Deep Learning under Adversarial Attacks and Generative Models

1. Outline

Adversarial attacks in machine learning are a set of techniques which attempt to fool models through malicious input. This technique is applied to a variety of tasks viz. attack or cause a malfunction in standard machine learning based models. A well cited example of such an attack in a deep learning based image recognition system, is addition of noise to an image of a primate which leads to its misclassification as a dog. Deep learning is a genre of machine learning algorithms that attempt to solve tasks by data driven modelling of abstractions following a stratified description paradigm using non-linear transformation architectures. Put in simple terms, a deep learning system for classifying primate vs. dogs would be constructed using a large collection of labelled images of primates and dogs, fed to a multi-layer deep neural network (DNN), computing the error in its inference, and backpropagating the error through the DNN in order to update the free parameters in it to minimize the inference error. Intuitively, the earlier layers in the DNN that are close to the image would start with aggregating pixel wide information using multiple attributes termed as low level features viz. edges, gradients, etc.; which would be hierarchically aggregated to obtain complex features viz. blobs, lines, curves, etc.; that are further aggregated to make a decision in favor of presence or absence of an object. Being entirely data driven, while these hierarchical abstractions are not always intuitively explainable, it turns out that more often than not, adversarial attacks on these models tend to make the models behave very erratically which cannot be humanly explained viz. in the cited case where the addition of not so visually corruptive very low noise tends to make the model classify a primate as a dog.

While these challenges persist in destabilizing deep learning based systems when deployed in practical scenario, recent research associated with generative modeling has shown the affirmative might of using such attacks in training robust deep learning system. The general perception is to associate adversarial learning only with generative modelling, while a lot of contributions essentially use adversarial learning in order to address the perception-distortion tradeoff in cost functions. This has helped us generate images from random vectors using generative adversarial networks (GAN), develop single image super-resolution algorithms (Deep SR GAN), adversarial transformation to images, semantic segmentation under adversarial losses.

This tutorial will focus on understanding the perception-distortion tradeoff and its mathematical constructs with respect to a loss function. Subsequently this understanding would be used to discern the computational mechanics of implementing adversarial losses within generative models, regression learning problems like super-resolution and image-to-image transformation, and classification problems like semantic segmentation. The information would be delivered through standard lectures and intertwined with some hands-on Python based implementations which can be carried out by participants on their standard laptops. Setup instructions would be provided to the audience prior to the tutorial sessions.

Relevance to ISBI audience

Deep learning has been observed to be steadily growing for solving problems in the field of biomedical and biological imaging and image analysis with the number of articles authored at leading conferences like ISBI growing at an overwhelming 8-10 folded rate per year. ISBI 2018 onwards has been witnessing some interesting contributions using adversarial learning for lesion generation, image transformation, stain deconvolution, computed super-resolution, 2D to 3D image transformation, semantic segmentation, etc. This growth in academic interest is also sustained industrially with players in the medical imaging sector gaining steady interest in using adversarial deep learning for solving their problems as well. While we see some of the adversarially trained deep learning models being industrially employed, the sheer volume of data to be processed, the multiscale nature of medical imagery, and the multiple update rules

and intermittently switching loss functions to be implemented, challenges many young researchers from getting started in the field of adversarial deep learning for biomedical and biological imaging and image analysis. Furthermore this also needs a good understanding of the inherent mathematical constructs of associated with its signal flow.

Classically, learning has been a playfield in computer sciences where machine vision has been an early adopter of most of these developments. However, the image formation process and appearance models being quite different for biomedical and biological images, the field has more early adopters willing to undertake a PhD in the area, who are from generally biomedical sciences, electrical sciences, physics, biology and medicine, and are not well trained in machine learning and algorithmic nuances as an adopters with early degree in computer sciences. This has led to the prevailing confusion and reluctance to the trust extended to uptake of deep learning in the community, as has been witnessed in the machine vision, speech and text analytics and data mining communities.

