

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

نمونه‌سازی تقابلی با استفاده از شبکه‌های مولد تقابلی

پروژه کارشناسی مهندسی کامپیوتر
یگانه مرشدزاده

استاد راهنما: دکتر ناصر مزینی

دانشگاه علم و صنعت ایران
مهرماه ۱۴۰۰

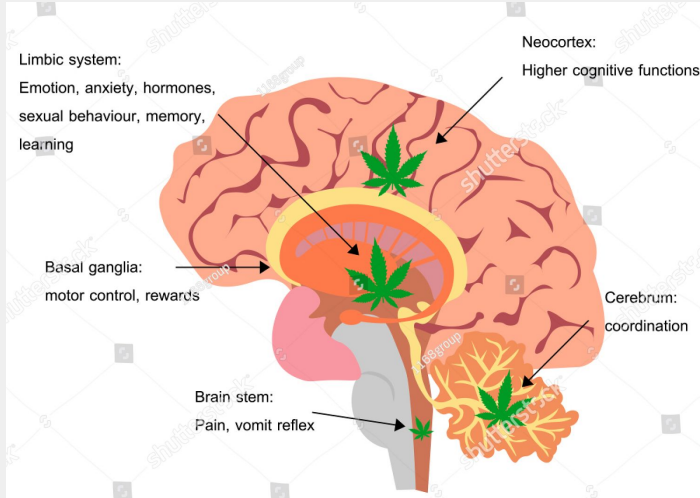
طرح کلی ارائه

1. مقدمه
2. پیش زمینه
3. کارهای مرتبط
4. روش حل مسئله
5. آزمایش ها و نتایجها
6. نتیجه گیری و پیشنهادها

