very likely* > *likely* > *plausible* > *technically possible* > *impossible*.

Cross entropy The model is trained to maximize the probability of the correct candidate, normalized over all candidates in the set (leading to a cross entropy log-loss between the posterior distribution of the scores and the true labels):

$$P(h_i|p) = \frac{\exp F(p, h_i)}{\sum_{j=1}^{N} \exp F(p, h_j)} . \quad (1)$$

Margin-based As we have argued, the cross entropy loss used with Equation 1 is problematic. Instead we propose to use the following margin-based loss (Weston and Watkins, 1999; Li et al., 2018):

$$L = \frac{1}{N} \sum_{h>h'} \max\{0, (\xi - F(p, h) + F(p, h'))\} \quad (2)$$

where F is a score function; N is the number of pairs of hypotheses where the first is more plausible than the second under the given premise p; $h > h'$ means that h ranks before (i.e., is more plausible than) h' under premise p; and ξ is a margin hyperparameter denoting the desired scores difference between these two hypotheses.

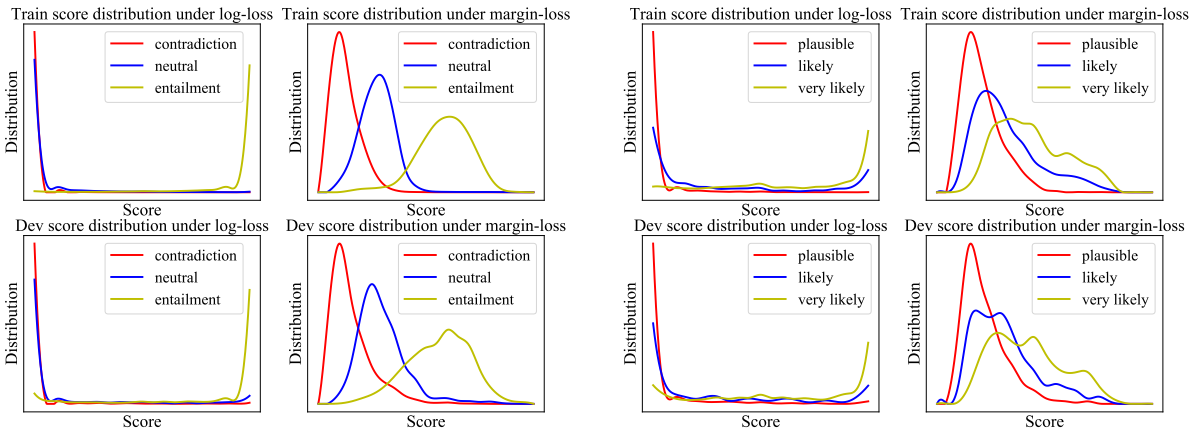
4 Experiments and Analysis

Dataset We consider MNLI, then JOCI, and finally COPA. We cast all these datasets into a format comprising of (p, h, h'), where h is more plausible than h' under the context of premise p.

MNLI In MNLI, each premise p is paired with multiple hypotheses, where we cast the label on each hypothesis as a relative plausibility judgment, where *entailment* > *neutral* > *contradiction* (we label them as 2, 1, and 0). We construct two 2-choice plausibility tasks from MNLI:

$$\text{MNLI}_1 = \{(p, h, h') \mid y_{p,h} > y_{p,h'}\}$$
$$\text{MNLI}_2 = \{(p, h, h') \mid (y_{p,h}, y_{p,h'}) \in \{(2, 1), (1, 0)\}\}$$

MNLI_1 is comprised of all pairs 2/1, 2/0 and 1/0; whereas MNLI_2 removes the presumably easier 2/0 pairs. For MNLI_1, all examples in the training set are constructed from the original MNLI training dataset, and all examples in the development set are from the original MNLI matched development dataset. For MNLI_2, all examples in our training and dev sets are from the original MNLI training dataset, where the same premise exists in both training and dev. But by adversarial design, each neutral hypothesis appears either



(a) MNLI_1 (b) JOCI_1

Figure 2: Train and development score distribution after training with a cross entropy log-loss and a margin-loss.

Dataset	Train	Eval
MNLI_1	410k	**dev:** 8.2k
MNLI_2	142k	**dev:** 130k
JOCI_1	8.7k	**dev:** 3.0k
JOCI_2	2.3k	**dev:** 1.9k
COPA	500	**test:** 500

Table 1: Statistics of various plausibility datasets. All numbers are numbers of (p, h, h') triplets.