This tutorial is aimed at introducing enthusiastic early adopters to the following concepts: (i) notion of hierarchical embeddings in biomedical and biological images, and understanding hierarchical knowledge transfer in a DNN, (ii) nature of loss functions and the differences between distortion and perception losses, (iii) some common use of perception losses while learning with adversarial attacks viz. generative models, super-resolution, semantic segmentation, (iv) explanations on stability and generalizability of learning with adversarial attacks including learnable perception loss functions.

2. Description of material to be covered (4000 characters)

1. **Introduction to the concept of learning hierarchical embeddings in medical images using deep neural networks.** This part of the talk would include summary of works which describe classical approaches of learning DNNs and the 3 primary aspects of ML which include (a) handling data and challenges associated with it, (b) complexity of creating a model for the need for hierarchical embedding, (c) loss functions used for optimizing performance with update of weights and understanding of the difference between distortion and perception loss.
2. **Adversarial attacks in machine learning.** This part of the talk would tell the story of some of the classical adversarial attacks which are used to test the integrity and reliability of a DL based system.
3. **Adversarial autoencoders (AAE) and Generative adversarial networks (GAN).** This part of the talk will discuss the concept of generating images from random variables, explain the parallels between generative networks and autoencoders, distribution modelling and learning distribution of latent variables. This would also be accompanied with Python based hands on practice tutorials for generating pathologies viz. lesions.
4. **Discriminators for modelling adversarial losses - Vanilla - Relativistic - Turing Test discriminator.** This part of the talk would focus on different types of practical implementation of discriminators, and how the learning dynamics changes across these different types of implementation. This would also mathematically explain the difference between distortion losses like mean square error (MSE), cross entropy loss (CE Loss) and perception losses learnt with discriminators.
5. **Practical implementation of adversarial learning.** This part of the talk will discuss some of the recent contributions to (a) single image super-resolution using deep neural networks viz. for optical microscopy, (b) image to image transformation viz. ultrasound image simulation via. generative transformation, (c) semantic segmentation viz. lesion segmentation in MRI volumes with accurate boundary delineation capability.

3. Contact Information and Biosketch of Presenter

Debdoot Sheet, *PhD, SMIEEE*
Assistant Professor, Department of Electrical Engineering and Centre for Artificial Intelligence
Indian Institute of Technology Kharagpur
Kharagpur, WB 721302, India

Email: debdoot@ee.iitkgp.ac.in
Phone: +91 3222 283 082 (Office), +91 94740 00086 (Mobile)
url: www.facweb.iitkgp.ac.in/~debdoot/

Biosketch:

Dr. Debdoot Sheet is an Assistant Professor of computational medical imaging and machine learning with core appointment at the Department of Electrical Engineering and associate appointment at the Centre for Artificial Intelligence at the Indian Institute of Technology Kharagpur, India and founder of medical imaging AI on the cloud company SkinCurate Research. He was born in Kharagpur, India in 1986 and has spent his life across Kharagpur and Kolkata in India and Munich in Germany.

Dr. Sheet received the B. Tech degree in electronics and communication engineering in 2008 from the West Bengal University of Technology, Kolkata. He obtained the MS and PhD degrees from the Indian Institute of Technology Kharagpur in 2010 and 2014 respectively. His current research interests include computational medical imaging, machine learning, deep learning, fairness- accountability- trust-explainability (FATE) of artificial intelligence (AI). He is also a DAAD alumni and was a visiting scholar at the Technical University of Munich during 2011-12. He is also a recipient of the IEEE Computer Society Richard E. Merwin Student Scholarship in 2012, the Fraunhofer Applications Award at the Indo-German Grand Science Slam in 2012, Won the GE Edison Challenge 2013, IEM Kolkata Distinguished Young Alumnus Award 2016. He is a senior member of IEEE, member of SPIE and ACM, life members of BMESI and IUPRAI, and serves as Regional Editor of IEEE Pulse since 2014.