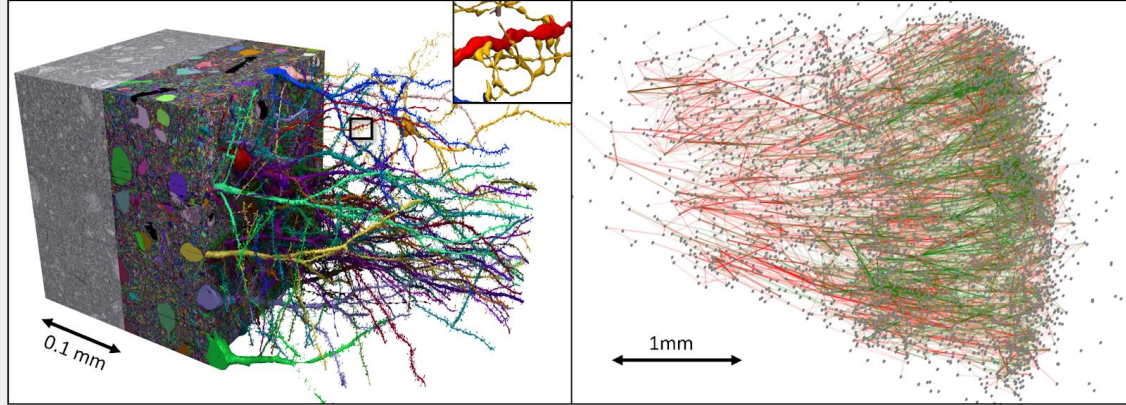
1. مقدمه

مثالهایی از کاربرد یادگیری عمیق

● بازسازی مدارهای مغزی

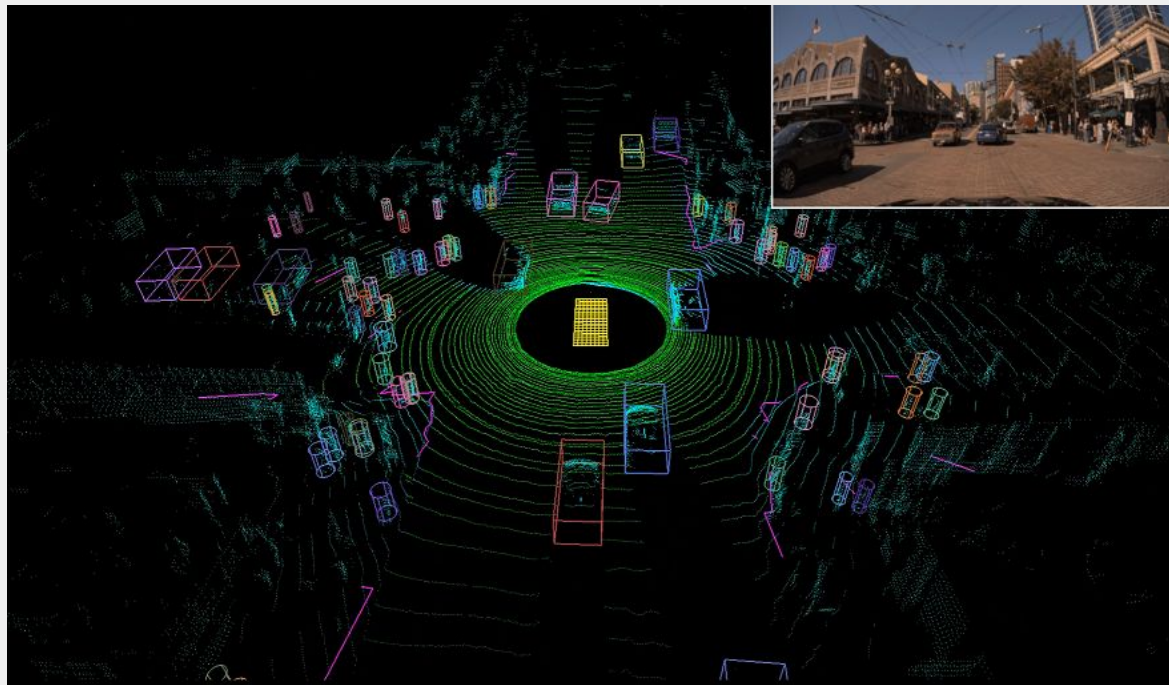


<https://www.shutterstock.com/image-vector/cannabis-brain-effect-on-each-part-1611769504>



<https://ai.googleblog.com/2021/06/a-browsable-petascale-reconstruction-of.html>

مثالهایی از کاربرد یادگیری عمیق

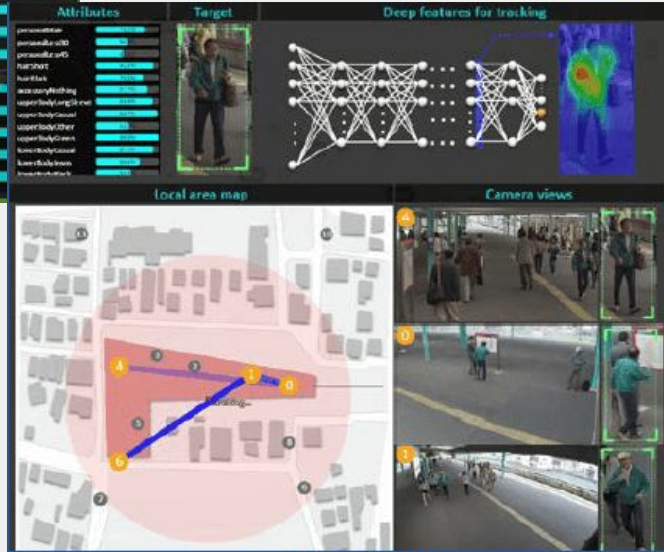


- بازسازی مدارهای مغزی
- آنالیز و بررسی جهش‌های DNA
- بینایی ماشین
- ماشین‌های خودران

https://research.nvidia.com/publication/2020-06_MVLidarNet

مثالهایی از کاربرد یادگیری عمیق

- بازسازی مدارهای مغزی
- آنالیز و بررسی جهش‌های DNA
- بینایی ماشین
- ماشین‌های خودران
- سیستم‌های دیدبانی و مراقبت
- ربات‌ها و هواپیماهای بدون سرنشین



