HOW TO PROBE: SIMPLE YET EFFECTIVE TECHNIQUES FOR IMPROVING POST-HOC EXPLANATIONS

Siddhartha Gairola[1,2], Moritz Böhle[1], Francesco Locatello[1,2], Bernt Schiele[1]

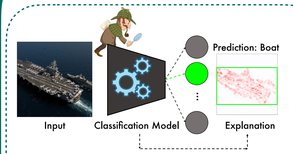
[1]Max Planck Institute for Informatics, Saarland Informatics Campus
[2]Institute of Science and Technology Austria

Paper

Paper

ICLR 2025

Motivation



Post-hoc attribution methods are a popular tool for "explaining" DNNs

Key Question

"To what extent does the training objective influence the explanation quality of pre-trained models?"

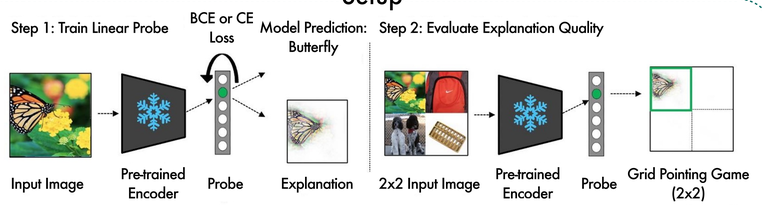
Why This Matters?

Post-hoc attribution methods are **popular** but **training-agnostic**.

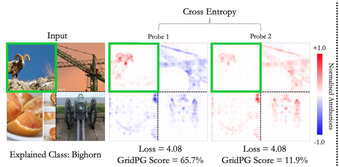
Diversity in pre-training requires revisiting that assumption.

What we do?

Setup



Softmax Invariance-Issue



Positive and Negative attributions NOT calibrated.

$$\frac{\exp(\hat{y}_{c,i}+\delta_i)}{\sum_k \exp(\hat{y}_{k,i}+\delta_i)} = \frac{\exp(\hat{y}_{c,i})\exp(\delta_i)}{\sum_k \exp(\hat{y}_{k,i})\exp(\delta_i)} = \frac{\exp(\hat{y}_{c,i})}{\sum_k \exp(\hat{y}_{k,i})}.$$

CE Loss is **invariant** to adding a **shift** δ to all output logits.

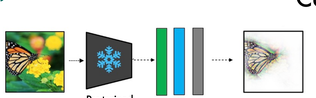
$$\frac{\exp(\mathbf{w}_c^T \mathbf{a}_i)}{\sum_k \exp(\mathbf{w}_k^T \mathbf{a}_i)} = \frac{\exp(\mathbf{w}_c^T \mathbf{a}_i + \Delta \mathbf{w}^T \mathbf{a}_i)}{\sum_k \exp(\underbrace{\mathbf{w}_k^T \mathbf{a}_i}_{\hat{y}_{k,i}} + \underbrace{\Delta \mathbf{w}^T \mathbf{a}_i}_{\delta_i})} = \frac{\exp(\hat{y}_{c,i}+\delta_i)}{\sum_k \exp(\hat{y}_{k,i}+\delta_i)}$$

Image-specific shifts result from $\mathbf{w}_k \rightarrow \mathbf{w'}_k = \mathbf{w}_k + \mathbf{\Delta w}$.

$$\mathcal{L}_{\mathrm{BCE},i} = -\sum t_{c,i} \log(\sigma(\hat{y}_{c,i})) + (1 - t_{c,i}) \log(1 - \sigma(\hat{y}_{c,i}))$$

BCE penalizes non-target shifts, **improving** class localization.

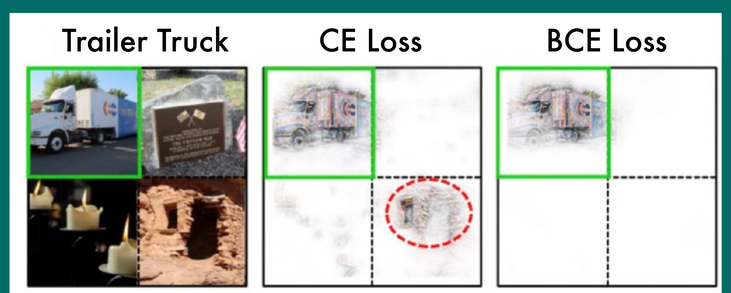
Complex Probes



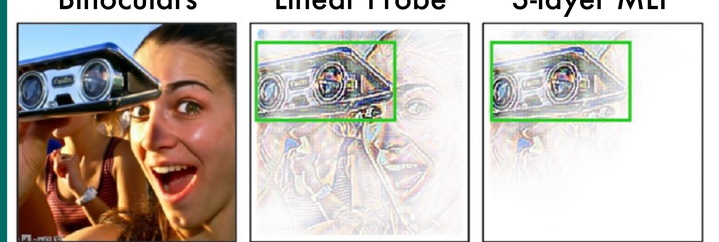
Conventional MLP: $\mathbf{f}(\mathbf{x};\theta) = \mathbf{l}_L \circ \mathbf{l}_{L-1} \circ \ldots \circ \mathbf{l}_2 \circ \mathbf{l}_1(\mathbf{x})$

B-cos Layer: $\mathbf{l}_l^*(\mathbf{a}_l;\mathbf{W}_l) = |c(\mathbf{a}_l;\mathbf{W}_l)|^{\mathrm{B}-1} \odot \mathbf{W}_l\,\mathbf{a}_l$

Interpretability Isn't Plug-and-Play



BCE probes localize **better** than **CE** probes.



B-cos MLPs lead to **improved** explanations.

Training of the **classification head** matters more than pre-training.

Just **<10%** of parameters can define explanation quality.

What we find?

Evaluation Setting

Pre-training Frameworks

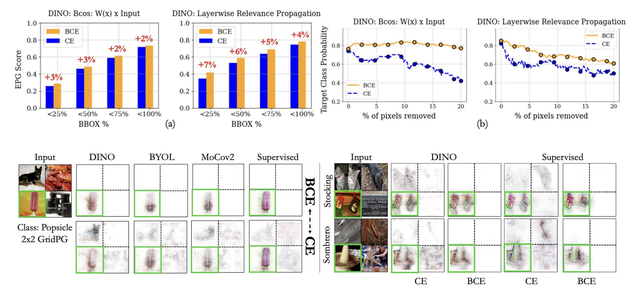
1. Fully supervised Learning
2. Vision-Language Learning (CLIP [1])
3. Self-supervised Learning

(MoCov2, BYOL, DINO)

Explanation Methods

1. Gradient Based: LRP, IxG, IntGrad
2. Activation Based: GradCAM
3. Perturbation Based: LIME, ScoreCAM
4. Inherently Interpretable: B-cos
5. Vision Transformer Based: CheferLRP, etc.

BCE Probes localize better than CE probes.



B-cos MLPs improve class-specific localization.

References

1. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) ICML 2021.
2. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.B.: Momentum contrast for unsupervised visual representation learning. CVPR 2020
3. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M.:Bootstrap your own latent a new approach to self-supervised learning. NeurIPS 2020
4. Caron, M., Touvron, H., Misra, I., J'egou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. CVPR 2021.