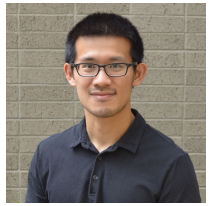
Predicting Fine-tuning Performances with Probing

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.

People can develop high-performing DNNs

Rank	Name	Model
1	Liam Fedus	SS-MoE
2	Microsoft Alexander v-team	Turing NLR v5
3	ERNIE Team - Baidu	ERNIE 3.0
4	Zirui Wang	T5 + UDG, Single Model (Google Brain)
5	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4
6	SuperGLUE Human Baselines	SuperGLUE Human Baselines

Models are evaluated by fine-tuning performances

There is a common theme in the developments of DNN models:

- Explore novel techniques to train models.
- Evaluate the effectiveness by **fine-tuning**.
  - i.e., Attach a classification layer. Optimize this layer and the network *together* on the target dataset.
- Report the quality of the models using (mostly) fine-tuning results.

Probing can also evaluate DNN models

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
Computation-heavy (e.g., 20h for QQP)	Lighter (e.g., 1 hr CPU time for 7 tasks)

Question: Can probing be used in model developments?

Challenges:

Feasibility: The probing results appear disjointed to fine-tuning results -- are they relevant?

Operation: There are many probing configurations. Where should we probe?

This project:

Feasibility: We give a positive answer.

Operation: We provide some empirical answers.

Predict the fine-tuning performance with probing

- K models are tested on some probing tasks.
- The probing accuracies of the k^{th} model are written in the vector $\mathbf{S}^{(k)}$.
- On fine-tuning task T, the k^{th} model can reach performance $\mathcal{A}_T^{(k)}$.
- We predict $\mathcal{A}_T$ from $\mathbf{S}$ using a linear regressor (parameterized by θ).

$$\theta_* = \text{argmin}_\theta \Sigma_k ||\theta^T \mathbf{S}^{(k)} - \mathcal{A}_T^{(k)}||^2$$

$$\text{RMSE} = \sqrt{\frac{1}{K} \Sigma_k ||{\theta_*}^T \mathbf{S}^{(k)} - \mathcal{A}_T^{(k)}||^2}$$

Control setting

We need to control for the artefacts. Why?

- Suppose our regressor using $\mathbf{S}$ gets $\mathrm{RMSE} = 0.01$ on both T_1 and T_2.
- But T_2 appears slightly "harder"...
- Random features can get $\mathrm{RMSE}_c = 0.02$ for T_1 but only 0.10 for T_2.
- Then $\mathbf{S}$ provides more predictability for T_2 than T_1, but RMSE itself can't tell.

So we instead measure and report the RMSE_reduction:

$$\mathrm{RMSE\_reduction} = \frac{\mathrm{RMSE}_c - \mathrm{RMSE}}{\mathrm{RMSE}_c} \times 100$$

Models

- 5 Transformer-based models from huggingface: `roberta-base`, `xlm-roberta-base`, `microsoft/deberta-base`, `albert-base-cased`, `xlnet-base-cased`
- Corrupt by MLM on scrambled Wikipedia for 500, 1k, 2k, 4k, 6k steps.
  - Except `xlnet` (since MLM doesn't apply to it)
  - In an ablation study (§ 5.9), we show that this procedure produces sufficiently diverse models.
- There are 25 models in total.