<https://www.asmag.com/showpost/23474.aspx>

مثالهایی از کاربرد یادگیری عمیق

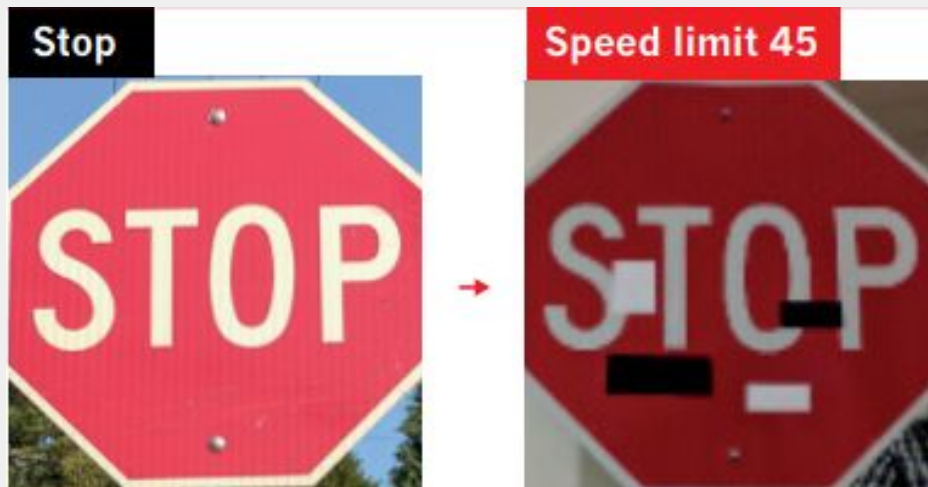


<https://ebusinessinstitute.com/voice/>



- بازسازی مدارهای مغزی
- آنالیز و بررسی جهش‌های DNA
- بینایی ماشین
- ماشین‌های خودران
- سیستم‌های دیدبانی و مراقبت
- ربات‌ها و هواپیماهای بدون سرنشین
- شناسایی بدافزارها
- تشخیص صدا و صحبت
- متوجه شدن زبان طبیعی
- تشخیص دستورهای صوتی

مثال از نمونه تقابلی

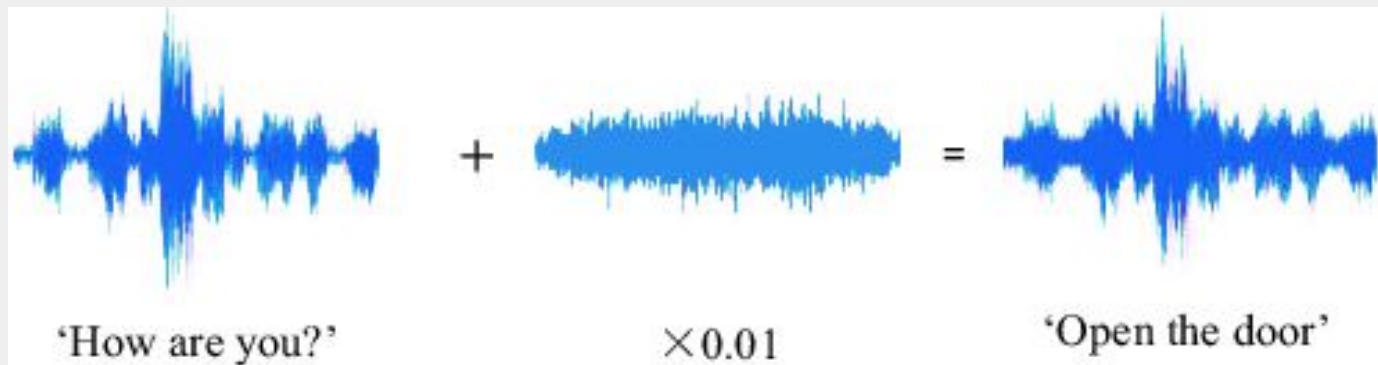


<https://www.nature.com/articles/d41586-019-03013-5>

<https://internetofbusiness.com/opinion-why-driverless-cars-will-force-an-insurance-u-turn/>



