Do Invariances in Deep Neural Networks Align with Human Perception?

Human invariances?

How do we measure alignment between model and invariance to $\ell_2$?

Eg: adversarial robustness tests today are special cases. Robustness benchmarks used.

Ideal Robustness Evaluation

In order for models to be robust, they must learn invariances in unexpected ways.

Lack of human-like invariances makes models fail in surprise ways.

Learning the Right Invariances is Necessary for Robustness

$\|f(x) - f(x')\| = \|f(x) - f(y)\|$ ✔

$\|f(x) - f(x')\| = \|f(y) - f(y')\|$ ✗

Our finding: residual connections + adversarial data augmentation + self-supervised training w contrastive loss = higher alignment.

We identify challenges and highlight the key role of loss used to generate identically represented inputs in measuring alignment with human perception.

We propose an improved method to measure alignment that does not require labelled data and can scale well.

Our Proposal for Measuring Alignment

Representation Inversion

Mahendran & Vedaldi, 2015

$\arg\min_{x'} \|x - x'\|_2$

$\mathcal{L}(x|x') = \|m(x) - m(x')\|_2 + \lambda \|\mathcal{R}(x)\|_2$

*For scalability, we show that we can use perceptual distance LPIPS (Zhang et al., 2018) to automate these judgments.

Prior works on measuring alignment have seemingly contradictory takeaways since they do not explicitly engage with the choice of $\mathcal{R}$.

Our finding: human-Aligned $\mathcal{R}$? Adversarial $\mathcal{R}$? Regularizer-free $\mathcal{R}$?

Regularizer-free

Human-Aligned

Adversarial

Importance of Loss Used to Generate Identically Represented Inputs

Regularizer-free

Human-Aligned

Adversarial

Our finding: We identify challenges and highlight the key role of loss used to generate identically represented inputs in measuring alignment with human perception.

What Components of DL Pipeline Contribute to Alignment?

Prior works on measuring alignment have seemingly contradictory takeaways since they do not explicitly engage with the choice of $\mathcal{R}$.

Our finding: Human-Aligned $\mathcal{R}$? Adversarial $\mathcal{R}$? Regularizer-free $\mathcal{R}$?

Regularizer-free

Human-Aligned

Adversarial

Importance of Loss Used to Generate Identically Represented Inputs

Regularizer-free

Human-Aligned

Adversarial

Our finding: We identify challenges and highlight the key role of loss used to generate identically represented inputs in measuring alignment with human perception.

What Components of DL Pipeline Contribute to Alignment?

Prior works on measuring alignment have seemingly contradictory takeaways since they do not explicitly engage with the choice of $\mathcal{R}$.

Our finding: Human-Aligned $\mathcal{R}$? Adversarial $\mathcal{R}$? Regularizer-free $\mathcal{R}$?

Regularizer-free

Human-Aligned

Adversarial

Importance of Loss Used to Generate Identically Represented Inputs

Regularizer-free

Human-Aligned

Adversarial

Our finding: We identify challenges and highlight the key role of loss used to generate identically represented inputs in measuring alignment with human perception.

What Components of DL Pipeline Contribute to Alignment?

Prior works on measuring alignment have seemingly contradictory takeaways since they do not explicitly engage with the choice of $\mathcal{R}$.

Our finding: Human-Aligned $\mathcal{R}$? Adversarial $\mathcal{R}$? Regularizer-free $\mathcal{R}$?

Regularizer-free

Human-Aligned

Adversarial

Importance of Loss Used to Generate Identically Represented Inputs

Regularizer-free

Human-Aligned

Adversarial

Our finding: We identify challenges and highlight the key role of loss used to generate identically represented inputs in measuring alignment with human perception.

What Components of DL Pipeline Contribute to Alignment?

Prior works on measuring alignment have seemingly contradictory takeaways since they do not explicitly engage with the choice of $\mathcal{R}$.

Our finding: Human-Aligned $\mathcal{R}$? Adversarial $\mathcal{R}$? Regularizer-free $\mathcal{R}$?

Regularizer-free

Human-Aligned

Adversarial

Importance of Loss Used to Generate Identically Represented Inputs

Regularizer-free

Human-Aligned

Adversarial

Our finding: We identify challenges and highlight the key role of loss used to generate identically represented inputs in measuring alignment with human perception.