Attention-Focused Adversarial Training for Robust Temporal Reasoning

Lis Kanashiro Pereira[1], Kevin Duh[2], Fei Cheng[3], Masayuki Asahara[4], Ichiro Kobayashi[1]

Ochanomizu University[1], Johns Hopkins University[2], Kyoto University[3], NINJAL[4]

Overview

Goal: Improve model *generalization* and *robustness* for **NLU**

Adversarial Training

$$\min_{\theta} \mathbb{E}_{(x,y)\sim D}[\max_{\delta} l(f(x+\delta;\theta),y)]$$

- Better *generalization* and *robustness*
- In **NLP**: perturbations are usually *added to the embeddings only*
- Perturbations are usually *generated by running a fixed number of gradient steps*

Weaknesses:

- *Adding the perturbation to the embeddings only might not be optimal*
- *Other layers of the transformer based models can encode specific syntactic and semantic information*

Solution:

- Add the noise to a *combination of the model layers* instead of only to the embedding layer
- Adds the adversarial perturbation to multiple hidden states or attention representations of the model layer

ML-ALICE

Adversarial perturbations δ_1 and δ_2 added to layer *r*

$$\min_{\theta} \mathbb{E}_{(x,y)\sim D}[\max_{\delta_{1r}} l(f(x+\delta_{1r};\theta),y) + \alpha \max_{\delta_{2r}} l(f(x+\delta_{2r};\theta), f(x;\theta))],$$

- Adds the adversarial perturbation to multiple hidden states or attention representations of the model layer
- Encourages the model to generate **more diverse** perturbed input examples
- Adding the adversarial perturbation to **attention representations** performs **best.** We hypothesize they might be more robust and less sensitive to noise.

Zero-Shot Results

We fine-tune RoBERTa_BASE on the CosmosQA training data and evaluate it on various Temporal test sets

Methods	TimeML	MC-TACO		SCT
	Acc	EM	F1	Acc
Standard	49.35	10.59	32.27	87.28
FreeLB	41.87	12.53	24.59	88.99
SMART	42.77	9.38	27.11	86.21
ALICE	43.77	**15.02**	21.11	91.24
ML-ALICE	**55.23**	11.48	**43.64**	**91.98**

ML-ALICE is effective in enhancing model generalizability in out domains.

Experiments and Results

Methods	TEA		TimeML	MC-TACO		SCT	MATRES	
	Acc	F1	Acc	EM	F1	Acc	Acc	F1
Standard	95.20	89.40	80.86	39.79	68.63	92.95	90.54	87.80
FreeLB	95.75	90.62	82.75	44.37	71.52	92.68	90.54	87.78
SMART	95.32	89.77	82.25	46.77	73.07	92.89	91.26	88.77
ALICE	95.63	90.35	82.35	47.00	73.04	93.43	89.54	86.80
ML-ALICE (hidden)	95.69	90.52	83.35	49.25	**74.78**	93.53	90.26	87.59
ML-ALICE (attendtion)	**96.72**	**92.80**	**83.94**	**49.77**	73.93	**94.07**	**91.40**	**88.97**
RoBERTa_LARGE	95.99	91.11	81.06	**51.05**	**76.85**	**96.37**	91.12	88.93
Best Temporal Model Result in the literature	**-**	**-**	81.70	59.08	79.46	91.90	**-**	87.30

Default text encoder: RoBERTa_BASE

ML-ALICE with adversarial perturbations added to attention representations performs best among all models.