Analyzing the Generalization Error of Multimodal Deep Networks

15-300, Fall 2019

Shane Guan ⟨xguan⟩

November 2, 2019

1 Project Webpage

Tentatively, you can find the project webpage here.

2 Project Description

2.1 Who

I will be working with Profs LP Morency and Zico Kolter, as well as their students Paul Liang and Vaishnavh Nagarajan.

2.2 What

A modality is a type of sensory input, like vision, hearing, taste, lidar, etc. As can be imagined, robots frequently view the world through multiple modalities, so the field of machine learning on multiple modalities is very applicable there.

Suppose you wished to train a robot to lip-read. Clearly, you would have to equip the robot with some sort of camera, as the task of lip-reading requires vision. However, while you are training

the robot, you might wonder if you can somehow speed up the training process by equipping the robot with ears, nose, and other sensors during training time (but still removing them at test time). Would the robot learn more about lip-reading through these other sensors? Would it perform better at test time when every sensor except the camera is removed, than if it were only trained using the camera?

In 2011, it was shown in the landmark paper "Multimodal Deep Learning" that, for the algorithm the authors designed for speech recognition, it performed better at lip-reading when it was trained on both audio and vision data (Ngiam, et al). Since then, there have been many other papers reporting similar benefits on unimodal tasks by training on multimodal data, such as a 2016 paper that showed that for one particular language model, it was better to also train on vision data (Kottur, et al).

What is missing is a theoretical understanding of when these multimodal benefits happen, like if they depend on the algorithm under study or on peculiarities of the dataset. I aim to fill in these gaps by deriving upper bounds on the generalization error (the error on the test set is one way to approximate this). I first aim to derive bounds on simple models and simple assumptions on the underlying data distribution. Then I will verify that these bounds work in practice by empirically measuring the derived bounds and seeing if they match up with the actual test error.

2.3 So What

Everyday we are pushing out more and more machine learning models for use in production in the world. If we do not understand on a theoretical basis why they work, then they can and will fail in subtle. In fact, there have already been several studies that showed how to attack popular state-of-the-art models using adversarial techniques (here's one from CMU). We need to have an understanding of when and how our models work, so we can provide guarantees on how often they

fail.

In addition, once we understand the cases in which various neural architectures work, this can possibly lead to insights into how to improve the network.

3 Project Goals

3.1 100%

Ideally, if everything goes according to plan, I will have derived a uniform convergence generalization bound on some multimodal network, and performed experiments to compare how the bound compares to the actual test error.

3.2 75%

If things go more slowly than expected, I will still aim to have empirically tested some of the hypothesis in the multimodal literature about the generalization error.

3.3 125%

If things go faster than expected, I might do the following. If the bounds I derived work in practice, I might use it to propose a new regularization term and then will empirically verify how well it works. If the uniform convergence bound I derived doesn't work in practice, I will try to derive a stability bound on networks, as recommended by a recent paper that suggested any uniform convergence bound on the test error of deep networks is vacuous (Nagarajan 2019).

4 Milestones

4.1 First Technical Milestone for 15-300

I will get up to speed on the various ways and algorithms for which training on multimodal improves the unimodal performance. Through this process, I will acquaint myself with the various hypothesis currently out there to explain this phenomenon.

I will also read up on popular techniques to analyze the generalization error, including the Rademacher complexity, VC dimension, PAC bounds, and covering number. This is so that I may understand the state-of-the-art and use them to derive my own bounds on the generalization error.

4.2 Biweekly Milestones for 15-400

I have identified the following sub-milestones for next semester.

- Jan 27, I will have understood in broad strokes what the framework for my analysis is, and which assumptions I am going to make about the multimodal situation.
- Feb 10, I will have made a sketch of my proof, perhaps appealing to intuition.
- Feb 24, I formalize the intuition and the proof
- Mar 16, I empirically test out the theory on toy datasets like MNIST and CIFAR-10 (must be multimodal), and compare it to previous derived bounds
- Mar 30, I test it on ImageNet and other more realistic datasets
- Apr 13, I write a first draft of a paper summarizing my results
- Apr 27, I make the paper camera ready.

5 Literature Search

What papers have you collected and/or read to help you in your project? Are you missing anything?

- Large Margin Deep Networks for Classification (read)
- On the depth of deep neural networks, a theoretical view (read)
- Robust Large Margin Deep Neural Networks (read)
- Uniform convergence may be unable to explain generalization in deep learning (read)
- Learnability, Stability, and Uniform Convergence
- Generalization Bounds for Universally Stable functions
- On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities (VC theory paper)
- An Information Theoretic Framework for Multi-view Learning
- Combining Labeled and Unlabeled Data with Co-Training
- Multi-view Learning Overview: Recent Progress and New Challenges
- Co-training and expansion: Towards bridging theory and practice

6 Resources Needed

For the experimental analysis of the generalization bounds, I will need a GPU cluster. I believe through my mentors I will have this covered.

7 References

7.1

Ngiam, Jiquan, et al. "Multimodal deep learning." Proceedings of the 28th international conference on machine learning (ICML-11). 2011.

7.2

Kottur, Satwik, et al. "Visual word2vec (vis-w2v): Learning visually grounded word embeddings using abstract scenes." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

7.3

Nagarajan, V., & Kolter, J. Z. (2019). Uniform convergence may be unable to explain generalization in deep learning. arXiv preprint arXiv:1902.04742.