Visual correspondence-based explanations improve AI robustness and human-AI team accuracy

Giang Nguyen* Mohammad Reza Taesiri* Anh Nguyen

* Equal contribution

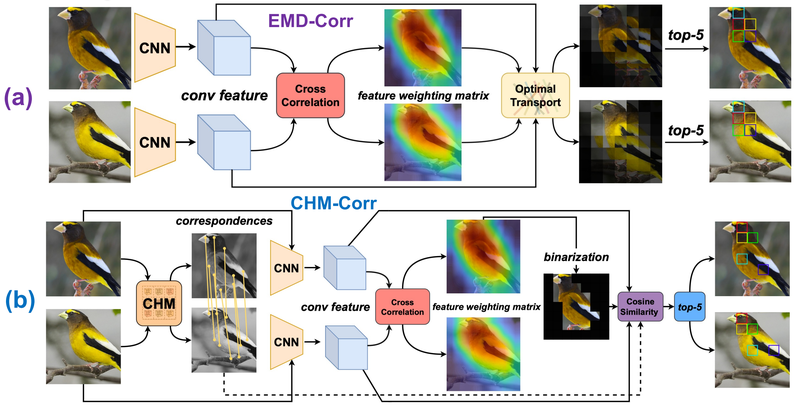
AUBURN UNIVERSITY · UNIVERSITY OF ALBERTA

Paper and code: https://aub.ie/visual-correspondence

Introduction

- Question: How to make prototype-based XAI models (1) accurate on in-distribution and out-of-distribution (OOD) data and (2) improve user accuracy?
- We propose **two novel architectures of explainable image classifiers** that first explain, and then predict (as opposed to post-hoc explanation methods).
- We improve OOD robustness and human-AI team accuracy on both ImageNet and CUB.

Correspondence-based classifiers

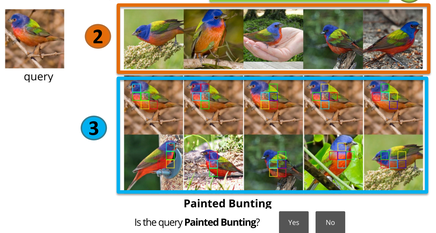


Both **(1)** re-rank *k*NN top-50 results using the patch-wise distance between the query and each candidate over the top-5 most similar patches; then **(2)** take the dominant class among the re-ranked top-20 as the predicted label.
(a) EMD-Corr: First compute patch-wise similarity, and then find correspondences via solving EMD; **(b) CHM-Corr:** First find correspondences via a matching network (CHM), and then compute patch-wise similarity.

Human studies

709 qualified, online participants
2 tasks – ImageNet and CUB
~60 users/method/task
21,270 test trials

6 explanation methods:
- ResNet-50: 1
- kNN, EMD-NN, CHM-NN: 1 + 2
- EMD-Corr, CHM-Corr: 1 + 3



Experimental Results

Test set	Pre-trained features	Training set	ResNet-50	kNN	EMD-Corr	CHM-Corr	CHM-Corr+
ImageNet	ImageNet	ImageNet	**76.13**	74.77	74.93	74.40	n/a
ImageNet-ReaL	ImageNet	ImageNet	**83.04**	82.05	82.32	81.97	n/a
ImageNet-R	ImageNet	ImageNet	36.17	36.18	**37.75**	37.62	n/a
ImageNet Sketch	ImageNet	ImageNet	24.09	24.72	25.36	**25.61**	n/a
DAmageNet	ImageNet	ImageNet	5.93	7.59	**8.16**	8.10	n/a
Adversarial Patch	ImageNet	ImageNet	55.04	59.30	59.43	**59.86**	n/a
CUB ← far →	ImageNet	ImageNet	n/a	54.72	**60.29**	53.65	49.63
CUB ← close →	iNaturalist	CUB	**85.83**	85.46	84.98	83.27	81.54

Top-1 Accuracy (%). ResNet-50 models' classification layer is fine-tuned on the training set. All other classifiers are non-parametric based on the pre-trained features and retrieve exemplars from the training set during testing.

1. kNN improves upon ResNet-50 on OOD datasets

- kNN decreases ResNet-50 accuracy on ImageNet (**–1.36** pts) but improves on *all* OOD datasets, e.g., on **+1.66** pts on DAmageNet and **+4.26** pts on Adversarial Patch.
- In CUB, kNN almost equalizes the fine-tuned model in performance (85.46% vs. 85.83%).

