THU-PM-256

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Teacher-generated spatial-attention labels boost robustness and accuracy of contrastive models

Yushi Yao*, Chang Ye*, Junfeng He+, Gamaleldin F. Elsayed+

* : Equal technical contribution + : Equal leadership and advising contribution

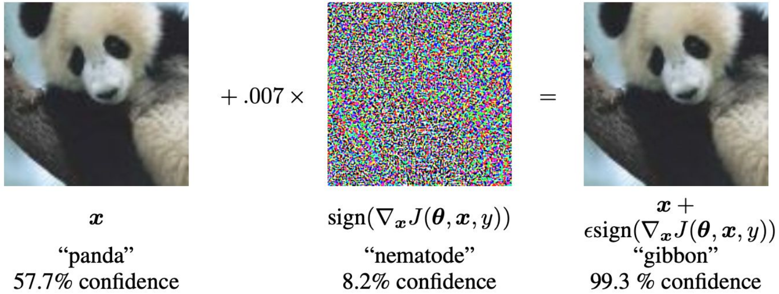
Google Research

Overview

- We create a dataset with spatial attention maps for the ImageNet benchmark by using a teacher model trained on human spatial attention labels.
- We use spatial-attention labels from the teacher model as an additional prediction target to train the contrastive model.
- The proposed method can learn better representation, leading to better accuracy and robustness for several downstream tasks.

Motivation

SALICON: Saliency in Context, Jiang et al, CVPR 2015



Explaining and harnessing adversarial examples, Goodfellow et al, ICLR 2015

Human visual system

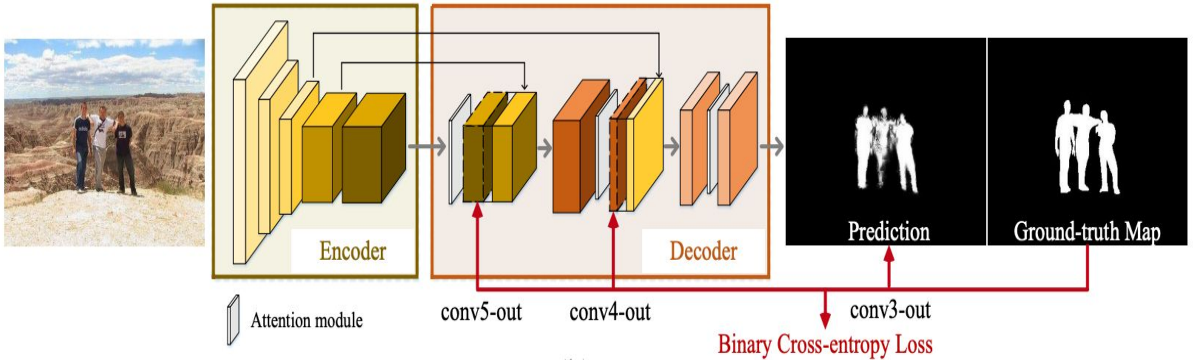
- Focus on specific region in visual scene that are useful to perform a specific vision task.

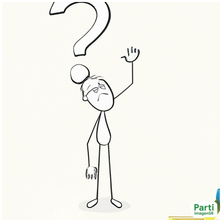
Machine visual system

- Attend to physically meaningless patterns.
- Tend to exploit features that are predictive but not causal

Hypothesis

Existing work of applying human spatial attention to supervised model

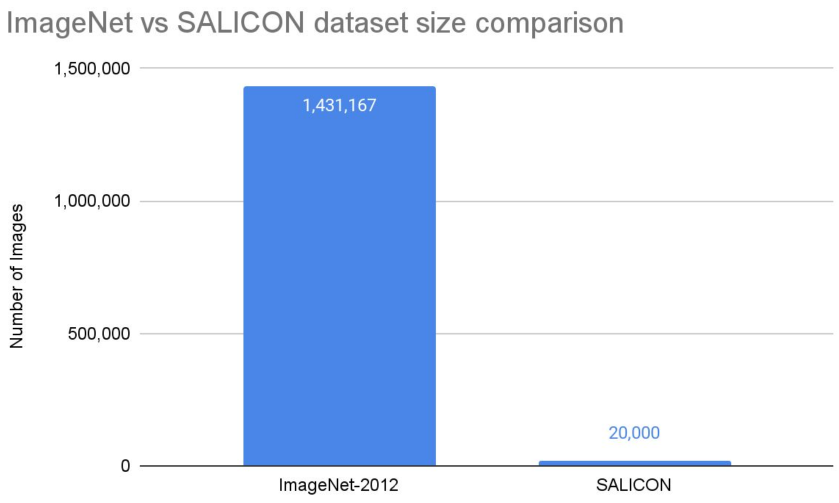




Would it also benefit to self-supervised model?

