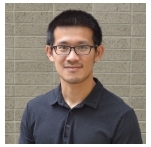
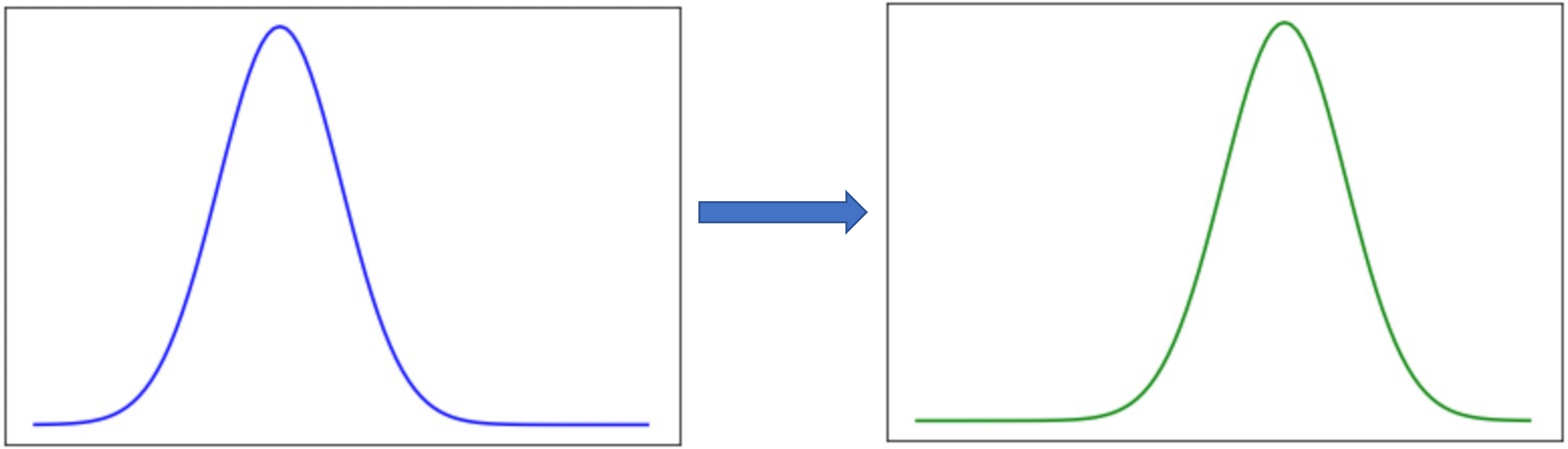
OOD-Probe: A Neural Interpretation of Out-of-Domain Generalization

Zining Zhu, Soroosh Shahtalebi, Frank Rudzicz



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., out-domain prediction accuracies.

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2020. Invariant Risk Minimization. *arXiv:1907.02893 [cs, stat]*, March. arXiv: 1907.02893.

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 self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054.

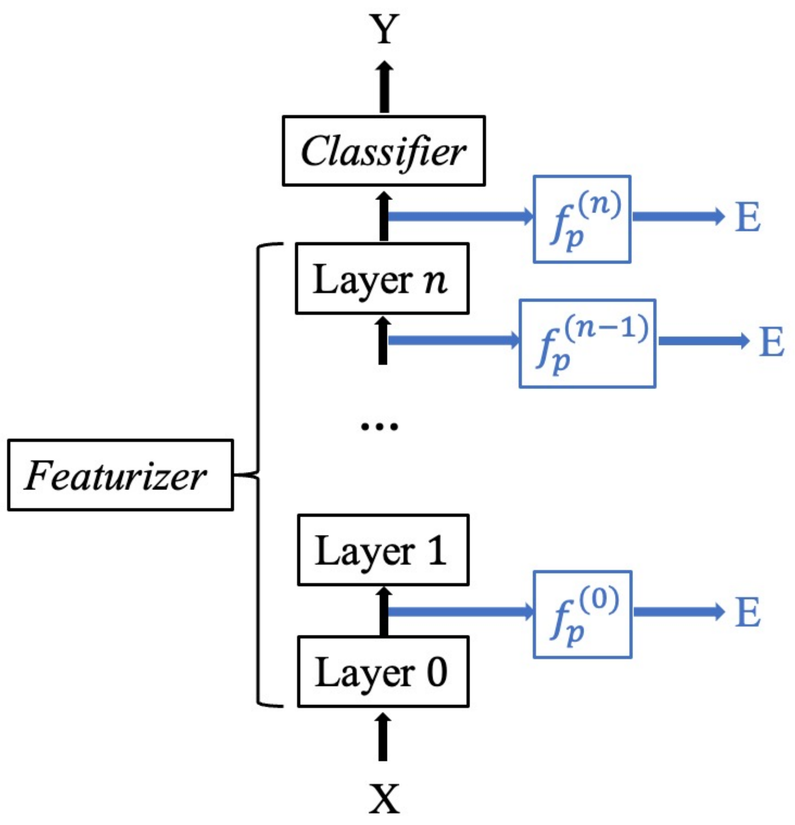
[3] Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT Rediscovers the Classical NLP Pipeline. In *ACL*, pages 4593–4601, Florence, Italy.