He has been closely associated with ISBI, attending regularly since 2013 and had also organized the Deep Learning tutorial and special session in ISBI 2016. His group regularly organizes Deep Learning workshops and topical schools across South Asia and Europe, and was also instructor to the 2 popular NPTEL MOOCs on Medical Image Analysis [1] and Deep Learning for Visual Computing [2], which are also available through YouTube and Github.

[1]. https://www.youtube.com/playlist?list=PLi7vCu7jEp8_nFoyZ-8exq5UYW_CAZ6zM
[2]. https://www.youtube.com/playlist?list=PLuv3GM6-gsE1Biyakccxb3FAn4wBLyfWf

During 2018-2019 his group has contributed across image generation, ultrasound simulation, compression, semantic segmentation, visual hashing using adversarial deep learning, for achieving performance beyond limits known in the past. This tutorial would be based on some of the related works presented in:

1. China, D., Tom, F., Nandamuri, S., Kar, A., Srinivasan, M., Mitra, P., & **Sheet, D.** (2019, April). UltraCompression: Framework for High Density Compression of Ultrasound Volumes using Physics Modeling Deep Neural Networks. In 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)
2. Tom, F., Sharma, H., Mundhra, D., Dastidar, T. R., & **Sheet, D.** (2019). Learning a Deep Convolution Network with Turing Test Adversaries for Microscopy Image Super Resolution. In 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)

3. Saha, O., Sathish, R., & **Sheet, D.** (2019). Learning with Multitask Adversaries using Weakly Labelled Data for Semantic Segmentation in Retinal Images. In 2019 International Conference on Medical Imaging with Deep Learning, PMLR 102 (pp. 414-426).
4. Sathish, R., Rajan, R., Vupputuri, A., Ghosh, N., & **Sheet, D.** (2019). Adversarially Trained Convolutional Neural Networks for Semantic Segmentation of Ischaemic Stroke Lesion using Multisequence Magnetic Resonance Imaging. In 41st International Engineering in Medicine and Biology Conference (EMBC).
5. Tom, F., & **Sheet, D.** (2018, April). Simulating patho-realistic ultrasound images using deep generative networks with adversarial learning. In 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018) (pp. 1174-1177). IEEE.

4. Description of expected audience

The tutorial is particularly aimed at enthusiastic early adopters of deep learning for biomedical and biological imaging and image analysis, who are trained in basics of machine learning, but do not have a clarity in functioning of adversarial learning and adversarially learnt loss functions. The target audience is expected to include students in research training (Master's, PhDs) and Post-Docs who are not trained with an undergraduate basic curriculum in machine learning (e.g. undergraduate training in electrical or biomedical sciences, physics, biological or medical sciences), but are interested to explore adversarial deep learning in their work. Audience is expected to have a basic experience of computer programming with Python and are expected to be conversant with commonly used terminologies in the biomedical imaging community. The audience is expected to be conversant in neural networks and its learning.

5. Description of the expected audience participation

The audience undertaking this tutorial would be attending 2 hours on the 5 outlined topics, and 1 hours of programming tutorials implementing 2 exercises in Python with Pytorch. One of the exercises would be on lesion generation using adversarial autoencoders and the other would be on adversarial loss for semantic segmentation, learnt using a vanilla discriminator and using the Turing test loss. Audience are expected to bring in their laptops preloaded with the software configurations to be provided ahead of time. Additionally cloud compute facility would be provided to them for use during this tutorial. (Subject to discussion with AWS or Google or Microsoft or Intel post decision on the proposal. Similar arrangements were made by the Instructor during earlier workshops/schools on deep learning.)

6. Description of the coursepacks

The following materials would be distributed to the participants:

- Slides to audience - released as PDFs (CC 4.0 SA license) via SlideShare/Google Drive link
- Python codes vide Jupyter Notebooks on Github (Apache / CC 4.0 license) (Similar to archives from earlier workshops on deep learning - https://github.com/iitkliv/DLMedia2017IITMandi , https://github.com/iitkliv/dlvcnptel)