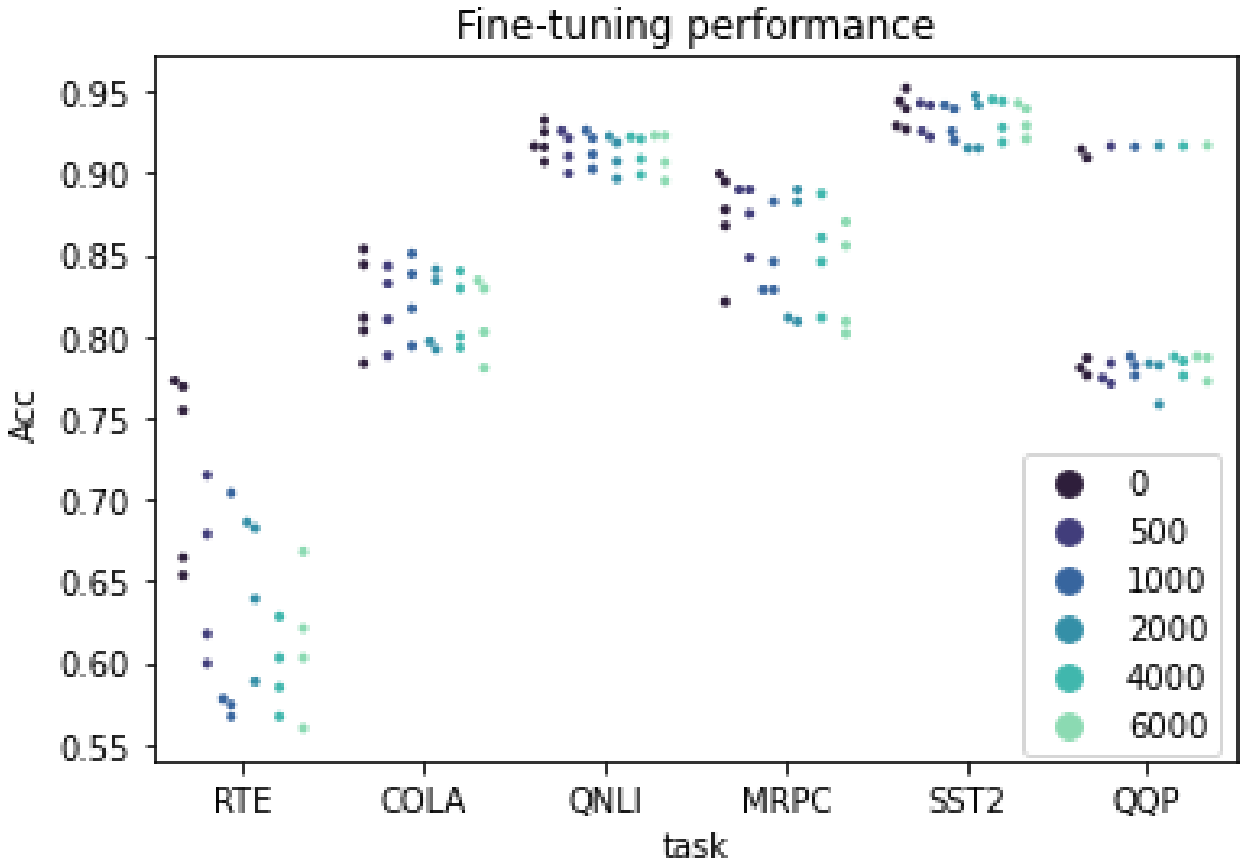
Probing tasks

We take 7 probing tasks from SentEval (Conneau and Kiela, 2018):

- `Bigram Shift`, `Coordination Inversion`, `Objective Number`, `Semantic Odd-Man Out`, `Past vs. Present`, `Subject Number`, `Tree Depth`.
- We subsample 1,200 data points per class.
  - Zhu et al., (2022) showed that several thousand samples already can have sufficient statistical powers.

Fine-tuning tasks

These tasks come from GLUE: `RTE`, `COLA`, `QNLI`, `MRPC`, `SST2`, `QQP`



Is there a "more useful" probing task than others? (§ 5.2)

There is no definitive answers, but depending on the tasks, there are some regularities.

	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

Are probing some layers more useful than others? (§ 5.3)

	RTE	COLA	MRPC	SST2	QNLI	QQP
bigram shift (BShift)	4,5	2,4,5	2,4,5,9	2,5,6	2,4,5	2,4,5
coordination inversion (CoordInv)	5,6,12	1,2,4,6	1,6	1,4,6	1,4-6	2-4,6
object number (ObjNumber)	1	1,3,8,11	1,3	1,3-5,8,11	1,3,8,11	1-5,12
semantic odd man out (SOMO)	4,5,8,12	2-6	3,4	3,5,6	2-6	2,5-9,12
past present (Tense)	1	1,3,5	1,5,6	1,11	1,3,5,8	1-5,8-11
subject number (SubjNum)	None	1,3-6,9	1	1,4	1	1,2,3,4
tree depth (TreeDepth)	1	1	1	1,3,5	1	1-3,7,8,11

Table 2: Layers with significant probing results ($p < .05$ from one-way ANOVA) with residual dof $= 12$.

What is the best that we can do? (§ 5.5)

With as few as 3 features, the maximum reachable $\mathrm{RMSE\_reduction}$ values are nontrivial.

Fine-tuning task	RTE	COLA	QNLI	MRPC	SST2	QQP
$\mathrm{RMSE\_reduction}$	41.69	75.66	47.56	72.59	80.52	76.77

Ablation: Use different probing methods (§ 5.6)

MLP-20 and RandomForest-100 are recommended.

	RTE	COLA	MRPC	SST2	QNLI	QQP
Highest-accuracy probe in §5.2 - §5.5	41.69	78.56	53.18	**72.59**	80.52	76.77
Specify one probing method (§5.6)						
DecisionTree	51.98	68.48	54.31	70.90	74.35	52.85
LogReg	45.28	78.34	44.87	70.26	83.13	73.98
MLP-10	48.50	72.12	45.88	65.87	73.82	81.97
MLP-20	47.37	74.94	**63.79**	69.22	79.10	**82.67**
RandomForest-10	50.64	74.08	50.17	68.2	75.19	59.66
RandomForest-100	**53.94**	**79.20**	53.21	71.60	**83.25**	72.72
SVM	51.71	74.01	57.92	71.44	76.78	73.03

Table 3: Maximum RMSE reductions using different probing configurations. The **bold-font** numbers are the maximum values in each column.

Other experiments

There are many other experiments, including:

- Use only one probing task (all 12 layers).
- Use smaller probing datasets (400 instead of 1200 per class)
- Uncertainty analysis.

Call for completing the "feedback loop"

Currently, probing is mostly used as *post-hoc* interpretations...

- Probing analysis is (in general) computationally friendly.
- Probing can give fine-grained diagnostics to empower model developments.
- Probing literature contain rich resource of "test material".

Probing analysis can be useful for model developments!

Summary

How can probing be useful for building DNN models?

- We show that probing results can predict an important intermediate signal, fine-tuning accuracy.
- We analyze the utility of different parameters in configuring the probing.