Dataset	Acc under log-loss	Acc under margin-loss
MNLI_1	**93.6**	93.4
MNLI_2	87.9	87.9
JOCI_1	86.6	**86.9**
JOCI_2	76.6	**78.0**

Table 2: Experimental results on MNLI_* and JOCI_*.

as the preferred (beating contradiction), or dispreferred alternative (beaten by entailment), which is flipped at evaluation time.

JOCI In JOCI, every inference pair is labeled with their ordinal inference Likert-scale labels 5, 4, 3, 2, 1. Similar to MNLI, we cast these to 2-choice problems under the following conditions:

$$\text{JOCI}_1 = \{(p, h, h') \mid y_{p,h} > y_{p,h'} \geq 3\}$$
$$\text{JOCI}_2 = \{(p, h, h') \mid (y_{p,h}, y_{p,h'}) \in \{(5, 4), (4, 3)\}\}$$

We ignore inference pairs with scores below 3, aiming for sets akin to COPA, where even the dispreferred option is still often semi-plausible.

COPA We label alternatives as 1 and 0.

Table 1 shows the statistics of these datasets.

Setup We fine-tune the BERT-base uncased model (Devlin et al., 2018) using our proposed margin-based loss. For MNLI and JOCI datasets, the margin hyperparameter $\xi = 0.2$. Since COPA does not contain a training set, we use the original dev set as the training set, and perform 10-fold cross validation to find the best hyperparameter: $\xi = 0.37$, learning rate $\eta = 3 \times 10^{-5}$, training 3 epochs.

Method	Acc (%)
PMI (Jabeen et al., 2014)	58.8
PMI_EX (Gordon et al., 2011)	65.4
CS (Luo et al., 2016)	70.2
CS_MWP (Sasaki et al., 2017)	71.2
BERT_{log} (ours)	73.4
$\text{BERT}_{\text{margin}}$ (ours)	**75.4**

Table 3: Experimental results on COPA test set.

Results on Recast MNLI and JOCI Table 2 shows results on the recast MNLI and JOCI datasets. We find that for the two synthetic MNLI datasets, margin-loss performs similarly to cross entropy log-loss. Shifting to the JOCI datasets, with less extreme (contradiction/entailed) hypotheses, especially in the adversarial JOCI_2 variant, we find margin-loss outperforming log-loss.

Though log-los and margin-loss give close quantitative results on predicting the more plausible (p, h) pairs, they do so in different ways, confirming our intuition. From Figure 2 we find that the log-loss always predicts the more plausible (p, h) pair with very high probabilities close to 1, and predicts the less plausible (p, h) pair with very low probabilities close to 0. Figure 2, showing a per-premise normalized score distribution from margin-loss, is more reasonable and explainable: hypothesis with different plausibility are distributed hierarchically between 0 and 1.

COPA Result Table 3 shows our results on COPA. Compared with previous state-of-the-art knowledge-driven baseline methods, a BERT model trained with a cross entropy log-loss achieves better performance. When training the

Dataset	Premise	Hypotheses	Gold	Log	Margin
$MNLI_1$	*a. I just stopped where I was.*	1. *I stopped in my tracks*	2	0.919	0.568
		2. *I stopped running right where I was.*	1	0.0807	0.358
		3. *I continued on my way.*	0	1.71×10^{-8}	0.0739
$MNLI_1$	*b. An organization's activities, core processes and resources must be aligned to support its mission and help it achieve its goals.*	1. *An organization is successful if its activities, resources and goals align.*	2	0.505	0.555
		2. *Achieving organizational goals reflects a change in core processes.*	1	0.495	0.257
		3. *A company's mission can be realized even without the alignment of resources.*	0	3.48×10^{-5}	0.187
$JOCI_1$	*c. A few people and cars out on their daily commute on a rainy day.*	1. *The commute is a journey.*	5	0.994	0.473
		2. *The commute is bad.*	4	5.79×10^{-3}	0.230
		3. *The commute becomes difficult.*	3	1.28×10^{-3}	0.157
$JOCI_1$	*d. Cheerleaders in red uniforms perform a lift stunt.*	1. *The stunt is a feat.*	5	0.508	0.304
		2. *The stunt is no fluke.*	4	0.486	0.279
		3. *The stunt is dangerous.*	3	2.72×10^{-4}	0.166
		4. *The stunt is remarkable.*	3	4.13×10^{-3}	0.153
		5. *The stunt backfires.*	3	2.36×10^{-4}	0.107
COPA	*e. She jumped off the diving board.*	1. *The girl landed in the pool.*	1	0.972	0.520
	e′. She ran on the pool deck.		0	0.028	0.480
COPA	*f. The student knew the answer to the question.*	1. *He raised his hand.*	1	0.982	0.738
		2. *He goofed off.*	0	0.018	0.262

Table 4: Examples of premises and their corresponding hypotheses in various plausibility datasets, with gold labels and scores given by the log-loss and margin-loss trained models.