[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]

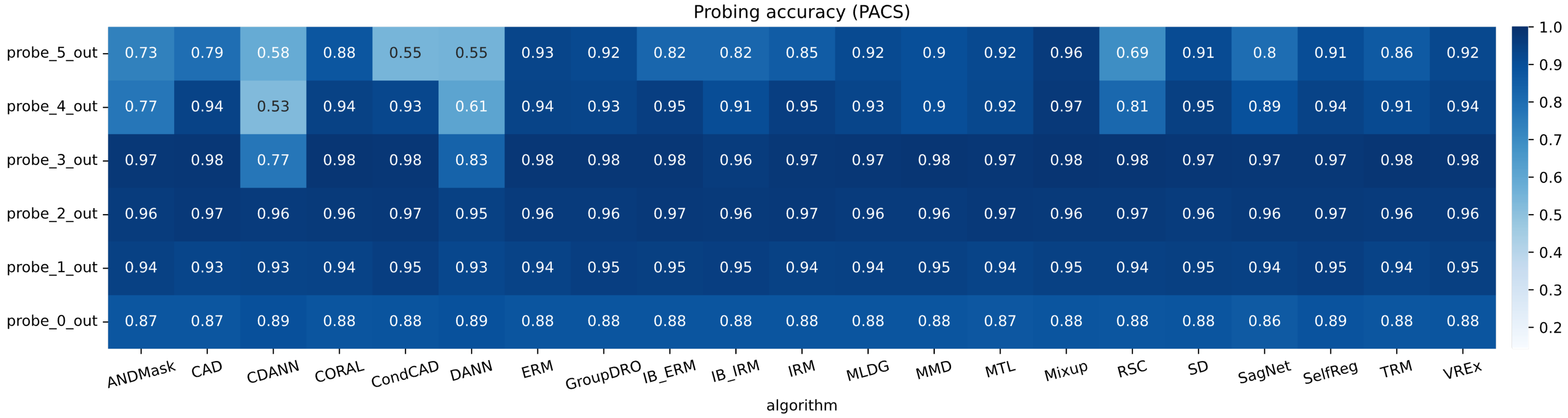
- Difference choices of performance metrics are relevant to different information-theoretic aspects.
  - Please refer to the paper for details.
- In this paper, we use accuracy.
  - So the probing performance and the generalization performance can be easily compared.

[5] Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. In *ICLR*.

