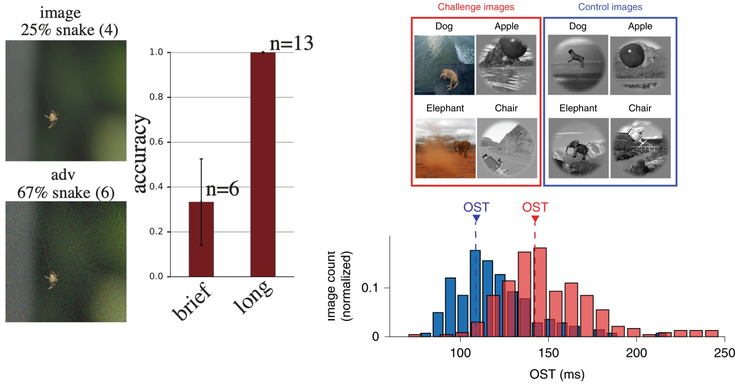
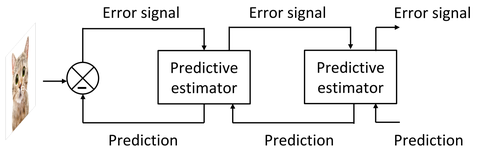
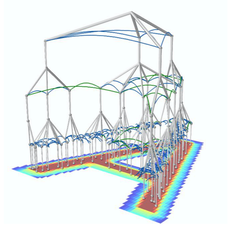
Motivation

- Gaps between human visual perception and artificial neural networks
  - Amount of data needed to learn
  - Robustness (corruption, adversarial, etc.)
- Recurrent feedback in the human brain
  - Longer object solution time for challenging images
  - Adversarial samples fool humans under limited time



- Modeling tools
  - Computational neuroscience: Predictive coding, etc.
  - Control: Kalman filters, Feedback control, etc.
  - Deep learning: generative models, attention mechanism, etc.

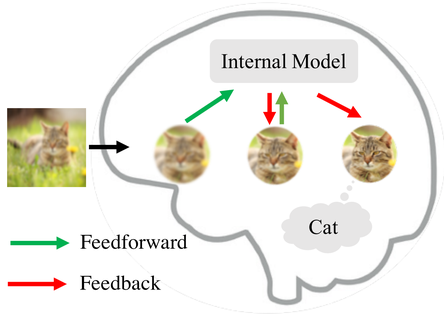




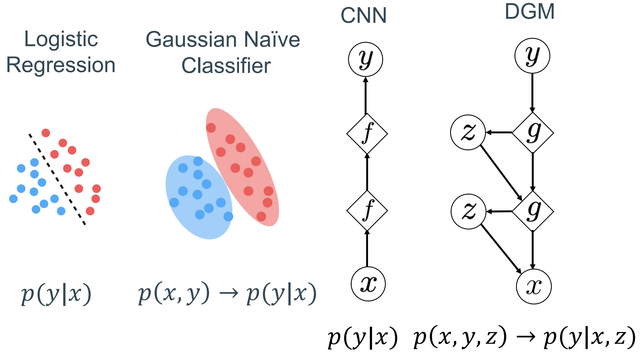
Goal: Introduce biologically inspired components into deep learning models in a principled way to create generalizable AI.

Neural networks with recurrent generative feedback

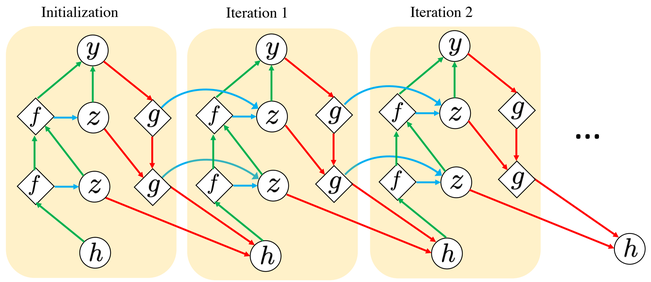
- Self-consistency



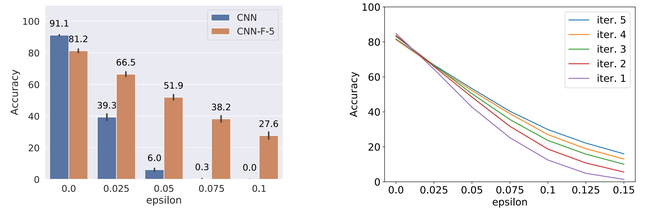
- Generative Classifier



- Iterative inference and online updates in CNN-F

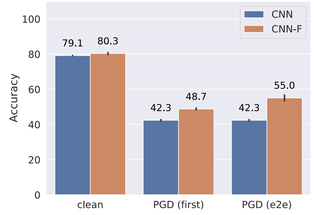


- Generative feedback promotes robustness



- Standard training on Fashion-MNIST improve adversarial robustness
- More iterations improves accuracy against larger perturbation magnitude.

- Adversarial training

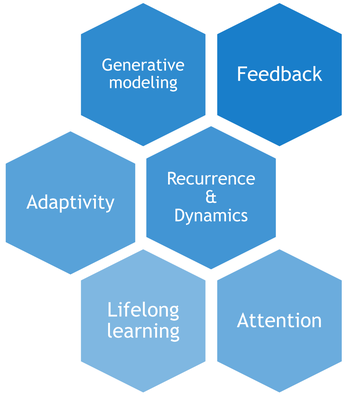


- Adversarial training on CIFAR-10.
- Attack the first forward pass (first) or the last forward output (e2e).
- Further improves robustness of CNN (ϵ=8/255).

Future work

We would like a model to:

Learn efficiently	• Quickly incorporate new data • Extract task-relevant information
Robust	• Robust to unseen corruptions • Fails gracefully with interpretation
Compositionality	• Understand the world • Generalization and creativity
Uncertainty quantification	• Likelihood or typicality estimation



References

Gamaleldin F. Elsayed et al. (2018) "Adversarial Examples that Fool both Computer Vision and Time-Limited Humans." In NeurIPS.

Kohitij Kar et al. (2019) "Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior." In Nature Neuroscience

Dileep George et al. (2017) "A generative vision model that trains with high data efficiency and breaks text-based CAPTCHAs." In Science

Andrew Ng et al. (2002) "On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes." In Neurips.

Tan Nguyen et al. (2018) "A Bayesian perspective of convolutional neural networks through a deconvolutional generative model." arXiv:1811.02657.

Yujia Huang et al. (2020) "Neural networks with recurrent generative feedback." In Neurips.