Tradeoff between Accuracy and Robustness in Multimodal Learning
15-300, Fall 2020

Yun (Catherine) Cheng

November 6, 2020

1 Project Description

I will be working with Professor Louis-Philippe Morency in the Language Technologies Institute and his graduate student Paul Liang in the Machine Learning Department on the tradeoff between accuracy and robustness in Multimodal Learning. We will first identify sources of unrobustness. Then on this basis, we will work on developing evaluation metrics for robustness in building a standardized benchmark. Finally, we will evaluate and analyze existing methods using this benchmark.

The problem we aim to solve first is to identify the sources of unrobustness by investigating the model's behaviors on artificial noise from different modalities. Adversarial examples are input data modified with human-imperceptible changes. However, it may strongly affect the prediction results because the model's decisions rely on one or a few insignificant features. Thus, the unpredictable errors that a model exhibits in facing adversarial examples demonstrate the unrobustness of the model.

We will experiment with various types of imperfect data. We are interested in both unimodal and multimodal noises. For perturbations in language processing, misspelling, grammatical mistakes, and drop-out words are common. In visual tasks, the noises can be in the form of blurred, distorted, rotated, and transformed images. For testing multimodal noises, we will focus on the potential correlation between noises from different modalities and their combined effects on the performance of the model.

After that, our goal is to integrate the evaluation metrics of robustness into a standardized benchmark for multimodal evaluation. Existing benchmarks are too specific to individual tasks. Furthermore, when comparing a model to the benchmark, we cannot tell whether the improvement comes from a unimodal or multimodal feature. Thus, in this project, we attempt to build a benchmark that can load multimodal data easily, make multimodal predictions, and evaluate different metrics of a model. The ultimate ideal benchmark will access performance, complexity, robustness, and fairness.

The potential contribution of this project to the area of Multimodal Learning is two-folded. On one side, the robustness issue is common in training multimodal models because of imperfect data. A robust model can tolerate missing modalities via translation - the ability to infer the missing modalities from the learned modality. On the other side, the methods are currently evaluated in a task-specific manner. Therefore, the proposal of a standardized benchmark will effectively save time and effort in re-exploring another domain for similar methods. The challenges of this project will include balancing accuracy and robustness in making the evaluation metrics. As previous studies suggest, we cannot attain accuracy and robustness at

the same time. Therefore, when making the evaluation metrics for robustness, it is essential to take account of trading off robustness against accuracy. To that end, we will find or characterize the desired properties that balance this tradeoff.

2 Project Goals

2.1 75% Project Goal

- Identify sources of unrobustness in language, video, and audio processing tasks.
- Test with unimodal and multimodal noises.
- Study the correlation between noises from different modalities.

2.2 100% Project Goal

- Make evaluation metrics for robustness.
- Propose a standardized benchmark for multimodal evaluation using the evaluation metrics.
- Study the tradeoff between accuracy and robustness and characterize the desired properties that balance this tradeoff.

2.3 125% Project Goal

- Evaluate existing methods using the proposed benchmark.
- Perform fine-grained analysis on undesired performance of the evaluated methods.

3 Project Milestones

3.1 First Technical Milestone

By the last day of this semester in 15-300, I should have read a few papers about the tradeoff between accuracy and robustness in unimodal and multimodal learning as well as previous works that attempt to address the robustness issue. I should be familiar with relevant terminologies in studying the robustness issue, recent challenges in balancing tradeoff between accuracy and robustness, and attempted solution to these challenges.

3.2 First Biweekly Milestone: February 1st

By the first biweekly milestone, I hope to have some idea about sources of unrobustness in language, video and audio processing tasks by experimenting with unimodal and multimodal noises. I hope to compare the significance of these artificial noises in determining the robustness of a model.

3.3 Second Biweekly Milestone: February 15th

By the second biweekly milestone, I hope to gain understanding in how to balance the tradeoff between accuracy and robustness in the context of multimodal learning. I also hope to find a few key desired properties that are shared among robust models.

3.4 Third Biweekly Milestone: March 1st

By the third biweekly milestone, I hope to propose the first evaluation metrics of robustness based on my understanding of robustness in theory. I will also test my hypothesized metrics on models with and without known robustness issue to see if they are good candidates in differentiating robust from unrobust models.

3.5 Fourth Biweekly Milestone: March 22nd

By the fourth biweekly milestone, I anticipate to improve the first evaluation metrics with more theoretical and empirical results.

3.6 Fifth Biweekly Milestone: April 5th

By the fifth biweekly milestone, if I have obtained a set of reasonable evaluation metrics, I hope to integrate them into the proposed standardized benchmark. If the evaluation metrics are still less than desired, I should continue working on improving them.

3.7 Sixth Biweekly Milestone: April 19th

By the sixth biweekly milestone, I hope to have made decent progress in the standardized benchmark. I will expect other glaring issues exposed by testing in this stage, so I will work on addressing them.

3.8 Seventh Biweekly Milestone: May 3nd

By the seventh and final biweekly milestone, I hope I have almost (if not) finished the benchmark. If time permits, we will compare existing methods to the benchmark and analyze the cause for the unrobustness issues within these methods.

4 Literature Search

As of this writing, as recommended by my advisor and mentor, I have started some reading on relevant works in the tradeoff between accuracy and robustness in a general setting. These works are cited as following [1],[2],[3],[4],[5],[6]. I will continue reading about robustness issue in such speicific areas of multimodal application as Visual Question Answering (VQA) and corresponding attempted solutions. These works are cited as following [7],[8],[9].

5 Resources Needed

We will be using Python and the deep learning framework Keras and Pytorch.

References

[1] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy, 2017.

[2] Paul Pu Liang, Zhun Liu, Yao-Hung Hubert Tsai, Qibin Zhao, Ruslan Salakhutdinov, and Louis-Philippe Morency. Learning representations from imperfect time series data via tensor rank regularization, 2019.

[3] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy, 2019.

[4] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy, 2019.

[5] Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabas Poczos. Found in translation: Learning robust joint representations by cyclic translations between modalities, 2020.

[6] Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John Duchi, and Percy Liang. Understanding and mitigating the tradeoff between robustness and accuracy, 2020.

[7] Jia-Hong Huang, Modar Alfadly, Bernard Ghanem, and Marcel Worring. Assessing the robustness of visual question answering, 2019.

[8] Xiujun Li, Chunyuan Li, Qiaolin Xia, Yonatan Bisk, Asli Celikyilmaz, Jianfeng Gao, Noah Smith, and Yejin Choi. Robust navigation with language pretraining and stochastic sampling, 2019.

[9] Jia-Hong Huang, Cuong Duc Dao, Modar Alfadly, and Bernard Ghanem. A novel framework for robustness analysis of visual qa models, 2018.