

OOD-Probe: A Neural Interpretation of Out-of-Domain Generalization

Zining Zhu, Soroosh Shahtalebi, Frank Rudzicz







1

DNNs can generalize between many domains



Why do DNNs generalize?

Many generalization algorithms are based on invariance principles [1]

- The learned representations remain invariant across domains.
- Let the optimal performance match on different domains.

There are usually some trade-off between the two clauses.

- What are the extent of these trade-offs?
- Nowadays, the OOD generalization algorithms are only evaluated by e.g., outdomain prediction accuracies.

An interpretability method: probing classifier

- In NLP, the desire to understand the model intrinsics led to many interpretability methods.
 - Probing classifier is a popular one.
- Probing has revealed many interesting findings about DNNs:
 - About linguistic structure. [2]
 - About an intrinsic pipeline that does "feature extraction -> semantics". [3]
 - About how DNNs respond to anomalies. [4]
 - Many others...

[2] Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by selfsupervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054.

[3] Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT Rediscovers the Classical NLP Pipeline. In ACL, pages 4593–4601, Florence, Italy.

[4] Bai Li, Zining Zhu, Guillaume Thomas, Yang Xu, and Frank Rudzicz. 2021. How is BERT surprised? Layerwise detection of linguistic anomalies. In ACL pages 4215–4228, Online.

OOD-Probe

We use probes f_p to predict the domain attribute E from DNN representations.

- OOD-Probe does not affect the DNN training.
- Minimal computing overheads.
- Wide applicability to OOD generalization algorithms.



What does probe results entail?

"Is there information about _____ here in this model?" [5]

- Difference choices of performance metrics are relevant to different information-theoretic aspects.
 - Please refer to the paper for details.
- In this paper, we use accuracy.
 - So the probing performance and the generalization performance can be easily compared.

Data, model, and algorithms

- Data:
 - RotatedMNIST, ColoredMNIST, VLCS, PACS
- Model:
 - 5-layer CNN for *MNIST, ResNet-18 for VLCS and PACS.
- Algorithms:
 - 21 OOD generalization algorithms on DomainBed.
 - Trained using the default hyperparameters.

An "increase - decrease" trend

									Р	robing a	accurac	y (PACS	5)									-10
probe_5_out -	0.73	0.79	0.58	0.88	0.55	0.55	0.93	0.92	0.82	0.82	0.85	0.92	0.9	0.92	0.96	0.69	0.91	0.8	0.91	0.86	0.92	- 0.9
probe_4_out -	0.77	0.94	0.53	0.94	0.93	0.61	0.94	0.93	0.95	0.91	0.95	0.93	0.9	0.92	0.97	0.81	0.95	0.89	0.94	0.91	0.94	- 0.8
probe_3_out -	0.97	0.98	0.77	0.98	0.98	0.83	0.98	0.98	0.98	0.96	0.97	0.97	0.98	0.97	0.98	0.98	0.97	0.97	0.97	0.98	0.98	- 0.7 - 0.6
probe_2_out -	0.96	0.97	0.96	0.96	0.97	0.95	0.96	0.96	0.97	0.96	0.97	0.96	0.96	0.97	0.96	0.97	0.96	0.96	0.97	0.96	0.96	- 0.5
probe_1_out -	0.94	0.93	0.93	0.94	0.95	0.93	0.94	0.95	0.95	0.95	0.94	0.94	0.95	0.94	0.95	0.94	0.95	0.94	0.95	0.94	0.95	- 0.4 - 0.3
probe_0_out -	0.87	0.87	0.89	0.88	0.88	0.89	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.87	0.88	0.88	0.88	0.86	0.89	0.88	0.88	- 0.2
I	ANDMask	CAD	CDANN	CORAL	CondCAD	DANN	ERM	GroupDRC	IB_ERM		IRM algorithm	MLDG	MMD	MTL	Mixup	RSC	SD	SagNet	SelfReg	TRM	VREX	

This trend varies across datasets

Probing accuracy (RotatedMNIST)															-10							
probe_4_out -	0.52	0.53	0.55	0.26	0.31	0.54	0.53	0.53	0.26	0.33	0.45	0.55	0.28	0.53	0.59	0.54	0.46	0.51	0.49	0.51	0.54	- 0.9
probe_3_out -	0.85	0.85	0.87	0.85	0.84	0.87	0.87	0.86	0.83	0.71	0.81	0.87	0.83	0.86	0.9	0.87	0.79	0.86	0.83	0.86	0.86	- 0.8 - 0.7
probe_2_out -	0.89	0.88	0.9	0.9	0.88	0.9	0.9	0.9	0.9	0.82	0.83	0.91	0.89	0.89	0.95	0.89	0.9	0.89	0.9	0.9	0.89	- 0.6 - 0.5
probe_1_out -	0.94	0.95	0.97	0.96	0.95	0.96	0.96	0.96	0.96	0.88	0.89	0.96	0.96	0.96	0.98	0.96	0.97	0.96	0.96	0.96	0.96	- 0.4 - 0.3
probe_0_out -	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.98	0.98	1	0.99	0.99	1	0.99	0.99	1	0.99	0.99	0.99	- 0.2
	ANDMask	CAD	CDANN	CORAL	CondCAD	DANN	ERM	GroupDRC	B_ERM		IR ^M algorithm	MLDG	MMD	MTL	Mixup	RSC	SD	SagNet	SelfReg	TRM	VREX	- 0.1

9

Probing results and DG performances

- Strong correlations between DG accuracy and probing accuracies on lower layers for RotatedMNIST.
- This trend is less visible on ColoredMNIST, VLCS, and PACS.



Conclusion

- We propose OOD-Probe, a general method to understand the mechanisms of generalization in DNNs.
- OOD-Probe shows some interesting findings, including:
 - Middle blocks in ResNet-18 encode domain information the most linearly.
 - Bottom layers in CNNs encode these information the most linearly.
 - Probing results sometimes correlate to the OOD generalization performances.