Deep Adversarial Learning for NLP

William Yang Wang
UC Santa Barbara
william@cs.ucsb.edu

Sameer Singh
UC Irvine
sameer@uci.edu

Jiwei Li
Shannon.ai
jiwei_li@shannonai.com

Abstract

Adversarial learning is a game-theoretic learning paradigm, which has achieved huge successes in the field of Computer Vision recently. It is a general framework that enables a variety of learning models, including the popular Generative Adversarial Networks (GANs). Due to the discrete nature of language, designing adversarial learning models is still challenging for NLP problems.

In this tutorial, we provide a gentle introduction to the foundation of deep adversarial learning, as well as some practical problem formulations and solutions in NLP. We describe recent advances in deep adversarial learning for NLP, with a special focus on generation, adversarial examples & rules, and dialogue. We provide an overview of the research area, categorize different types of adversarial learning models, and discuss pros and cons, aiming to provide some practical perspectives on the future of adversarial learning for solving real-world NLP problems.

1 Tutorial Description

Adversarial learning (AdvL) is an emerging research area that involves a game-theoretical formulation of the learning problem. Recently, with the introduction of Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), we have observed some stunning results in the area of image synthesis in Computer Vision (Brock et al., 2018).

Comparing to images, even language is discrete, the general family of adversarial learning methods still have gained significantly more attentions in NLP in recent years[1]. In contrast to the focus of GANs in Computer Vision, Natural Language Processing researchers have taken a broader approach to adversarial learning. For example, three core technical subareas for adversarial learning include:

- **Adversarial Examples**, where researchers focus on learning or creating adversarial examples or rules to improve the robustness of NLP systems. (Jia and Liang, 2017; Alzantot et al., 2018; Iyyer et al., 2018; Ebrahimi et al., 2018a,b; Shi et al., 2018b; Chen et al., 2018; Farag et al., 2018; Ribeiro et al., 2018; Zhao et al., 2018)
- **Adversarial Training**, which focuses on adding noise, randomness, or adversarial loss during optimization. (Wu et al., 2017; Wang and Bansal, 2018; Li et al., 2018a; Yasunaga et al., 2018; Ponti et al., 2018; Kurita et al., 2018; Kang et al., 2018; Li et al., 2018c; Masumura et al., 2018)
- **Adversarial Generation**, which primarily includes practical solutions of GANs for processing and generation natural language. (Yu et al., 2017; Li et al., 2017; Yang et al., 2018; Wang and Lee, 2018; Xu et al., 2018)

Additionally, we will also introduce other technical focuses such as negative sampling and contrastive estimation (Cai and Wang, 2018; Bose et al., 2018), adversarial evaluation (Elliott, 2018), and reward learning (Wang et al., 2018c). In particular, we will also provide a gentle introduction to the applications of adversarial learning in different NLP problems, including social media (Wang et al., 2018a; Carton et al., 2018), domain adaptation (Kim et al., 2017; Alam et al., 2018; Zou et al., 2018; Chen and Cardie, 2018; Tran and Nguyen, 2018; Cao et al., 2018; Li et al., 2018b), data cleaning (Elazar and Goldberg, 2018; Shah et al., 2018; Ryu et al., 2018; Zellers et al., 2018), information extraction (Qin et al., 2018; Hong et al., 2018; Wang et al., 2018b; Shi et al., 2018a; Bekoulis et al., 2018), and information retrieval (Li and Cheng, 2018).

Adversarial learning methods could easily combine any representation learning based neural networks, and optimize for complex problems in NLP. However, a key challenge for applying deep adversarial learning techniques to real-world sized NLP problems is the model design issue. This tutorial draws connections from theories of deep adversarial learning to practical applications in NLP.

In particular, we start with the gentle introduction to the fundamentals of adversarial learning. We further

[1] Through a simple ACL anthology search, we found that in 2018, there were 20+ times more papers mentioning "adversarial", comparing to 2016. Meanwhile, the growth of all accepted papers is 1.39 times during this period.

Proceedings of NAACL HLT 2019: Tutorial Abstracts, pages 1–5
Minneapolis, Minnesota, June 2, 2019.

discuss their modern deep learning extensions such as Generative Adversarial Networks (Goodfellow et al., 2014). In the first part of the tutorial, we also outline various applications of deep adversarial learning in NLP listed above. In the second part of the tutorial, we will focus on generation of adversarial examples and their uses in NLP tasks, including (1) The inclusion and creation of adversarial examples for robust NLP; (2) The usage of adversarial rules for interpretable and explainable models; and (3) The relationship between adversarial training and adversarial examples. In the third part of the tutorial, we focus on GANs. We start with the general background introduction of generative adversarial learning. We will introduce an in-depth case study of Generative Adversarial Networks for NLP, with a focus on dialogue generation (Li et al., 2017).