BERT model with a margin-loss instead of a cross entropy log-loss, our method gets the new SOTA result on the established COPA splits, with an accuracy of 75.4%.[3]

Analysis Table 4 shows some examples from the $MNLI_1$, $JOCI_1$ and COPA datasets, with per-premised normalized scores (for sake of analysis across examples; relative magnitude of alternatives given a premise is unchanged).

For the premise a from $MNLI_1$, log-loss gives a very high score (0.919) for the entailment hypothesis $a.1$, while gives a very low score (0.0807) for the neutral hypothesis $a.2$, and an extreme low score (1.71×10^{-8}) for the contradiction hypothesis $a.3$. Though the log-loss can achieve high accuracy by making these extreme prediction scores, we argue these scores are unintuitive. For the premise b from $MNLI_1$, log-loss again gives a very high score (0.505) for the hypothesis $b.1$. But it also gives a high score (0.495) for the neutral hypothesis $b.2$. The contradiction hypothesis $b.3$ still gets an extreme low score (3.48×10^{-5}).

These are the two ways for the log-loss approach to make predictions with high accuracy: always giving very high score for the entailment hypothesis and low score for the contradiction hypothesis, but giving either very high or very low score for the neutral hypothesis. In contrast, the margin-loss gives more intuitive scores for these two examples. Also, we get similar observations from the $JOCI_1$ examples c and d.

The fifth example from COPA is asking for a more plausible 'cause' premise for the 'effect' hypothesis. Here, each of the two candidate premises e and e' is a possible answer. The log-loss gives very high (0.972) and very low (0.028) scores for the two candidate premises, which is unreasonable. Whereas the margin-loss gives much more rational ranking scores for them (0.52 and 0.48). For the sixth example f, which is asking for a more likely 'effect' hypothesis for the 'cause' premise, margin-loss still gets more reasonable prediction scores than the log-loss.

5 Conclusion

In this paper, we propose that margin-loss in contrast to log-loss is a more plausible training objective for COPA-style *plausibility* tasks. Through adversarial construction we illustrated that a log-loss approach can be driven to encode plausible statements (Neutral hypotheses in NLI) as either extremely likely or unlikely, which was highlighted in contrasting figures of per-premise normalized hypothesis scores. This intuition was shown to lead to a new state of the art in the original COPA task, based on a margin-based loss.

[3] We exclude a blog-posted OpenAI GPT result, provided without experimental conditions and not reproducible.

References

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Christopher Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Gregory N Hullender. 2005. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine learning (ICML-05)*, pages 89–96.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pages 129–136. ACM.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proc. EMNLP*, pages 670–680.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Andrew S Gordon, Cosmin A Bejan, and Kenji Sagae. 2011. Commonsense causal reasoning using millions of personal stories. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Shahida Jabeen, Xiaoying Gao, and Peter Andreae. 2014. Using asymmetric associations for commonsense causality detection. In *Pacific Rim International Conference on Artificial Intelligence*, pages 877–883. Springer.

Zhongyang Li, Xiao Ding, and Ting Liu. 2018. Constructing narrative event evolutionary graph for script event prediction. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4201–4207. AAAI Press.

Zhiyi Luo, Yuchen Sha, Kenny Q Zhu, Seung-won Hwang, and Zhongyuan Wang. 2016. Commonsense causal reasoning between short texts. In *Fifteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language understanding paper. pdf*.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Shota Sasaki, Sho Takase, Naoya Inoue, Naoaki Okazaki, and Kentaro Inui. 2017. Handling multiword expressions in causality estimation. In *IWCS 12th International Conference on Computational SemanticsShort papers*.

Jason Weston and Chris Watkins. 1999. Support vector machines for multi-class pattern recognition. In *ESANN 1999, 7th European Symposium on Artificial Neural Networks, Bruges, Belgium, April 21-23, 1999, Proceedings*, pages 219–224.

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. Ordinal common-sense inference. *Transactions of the Association of Computational Linguistics*, 5(1):379–395.