THEORY OF MACHINE LEARNING

LECTURE 22

NTK SUMMARY, REPRESENTATION LEARNING

ANNOUNCEMENTS

- **Homework 3 due on Monday April 11**
- **Project presentations:** starting in two weeks! (~18 projects)
  - Dates: April 19, 22, 26 (5 projects /class), couple online
- This week and next: representation learning, robustness

LAST WEEK

- Learning NNs is hard, often done via gradient descent
  - Topics skipped – "strongest" hardness results based on crypto [Klivans, Sherstov]

- Analyzing gradient descent:
  - Can one analyze dynamics of gradient descent? [can view as Kernel regression for a time-varying kernel]
  - Are there cases where we can reason about resulting solution? [for infinitely wide nets, kernel remains "fixed" – neural tangent kernel]

NTK REVIEW

Theorem. [Jacot, Gabriel, Hongler 18] [Arora, et al. 2019] A width ~ n^3 network (any number of layers) trained via GD from random initialization achieves zero training error. Moreover, the final solution is equivalent to solving a "Kernel regression" problem with a specific kernel.

- Kernel regression
- Any model training can be viewed as Kernel regression with time varying kernel
- With wide DNNs, kernel doesn't change much!

NTK EXPERIMENTS

[Arora, et al. 2019] What happens if we forget about NNs, compute closed form for NTK (determined only by number of layers, types of connections, activation function), perform kernel regression?

Depth	CNN-V	CNTK-V	CNN-GAP	CNTK-GAP
3	61.97%	64.67%	57.96%	70.47%
4	62.12%	65.52%	80.58%	75.93%
6	64.03%	66.03%	80.97%	76.73%
11	70.97%	65.90%	75.45%	77.43%
21	80.56%	64.09%	81.23%	77.08%

REPRESENTATION LEARNING

- General idea – neural networks are "hierarchical feature extractors"
- Circa 2000s – manual feature extraction (HOG, SIFT)
- NNs embed inputs -> "feature space" (alternative 'representation')

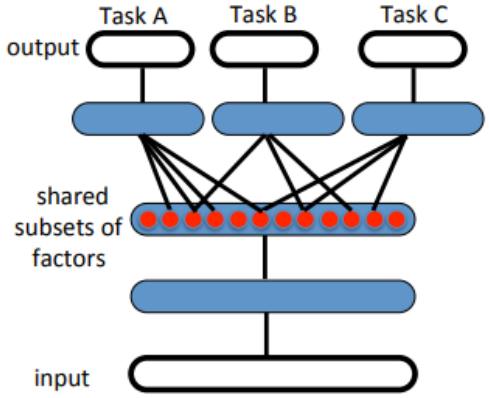

Fig. 1. Illustration of representation-learning discovering explanatory factors (middle hidden layer, in red), some explaining the input (semi-supervised setting), and some explaining target for each task. Because these subsets overlap, sharing of statistical strength helps generalization.

REPRESENTATION LEARNING

- What makes a good representation?
  - Contrastive (for classification)
  - "Disentangled" or orthogonal
  - Sparse "explanations" for phenomena
  - Hierarchically organized, explanatory

- Supervised vs Unsupervised

REPRESENTATION LEARNING

- Unsupervised learning of representation
  - Sparse coding / autoencoders
  - Self-supervision