# Learning-to-Learn with Neural Networks

**Reuben Feinman (reuben.feinman@nyu.edu)**
Center for Neural Science
New York University

**Brenden M. Lake (brenden@nyu.edu)**
Department of Psychology and Center for Data Science
New York University

### Abstract

The abstract should be one paragraph, indented 1/8 inch on both sides, in 9 point font with single spacing. The heading "**Abstract**" should be 10 point, bold, centered, with one line of space below it. This one-paragraph abstract section is required only for standard six page proceedings papers. Following the abstract should be a blank line, followed by the header "**Keywords:**" and a list of descriptive keywords separated by semicolons, all in 9 point font, as shown below.

**Keywords:** learning-to-learn; neural networks; inductive biases

## Introduction

Deep neural networks (DNNs) are machine learning techniques that impose a hierarchical architecture consisting of multiple layers of nonlinear processing units. In practice, DNNs achieve state-of-the-art performance for a variety of generative and discriminative learning tasks from domains including image processing, speech recognition, drug discovery and genomics.

Although DNNs are known to be robust to noisy inputs, they have been shown to be vulnerable to specially-crafted adversarial samples. These samples are constructed by taking a normal sample and perturbing it, either at once or iteratively, in a direction that maximizes the chance of misclassification. Figure 1 shows some examples of adversarial MNIST images alongside noisy images of equivalent perturbation size. Adversarial attacks which require only small perturbations to the original inputs can induce high-efficacy DNNs to misclassify at a high rate. Some adversarial samples can also induce a DNN to output a specific target class. The vulnerability of DNNs to such adversarial attacks highlights important security and performance implications for these models. Consequently, significant effort is ongoing to understand and explain adversarial samples and to design defenses against them.

## Experiments

This is where experiment information will go.

## Acknowledgements

This research was supported by a Henry M. McCracken fellowship and NYU start-up faculty funding. We thank Subhankar Ghosh for useful code and preliminary experiment ideas.

The source code repository for this paper can be found at `http://github.com/rfeinman/learning-to-learn`

## References

Ritter, S., Barrett, D. G. T., Santoro, A., & Botvinick, M. M. (2017). Cognitive psychology for deep neural networks: a shape bias case study. In *Proceedings of the 34th international conference on machine learning (icml '16)* (pp. 2940–2949).