Security and robustness

This includes the problem of "fooling" a ML model, and defending models against these attacks.

Attacks:

7a. Gradient-Based Adversarial Attacks
Fast Gradient Signed Method
Citation: Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy. Explaining and Harnessing Adversarial Examples. In Proceedings of the 2015 International Conference on Learning Representations (ICLR '15). [PDF]
[Notebook via Tensorflow] [Notebook via CleverHans] [Notebook via PyTorch]
[Also see CleverHans and Foolbox]
Projected Gradient Descent
Citation: Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In Proceedings of the 2018 International Conference on Learning Representations (ICRL '18). [PDF][Notebook via IBM ART - on audio!]
[Notebook via IBM ART, with defenses] [Also see CleverHans and Foolbox]

7b. One Pixel Attack
Citation: Jiawei Su, Danilo Vasconcellos Vargas, Sakurai Kouichi. One pixel attack for fooling deep neural networks. In IEEE Transactions on Evolutionary Computation 23.5 (2019). [PDF] [Github (includes two notebooks)]

7c. Boundary attack and hop-skip-jump
Citation: Wieland Brendel, Jonas Rauber, Matthias Bethge. Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models. In Proceedings of the 2018 International Conference on Learning Representations (ICLR 2018). [PDF] [Notebook via IBM] [Also in CleverHans and Foolbox]
Citation: Jianbo Chen, Michael I. Jordan, Martin J. Wainwright. HopSkipJumpAttack: A Query-Efficient Decision-Based Attack. In Proceedings of the 2020 IEEE Symposium on Security and Privacy (SP '20). [PDF] [Notebook]
[Notebook: attack on tesseract][Also see CleverHans]

7d. Adversarial patch
Citation: Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi, Justin Gilmer. Adversarial Patch. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017). [PDF] [Notebook (warning: uses Python 2)]
[Notebook] [Notebook - TF1][Notebook - TF2]

7e. Black-box attack by training a substitute model
Citation: Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, Ananthram Swami. Practical Black-Box Attacks against Machine Learning. In Proceedings of the 2017 ACM Asia Conference on Computer and

Communications Security, ASIA CCS 2017. [PDF]
[Example in CleverHans (warning: not a notebook!)]

7f. Model Extraction Attack on BERT
Citation: Kalpesh Krishna, Gaurav Singh Tomar, Ankur P. Parikh, Nicolas Papernot, Mohit Iyyer. Thieves on Sesame Street! Model Extraction of BERT-based APIs. In Proceedings of the 2020 International Conference on Learning Representations (ICLR '20). [PDF] [Github (official via authors)][Blog post]
(Warning: this may require a substantial amount of work to run on Colab!)

Defenses:

7g. Adversarial training
Citation: Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, Patrick McDaniel. Ensemble Adversarial Training: Attacks and Defenses. In Proceedings of the 2018 International Conference on Learning Representations (ICLR '18). [PDF]
Citation: Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. In Proceedings of the 2018 International Conference on Learning Representations (ICLR '18). [PDF]
[Implemented in IBM ART - there's a notebook showing how to use the library. To use Madry's protocol, change the trainer]