CS 4644/7643: Deep Learning
Project Proposal

Instructor: Danfei Xu

TAs: Aditya Singh (Head TA), Amogh Dabholkar, Yash Jakhotiya, Kyunhun Lee, Anshul Ahluwalia, Zachary Minot, Aaditya Singh, Charlie Gunn, Anshul Gupta, Ningyuan Yang

Discussions: https://piazza.com/gatech/fall2022/cs46447643

Due: Tuesday, September 27, 11:59pm

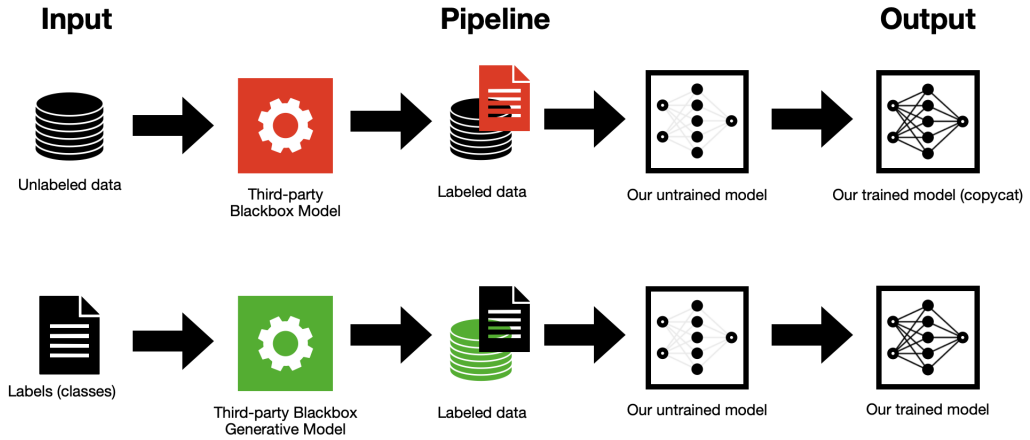
Team members: Andrea Covre, Connor Reitz, Jake Hopkins

Project Summary

The overarching topic of this project is model inference: can we infer a model by training it on data-sets classified by another black-box model? Ex: Model A predicts an image's label with 97 percent accuracy. Use Model A with unclassified data to produce classified data, and train Model B with this data and measure accuracy. As a baseline we could first use similar architectures to Model A to measure accuracy and then also consider different architectures afterwards.

Model inference has many interesting outcomes. On one hand, this could be considered model theft. Instead of paying for a model as a service, an unethical use case could be to create your own model by inferring this model and subsequently being able to use your own model for free. If model theft is shown to produce accurate results, further research could be done into prevention of model theft. On the other hand, this model inference could be used to cheaply classify data through using online, free-use models (such as DALLE-mini). This could help research areas or learning/AI startups that need a lower cost barrier to entry, as getting large, classified data-sets is a large cost that these groups face.

Proposed Methodology/Approach



We will have two cases:

- **Model is trained on data labeled by a third-party model**
  - Method:
    we will have unlabeled data collected from some source which will

be labeled by a pre-existing model. This data will then be used to train our models with different architectures.

  - Evaluation:
    we will compare the performance of our models against the model that originally labeled the data on a dataset unseen by all models.

- **Model is trained on data generated by a third-party model given the labels**

  - Method:
    we will have a list of classes (CIFAR-10 classes) which will be used to mine corresponding data samples from a pre-existing generative model (DALLE-mini). This data will then be used to train our classification models with different architectures.
  - Evaluation:
    we will compare the performance of our models trained on synthetic data against equivalent models trained on proper datasets.

Resources/Related Work

Resources

- Craiyon (ex DALL-E mini) to mine AI generated images
  https://www.craiyon.com
- PyTorch_CIFAR10 for a variety of pre-trained and trainable CIFAR-10 models
  https://github.com/huyvnphan/PyTorch_CIFAR10
- MNIST dataset
  https://www.kaggle.com/competitions/digit-recognizer/data
- Online MNIST Classifier
  https://cs.stanford.edu/people/karpathy/convnetjs/demo/mnist.html

Related Work

- J. Tremblay et al., "Training Deep Networks with Synthetic Data: Bridging the Reality Gap by Domain Randomization," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2018, pp. 1082-10828, doi: 10.1109/CVPRW.2018.00143.
  https://arxiv.org/pdf/1804.06516

- Tramèr, F., Zhang, F., Juels, A., Reiter, M.K., Ristenpart, T. (2016). Stealing Machine Learning Models via Prediction APIs. ArXiv, abs/1609.02943. https://arxiv.org/abs/1609.02943

- He, Y., Meng, G., Chen, K., Hu, X., He, J. (2019). Towards Privacy and Security of Deep Learning Systems: A Survey. arXiv preprint arXiv:1911.12562. https://arxiv.org/abs/1911.12562

Datasets

First, as a POC of our attempt to construct a model, we will attempt to steal a model by using the MNIST dataset. We've heard it been called the de facto 'Hello World!' of Machine Learning, so it seems only right for us to use this dataset here. If we can effectively re-create an online model using this dataset, we can try with some more complicated data.

Using Dall-E, we can essentially create a unique, infinite dataset of whatever images we wanted. We will try to generate car and plane images and use these on another online model to label them and essentially re-engineer that model. While Dall-e access may be difficult to be granted, we can use the Dall-e mini and repeatedly pull the images to create our dataset (with an automated process).