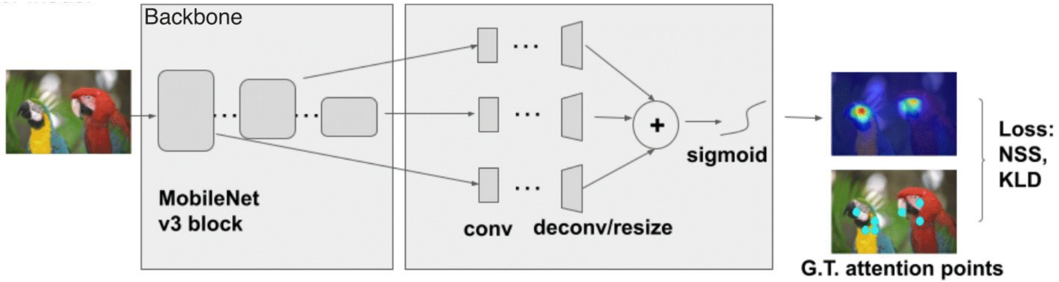
Understanding more about human and machine attention in deep neural networks, Lai et al, TMM 2020

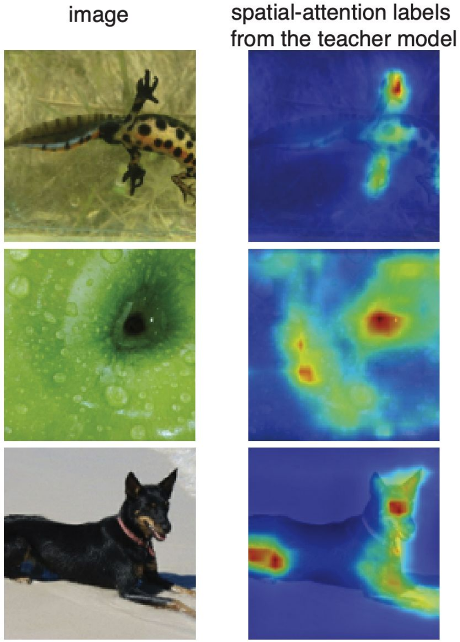
Challenge



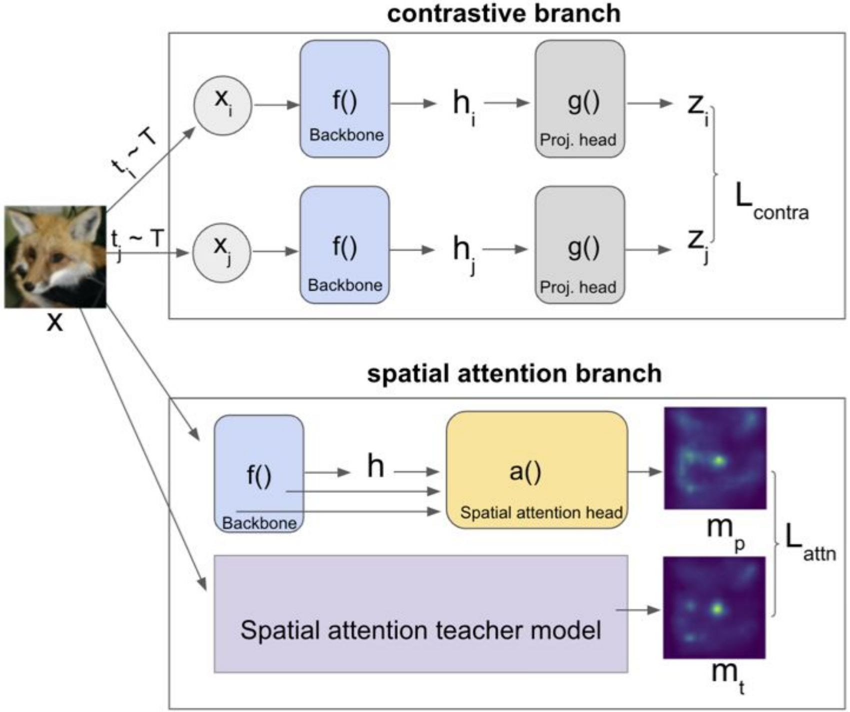
- No existing large human spatial attention dataset
- Expensive to collect to collect a large volume of human spatial attention data.