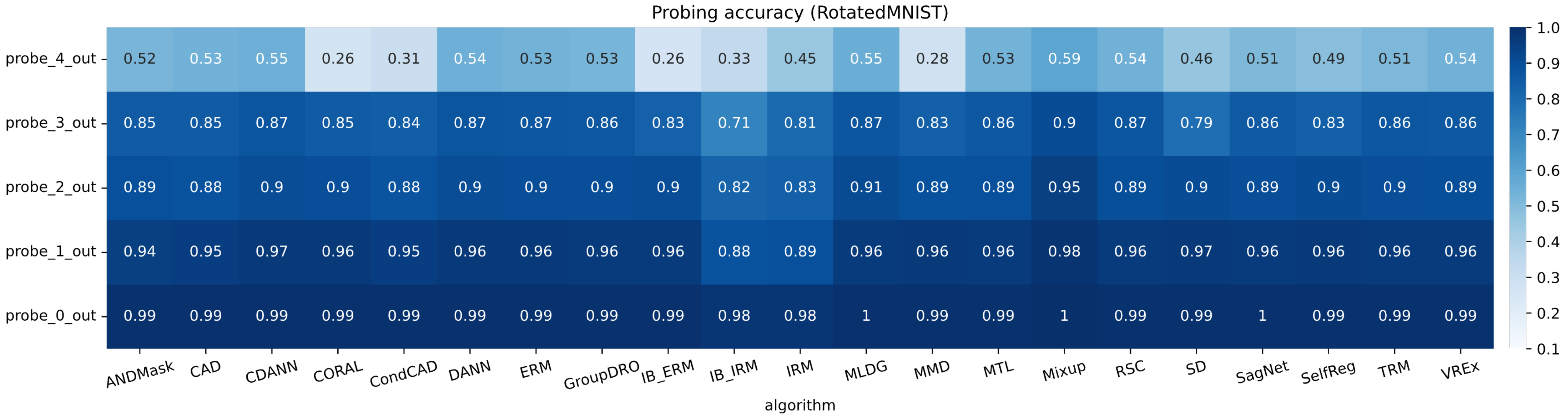
Data, model, and algorithms

- 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

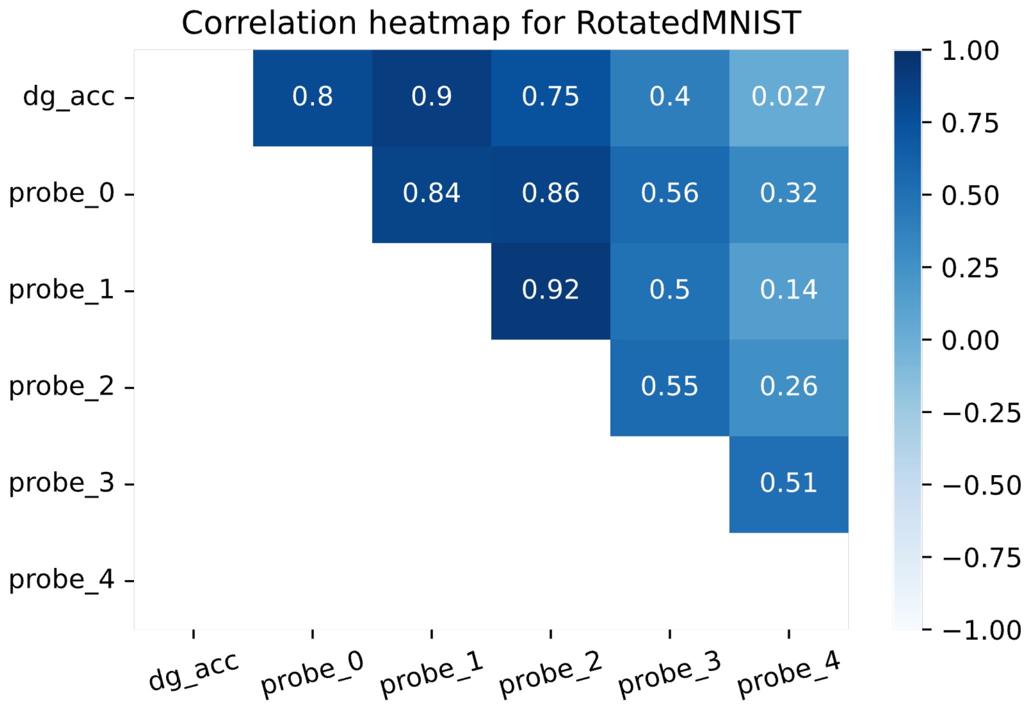


This trend varies across datasets



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.