This tutorial aims at introducing deep adversarial learning methods to researchers in the NLP community. We do not assume any particular prior knowledge in adversarial learning. The intended length of the tutorial is 3.5 hours, including a coffee break.

2 Outline

Noise-Robust Representation Learning, Adversarial Learning, and Generation are three closely related research subjects in Natural Language Processing. In this tutorial, we touch the intersection of all the three research subjects, covering various aspects of the theories of modern deep adversarial learning methods, and show their successful applications in NLP. This tutorial is organized in three parts:

- **Foundations of Deep Adversarial Learning**. First, we will provide a brief overview of adversarial learning (RL), and discuss the cutting-edge settings in NLP. We describe methods such as Adversarial Training (Wu et al., 2017), Negative Sampling, and Noise Contrastive Estimation (Cai and Wang, 2018; Bose et al., 2018). We introduce domain-adaptation learning approaches, and the widely used data cleaning and information extraction methods (Elazar and Goldberg, 2018; Shah et al., 2018; Ryu et al., 2018; Zellers et al., 2018; Qin et al., 2018; Hong et al., 2018; Wang et al., 2018b; Shi et al., 2018a; Bekoulis et al., 2018). In this part, we also introduce the modern renovation of deep generative adversarial learning (Goodfellow et al., 2014), with a focus on NLP (Yu et al., 2017; Yang et al., 2018; Wang and Lee, 2018; Xu et al., 2018).

- **Adversarial Examples for NLP** Second, we will focus on the designing practical adversarial examples for NLP tasks. In particular, we will provide an overview of recent methods, including their categorization by whether they are white (e.g. Ebrahimi et al., 2018a) or black box (e.g. Iyyer et al., 2018), character- (e.g. Belinkov and Bisk, 2018) or word-based (e.g. Alzantot et al., 2018), and the tasks they have been applied to. We will also provide an in-depth analysis of some of the general techniques for creating adversarial examples, such as gradient-based (e.g. Ebrahimi et al., 2018b), manually-designed (e.g. Jia and Liang, 2017), or learned (e.g. Zhao et al., 2018) perturbation techniques. Next, we will focus on practical applications of adversarial examples, such as existing work on adversarial rules for interpretable NLP (Ribeiro et al., 2018). To conclude this part, we discuss future directions and novel application areas for adversarial examples in NLP, including KB completion (Pezeshkpour et al., 2019).

- **An In-depth Case Study of GANs in NLP.** Third, we switch from the focuses of adversarial training and adversarial examples to generative adversarial networks (Goodfellow et al., 2014). We will discuss why it is challenging to deploy GANs for NLP problems, comparing to vision problems. We then focus on introducing SeqGAN (Yu et al., 2017), an early solution of textual models of GAN, with a focus on policy gradient and Monte Carlo Tree Search. Finally, we provide an in-depth case study of deploying two-agent GAN models for conversational AI (Li et al., 2017). We will summarize the lessons learned, and how we can move forward to investigate game-theoretical approaches in advancing NLP problems.

3 History

The full content of this tutorial has not yet been presented elsewhere, but some parts of this tutorial has also been presented at the following locations in recent years:

1. "*Deep Reinforcement Learning for NLP*", William Wang, Jiwei Li, and Xiaodong He presented at the ACL 2018 Tutorial, Melbourne, AU., Total attendance: 500 (the most popular tutorial).

2. "*Scalable Construction and Reasoning of Massive Knowledge Bases*", Xiang Ren, Nanyun Peng, William Wang. Tutorial at NAACL 2018, New Orleans, Total attendance: 300 (the most popular tutorial).

3. "*Questioning Question Answering Answers*", Sameer Singh, invited talk at the Machine Reading for Question Answering (MRQA) Workshop at ACL 2018 in Melbourne AU, Total attendance: 200 (one of the most popular workshops).

4. "*Teaching a Machine to Converse*", Jiwei Li, presented at OSU, UC Berkeley, UCSB, Harbin Inst. of Technology, total attendance: 500.

