Predicting Fine-tuning Performance with Probing

Zining Zhu[1,2], Soroosh Shahtalebi[2], Frank Rudzicz[1,2,3]
[1] University of Toronto [2] Vector Institute of Artificial Intelligence [3] Unity Health Toronto

Introduction

Fine-tuning is the "de-facto" method for evaluating the development of large neural NLP systems, but probing has become increasingly popular.

Fine-tuning	**Probing**
Task resembles deployment	Tasks are out-of-domain
Test cases are inclusive	Test cases are specific
Aim at high performance	Aim at faithful interpretations
Computationally heavy	Computationally friendly

Can probing be used in the development of DNN models? There are two questions:
1. **Feasibility**. The probing results appear disjointed to fine-tuning results -- are they relevant?
2. **Operation**. There are many probing configurations. Where should we probe?

Methodology

We use probing accuracies $\boldsymbol{S}$ to regress the fine-tuning performances $\boldsymbol{A}$ of K models:

$$\theta_* = \text{argmin}_\theta \Sigma_k \, ||\theta^T \boldsymbol{S} - \boldsymbol{A}||^2$$

- Root-mean-squared-error (RMSE) measures the quality of this regression.
- But random features could achieve small RMSE too (let's write it as $RMSE_c$)
- So we measure the reduction of RMSE:

$$RMSE_{\text{reduction}} = \frac{RMSE_c - RMSE}{RMSE_c} \times 100$$

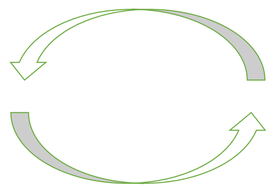
Experiments (selected)

	RTE	**COLA**	**MRPC**	**SST2**	**QNLI**	**QQP**
All layers one task (5.2)						
BShift	6.24	52.80	**53.18**	29.78	55.29	51.64
CoordInv	2.10	66.59	18.18	44.24	56.35	56.57
ObjNum	2.19	44.20	28.02	53.15	60.64	72.38
SOMO	30.90	44.75	29.39	29.28	38.64	55.68
Tense	3.07	48.42	34.65	22.29	41.37	75.58
SubjNum	-19.66	**78.56**	34.48	47.75	64.74	51.50
TreeDepth	4.37	53.03	9.54	46.98	62.79	54.67
One layer per task (5.4)	36.12	62.66	25.78	49.87	59.79	26.73
Only three features (5.5)	**41.69**	75.66	47.56	**72.59**	**80.52**	**76.77**
	CoordInv_1	ObjNum_2	TreeDepth_1	SubjNum_1	SubjNum_2	TreeDepth_6
	TreeDepth_1	SubjNum_2	SOMO_4	BShift_3	Tense_8	Tense_8
	BShift_12	TreeDepth_12	ObjNum_7	CoordInv_10	CoordInv_9	Tense_12

Table 1: RMSE reduction from baseline. A larger value shows the probing results more indicative of the fine-tuning performance. A small (or even negative) value means the probing results are not informative, compared to random features. The **bold-font** configurations are those with the highest RMSE reductions for predicting each fine-tuning task.

Results & Conclusion

Feasibility: Probing results are relevant to, and can be predictive of, the fine-tuning performance.
Operation: We provide some heuristics for setting up the probing tasks.
We call for completing a "feedback loop": use probing in developing large neural NLP systems:

Model development  Model interpretation using e.g., probing