Teacher model
for predicting human saliency





Contrastive model with spatial attention maps



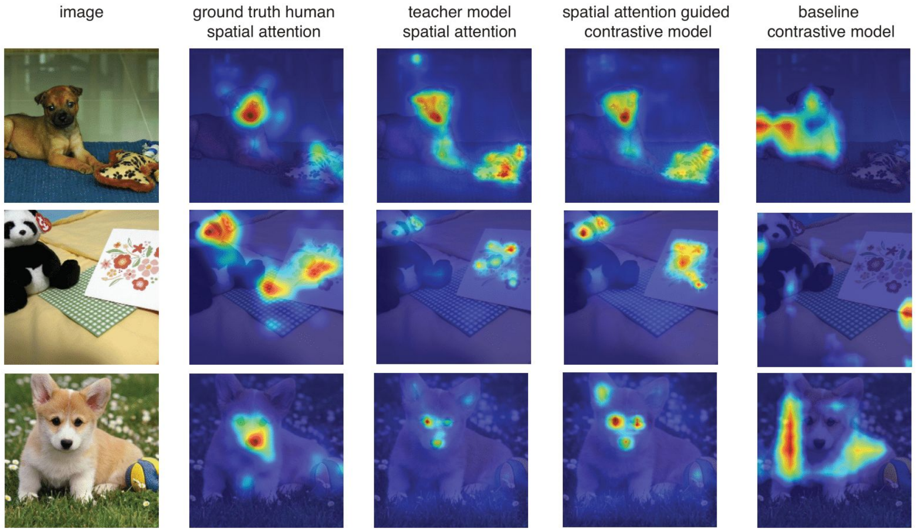
$$L = L_{contra} + L_{attn}$$

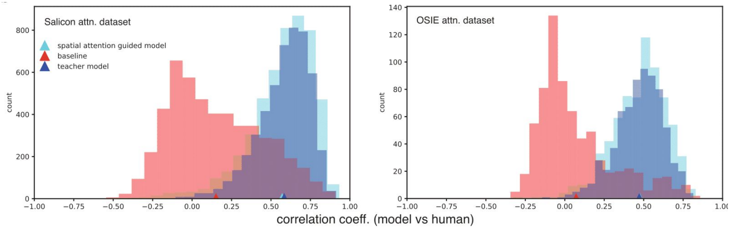
$$L_{contra} = -\sum_{i,j} \log \frac{\exp(sim(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} 1_{k \neq i} \exp(sim(z_i, z_k)/\tau)}$$

$$L_{attn} = \sum_i (\lambda KLD(m_i^p, m_i^t) - \beta NSS(m_i^p, p_i^t))$$

*NSS/KLD are typical loss for calculating saliency distance.

Results: Attention alignment between model and human





Summary:

- Baseline model is less correlated to human attention
- Spatial attention guided models are highly predictive of human attention

Results: Classification task

Model	Accuracy (%)
Contrastive	67.61±0.04
Contrastive attn. teacher	**68.23±0.08**
Contrastive attn. co-train	66.35±0.12
Supervised	75.91±0.10
Supervised attn. teacher	76.02±0.04
Supervised (ResNet-18)	69.17±0.07
Supervised (ResNet-18) attn. teacher	69.30±0.04

Summary

- Human spatial attention improves the SSL model's performance with teacher model.
- Human spatial attention also improves the SL model's performance but the gain is smaller
- Gain is smaller when using human spatial attention directly on SSL (co-train)

Reason:

- Contrastive model's representation is more general as the human attention collected is not task-specific for teacher model.
- Teacher model generalize its knowledge on human attention beyond the limited ground truth human attention data.

Results: Robustness

Model	Speckle Noise	Gaussian Blur	Spatter	Saturate
Contrastive	28.23±0.31	26.16±0.07	43.08±0.18	60.42±0.15
Contrastive attn. teacher	**29.15±0.65**	**27.10±0.35**	**44.04±0.08**	**60.50±0.02**

Image classification accuracy on ImageNet-C

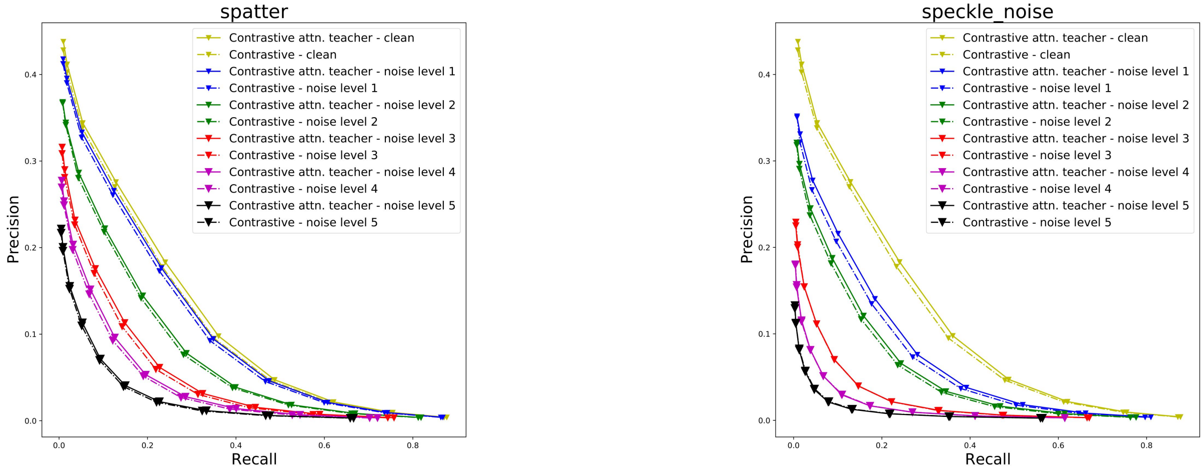


Image retrieval PR curve on ImageNet-C

Summary

- We provided a teacher model trained from scratch that can be used to generate pseudo-saliency labels for large data set
- Spatial attention guided models are highly predictive of human attention
- Spatial attention guided models are more accurate and robust than baselines