5. "*Local, Model-Agnostic Explanations of Machine Learning Predictions*", Sameer Singh, invited

talks and keynotes at various venues, such as UCSD, KAIST, UC Riverside, FICO, and Caltech, total attendance: 800.

4 Duration

The intended duration of this tutorial is 3.5 hours plus a half an hour break.

5 Information About the Presenters

William Wang is an Assistant Professor at the Department of Computer Science, University of California, Santa Barbara. He received his PhD from School of Computer Science, Carnegie Mellon University. He focuses on information extraction and he is the faculty author of KBGAN—the first deep adversarial learning system for knowledge graph reasoning. He has presented tutorials at ACL, NAACL, and IJCAI, with more than 60 published papers at leading conferences and journals including *ACL, EMNLP, NAACL, CVPR, ECCV, COLING, AAAI, IJCAI, CIKM, ICWSM, SIGDIAL, IJCNLP, INTERSPEECH, ICASSP, ASRU, SLT, Machine Learning, and Computer Speech & Language*, and he has received paper awards and honors from CIKM, ASRU, and EMNLP. Website: http://www.cs.ucsb.edu/~william/

Sameer Singh is an Assistant Professor of Computer Science at the University of California, Irvine. He is working on large-scale and interpretable machine learning applied to information extraction and natural language processing. Before UCI, Sameer was a Postdoctoral Research Associate at the University of Washington. He received his PhD from the University of Massachusetts, Amherst in 2014, during which he also interned at Microsoft Research, Google Research, and Yahoo! Labs. His group has received funding from Allen Institute for AI, NSF, Adobe Research, and FICO, and was selected as a DARPA Riser. Sameer has presented tutorials at WSDM and AAAI, and published extensively at top-tier machine learning and natural language processing conferences. Website: http://sameersingh.org/

Jiwei Li is the co-founder and CEO of Shannon.AI, an AI startup based in Beijing, China. He spent three years and received his PhD in Computer Science from Stanford University with Prof. Dan Jurafsky. His research focuses on deep learning in NLP applications, including dialogue, question answering, discourse analysis and information extraction. He has published more than 20 lead-author papers at ACL, EMNLP, NAACL and ICLR, and is the most prolific NLP/ML first author during 2012-2018. He is the lead author of the first study in deep reinforcement learning and adversarial learning for dialogue generation. He is the recipient of a Facebook Fellowship in 2015 and he is named Forbes 30 under 30 in China in 2018. Website: https://nlp.stanford.edu/~bdlijiwei/.

References

Firoj Alam, Shafiq Joty, and Muhammad Imran. 2018. Domain adaptation with adversarial training and graph embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. *arXiv preprint arXiv:1804.07998*.

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018. Adversarial training for multi-context joint entity and relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2830–2836, Brussels, Belgium. Association for Computational Linguistics.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations (ICLR)*.

Avishek Bose, Huan Ling, and Yanshuai Cao. 2018. Adversarial contrastive estimation. *ACL*.

Andrew Brock, Jeff Donahue, and Karen Simonyan. 2018. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.

Liwei Cai and William Yang Wang. 2018. Kbgan: Adversarial learning for knowledge graph embeddings. *NAACL*.

Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, and Shengping Liu. 2018. Adversarial transfer learning for chinese named entity recognition with self-attention mechanism. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Samuel Carton, Qiaozhu Mei, and Paul Resnick. 2018. Extractive adversarial networks: High-recall explanations for identifying personal attacks in social media posts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Hongge Chen, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, and Cho-Jui Hsieh. 2018. Attacking visual language grounding with adversarial examples: A case study on neural image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Xilun Chen and Claire Cardie. 2018. Multinomial adversarial networks for multi-domain text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 1226–1240.

Javid Ebrahimi, Daniel Lowd, and Dejing Dou. 2018a. On adversarial examples for character-level neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 653–663.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018b. Hotflip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 31–36.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21.

Desmond Elliott. 2018. Adversarial evaluation of multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2974–2978.

Youmna Farag, Helen Yannakoudakis, and Ted Briscoe. 2018. Neural automated essay scoring and coherence modeling for adversarially crafted input. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Yu Hong, Wenxuan Zhou, jingli zhang, Guodong Zhou, and Qiaoming Zhu. 2018. Self-regulation: Employing a generative adversarial network to improve event detection. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of NAACL-HLT*, pages 1875–1885.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031.

