Studying How to Efficiently and Effectively Guide Models with Explanations

Sukrut Rao* Moritz Böhle* Amin Parchami-Araghi Bernt Schiele
Max Planck Institute for Informatics, Saarland Informatics Campus
sukrut.rao@mpi-inf.mpg.de
sukrutrao.github.io/publication/using-explanations-to-guide-models

ICCV23 PARIS

Problem setting

Image Attribution

DNNs often rely on spurious features

→ relying on person to predict bike

→ relying on background features

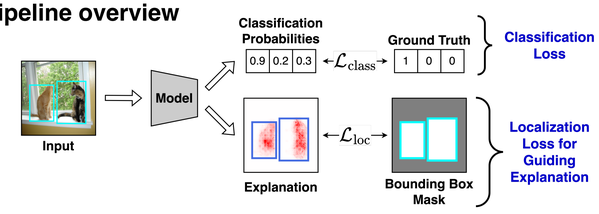
which could cause poor generalization

Want models to be 'right for the right reasons'

→ explicitly guide models via annotations

Guiding Models via Explanations

Pipeline overview



We study efficient and effective guidance

Low annotation costs
→ bounding boxes
→ coarse annotations
→ few annotations
→ guidance 'depth'

In-depth study across
→ loss functions
→ model types
→ attribution methods
→ guidance 'depth'
→ on large-scale datasets

Large-scale Datasets
PASCAL VOC, MS COCO

Attribution Methods
B-cos, GradCAM, IxG, IntGrad

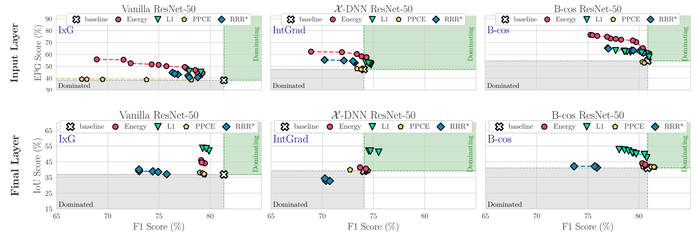
Localization Losses
Energy, L1, PPCE, RRR*

Guidance at various Layers
Input, Final, and Intermediate

Results

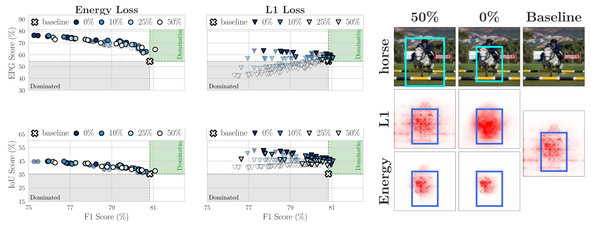
Extensive comparison: Layers, models, datasets, metrics

Observations
→ Energy loss ↑focus
→ L1 loss ↑bbox coverage
→ B-cos most 'guidable'
→ consistent across datasets
→ qualitative improvements



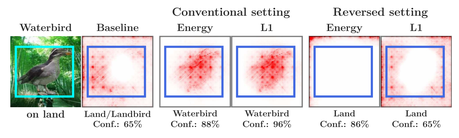
Qualitative comparison: focused model explanations

Robust guidance: coarse annotations

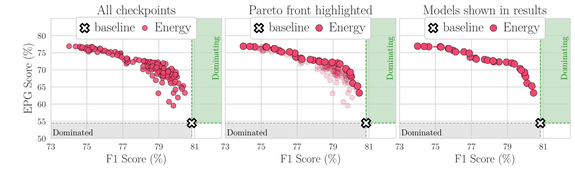


'Right' features: Waterbirds-100

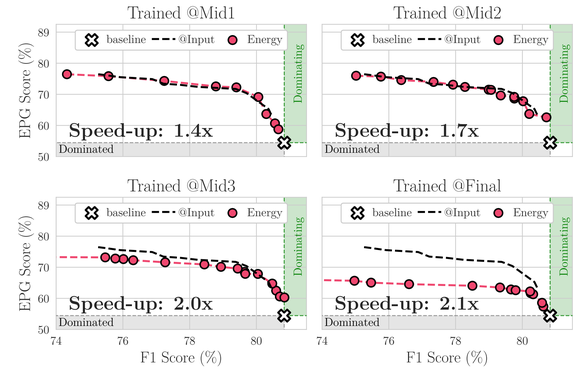
	Conventional		Reversed	
Model	Worst	Overall	Worst	Overall
Baseline	43.4 (±2.4)	68.7 (±0.2)	56.6 (±2.4)	80.1 (±0.2)
Energy	**56.1 (±4.0)**	**71.2 (±0.1)**	**62.8 (±2.1)**	**83.6 (±1.1)**
L_1	51.1 (±1.9)	69.5 (±0.2)	58.8 (±5.0)	82.2 (±0.9)



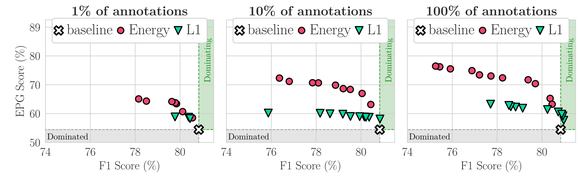
Fair comparison: Pareto curve evaluation



Efficient guidance: intermediate layers...



... and limited bounding box annotations



References

RRR (Ross et al., 2017), PPCE (Shen et al., 2021), EPG (Wang et al., 2020), L1 Loss (Gao et al., 2022), VOC (Everingham et al. 2010), COCO (Lin et al., 2014), B-cos (Böhle et al., 2022), Grad-CAM (Selvaraju et al., 2017), IxG (Shrikumar et al., 2017), IntGrad (Sundararajan et al., 2017), Waterbirds (Sagawa et al., 2019)