2. Using correspondences further improves kNN on ImageNet

- **EMD-Corr** and **CHM-Corr** further improve kNN by **+1.5** pts on ImageNet-R and ImageNet Sketch, and by **+0.5** pts on DAmageNet and Adversarial Patch.

(a) Explanations for kNN's `parachute` decision (top) and CHM-NN (bottom) (b) Explanations for CHM-Corr's `ibex` decision

kNN mislabels the image ***parachute*** because it matches the parachute scenes. CHM-Corr correctly labels the input as it matches the ***ibex*** image mostly using the animal's features, discarding the background information.

(a) Explanations for kNN's `toaster` decision (top) and EMD-NN (bottom) (b) Explanations for EMD-Corr's `hen` decision

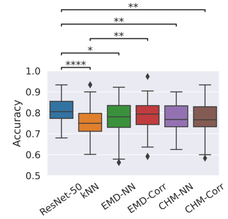
Operating at the image-level similarity, kNN incorrectly matches the input to the ***toaster*** images. EMD-Corr instead ignores the adversarial ***toaster*** patch in the input and only uses the head and neck patches of the ***hen*** to make decision

3. EMD-Corr is more robust than CHM-Corr

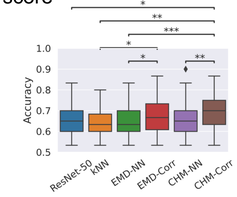
- Using ImageNet features, when tested on CUB, **EMD-Corr** improves kNN by **+5.57** pts.
- **CHM-Corr** fails on CUB due to low-quality CHM correspondences (pretrained on PF-PASCAL).
- Keypoints from humans (CHM-Corr+) worsens accuracy (**–3.5 to –10.5** pts) vs. **CHM-Corr**.

Method	ImageNet		CUB	
	Users	Accuracy	Users	Accuracy
ResNet-50	60	81.56 ± 5.54	60	65.50 ± 7.46
kNN	59	75.76 ± 8.55	59	64.75 ± 7.14
EMD-NN	57	77.72 ± 8.27	59	64.12 ± 7.07
EMD-Corr	59	78.87 ± 6.57	58	67.64 ± 7.44
CHM-NN	60	77.56 ± 6.91	60	65.72 ± 8.14
CHM-Corr	59	77.23 ± 7.56	59	69.72 ± 9.08

Per-user Top-1 accuracy (%) on ImageNet and CUB



(a) ImageNet



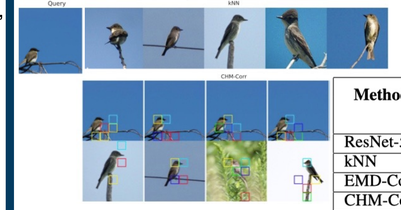
(b) CUB

4. Visual explanations hurt user ImageNet accuracy

- With only confidence scores and no explanations, users score 81.56%, better than users of other methods.

5. Correspondences-based explanations are effective for CUB users

- EMD-Corr and CHM-Corr users outperform all other users, **+2 pts** and **+4 pts** w.r.t. no visual explanations.
- Using correspondences, users more accurately ***reject*** AI's ***incorrect*** predictions.



Prediction: **Sayornis** GT: **Flycatcher**
All CHM-Corr users (3/3) rejected while all kNN users (4/4) wrongly accepted.

Method	ImageNet		CUB	
	AI-Only	Human-AI Team	AI-Only	Human-AI Team
ResNet-50	86.05	90.41	87.11	87.74
kNN	85.95	87.85	87.40	86.56
EMD-Corr	85.91	89.48	86.88	87.03
CHM-Corr	85.36	88.51	85.48	87.22

AI-only and human-AI team Accuracy (%) on ImageNet and CUB. Human-AI: AIs classify N% of inputs that they are confident, leaving the rest for humans to classify.

6. Explanations improve human-AI team accuracy

- Complementary human-AI team accuracy: human-AI team performs better than either AI-alone or human-alone.
- On ImageNet, we found complementary team performance across all explanation methods, improving by **+2 to +4.5** pts.
- On CUB, we also found complementary performance on all methods (except for kNN) but the improvements are smaller than those on ImageNet.

Discussion

1. Correspondence-based explanations show: (a) where models are looking at like attribution maps (e.g., GradCAM), (b) exemplars like nearest neighbors, c) part-to-part correspondences (e.g., ProtoPNet). How do we leverage such details for fine-grained and high-stakes downstream problems (e.g., X-ray classification)?
2. Incorporating human attention to guide AI → interactive human-AI collaboration.