Dongyeop Kang, Tushar Khot, Ashish Sabharwal, and Eduard Hovy. 2018. Adventure: Adversarial training for textual entailment with knowledge-guided examples. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Young-Bum Kim, Karl Stratos, and Dongchan Kim. 2017. Adversarial adaptation of synthetic or stale data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1297–1307.

Shuhei Kurita, Daisuke Kawahara, and Sadao Kurohashi. 2018. Neural adversarial training for semi-supervised japanese predicate-argument structure analysis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Bo Li and Ping Cheng. 2018. Learning neural representation for clir with adversarial framework. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1861–1870.

Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017. Adversarial learning for neural dialogue generation. *EMNLP*.

Juntao Li, Yan Song, Haisong Zhang, Dongmin Chen, Shuming Shi, Dongyan Zhao, and Rui Yan. 2018a. Generating classical chinese poems via conditional variational autoencoder and adversarial training. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018b. What's in a domain? learning domain-robust text representations using adversarial training. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*.

Zhongyang Li, Xiao Ding, and Ting Liu. 2018c. Generating reasonable and diversified story ending using sequence to sequence model with adversarial training. In *Proceedings of the 27th International Conference on Computational Linguistics*.

Ryo Masumura, Yusuke Shinohara, Ryuichiro Higashinaka, and Yushi Aono. 2018. Adversarial training for multi-task and multi-lingual joint modeling of utterance intent classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Pouya Pezeshkpour, Yifan Tian, and Sameer Singh. 2019. Investigating robustness and interpretability of link prediction via adversarial modifications. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Edoardo Maria Ponti, Ivan Vulić, Goran Glavaš, Nikola Mrkšić, and Anna Korhonen. 2018. Adversarial propagation and zero-shot cross-lingual transfer of word vector specialization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Pengda Qin, Weiran Xu, and William Yang Wang. 2018. Dsgan: Generative adversarial training for robust distant supervision relation extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Victoria, Australia. ACL.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging nlp models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 856–865.

Seonghan Ryu, Sangjun Koo, Hwanjo Yu, and Gary Geunbae Lee. 2018. Out-of-domain detection based on generative adversarial network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 714–718.

Darsh Shah, Tao Lei, Alessandro Moschitti, Salvatore Romeo, and Preslav Nakov. 2018. Adversarial domain adaptation for duplicate question detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1056–1063.

Ge Shi, Chong Feng, Lifu Huang, Boliang Zhang, Heng Ji, Lejian Liao, and Heyan Huang. 2018a. Genre separation network with adversarial training for cross-genre relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Haoyue Shi, Jiayuan Mao, Tete Xiao, Yuning Jiang, and Jian Sun. 2018b. Learning visually-grounded semantics from contrastive adversarial samples. In *Proceedings of the 27th International Conference on Computational Linguistics*.

Van-Khanh Tran and Le-Minh Nguyen. 2018. Adversarial domain adaptation for variational neural language generation in dialogue systems. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1205–1217.

Weichao Wang, Shi Feng, Wei Gao, Daling Wang, and Yifei Zhang. 2018a. Personalized microblog sentiment classification via adversarial cross-lingual multi-task learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Xiaozhi Wang, Xu Han, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2018b. Adversarial multi-lingual neural relation extraction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1156–1166.

Xin Wang, Wenhu Chen, Yuan-Fang Wang, and William Yang Wang. 2018c. No metrics are perfect: Adversarial reward learning for visual storytelling. In *ACL*.

Yaushian Wang and Hung-yi Lee. 2018. Learning to encode text as human-readable summaries using generative adversarial networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Yicheng Wang and Mohit Bansal. 2018. Robust machine comprehension models via adversarial training. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, volume 2, pages 575–581.

Yi Wu, David Bamman, and Stuart Russell. 2017. Adversarial training for relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1778–1783.

Jingjing Xu, Xuancheng Ren, Junyang Lin, and Xu Sun. 2018. Diversity-promoting gan: A cross-entropy based generative adversarial network for diversified text generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. 2018. Improving neural machine translation with conditional sequence generative adversarial nets. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

Michihiro Yasunaga, Jungo Kasai, and Dragomir Radev. 2018. Robust multilingual part-of-speech tagging via adversarial training. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. *AAAI*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Zhengli Zhao, Dheeru Dua, and Sameer Singh. 2018. Generating natural adversarial examples. In *International Conference on Learning Representations (ICLR)*.

Bowei Zou, Zengzhuang Xu, Yu Hong, and Guodong Zhou. 2018. Adversarial feature adaptation for cross-lingual relation classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 437–448.