مثال از نمونه تقابلی



https://www.researchgate.net/figure/An-illustration-of-machine-learning-adversarial-examples-Studies-have-shown-that-by_fig1_325370539

2. پیش زمینه

نمونه تقابلی و آموزش تقابلی

● نمونه تقابلی

- نسخه دستکاری شده از یک نمونه (مثلا تصویر) که عمدا و آگاهانه آشفته و دستکاری شده است تا شبکه را به اشتباه بیندازد.
- این نمونه‌ها با تغییرهای جزئی تقابلی در تصویر اصلی به وجود می‌آیند.
- تغییرها برای انسان‌ها قابل تشخیص نیستند و نمونه تقابلی مشابه نمونه اصلی به نظر می‌رسد.

● آموزش تقابلی

- استفاده از نمونه‌های تقابلی در کنار نمونه‌های تمیز و دست نخورده برای آموزش مدل‌های یادگیری عمیق

یک دسته‌بندی از انواع حملات تقابلی

● حمله جعبه سفید

- مدل هدف
- معماری مدل
- پارامترهای مدل
- روش آموزش
- داده‌های آموزش

● حمله جعبه سیاه

- مدل هدف (محدود)
- معماری مدل (محدود)
- روش آموزش (محدود)
- پیش‌بینی نهایی کلاس
- ~~پارامترهای مدل~~

● حمله جعبه نیمه‌سفید یا نیمه‌سیاه

- احتمال‌های پیش‌بینی

شبکه مولد تقابلی (شبکه زایای دشمن گونه - GAN)

- مولد (تولید کننده)

- نویز گوسی یا یکنواخت

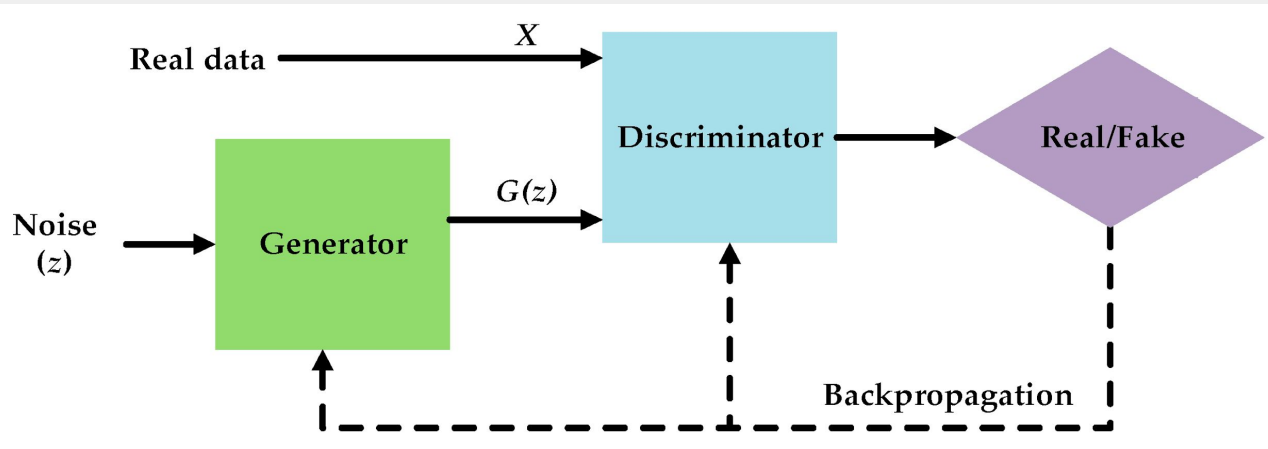
- تصویهای تولید شده بیشترین شباهت به تصویهای حقیقی و طبیعی مجموعهی داده

- تمیزدهنده (جدا ساز - تفکیک کننده - تشخیص دهنده)

- تصویر را به عنوان ورودی می گیرد

- و سعی در تشخیص تصویهای حقیقی




- از تصاویر جعلی توسط مولد دارد.



3. کارهای مرتبط

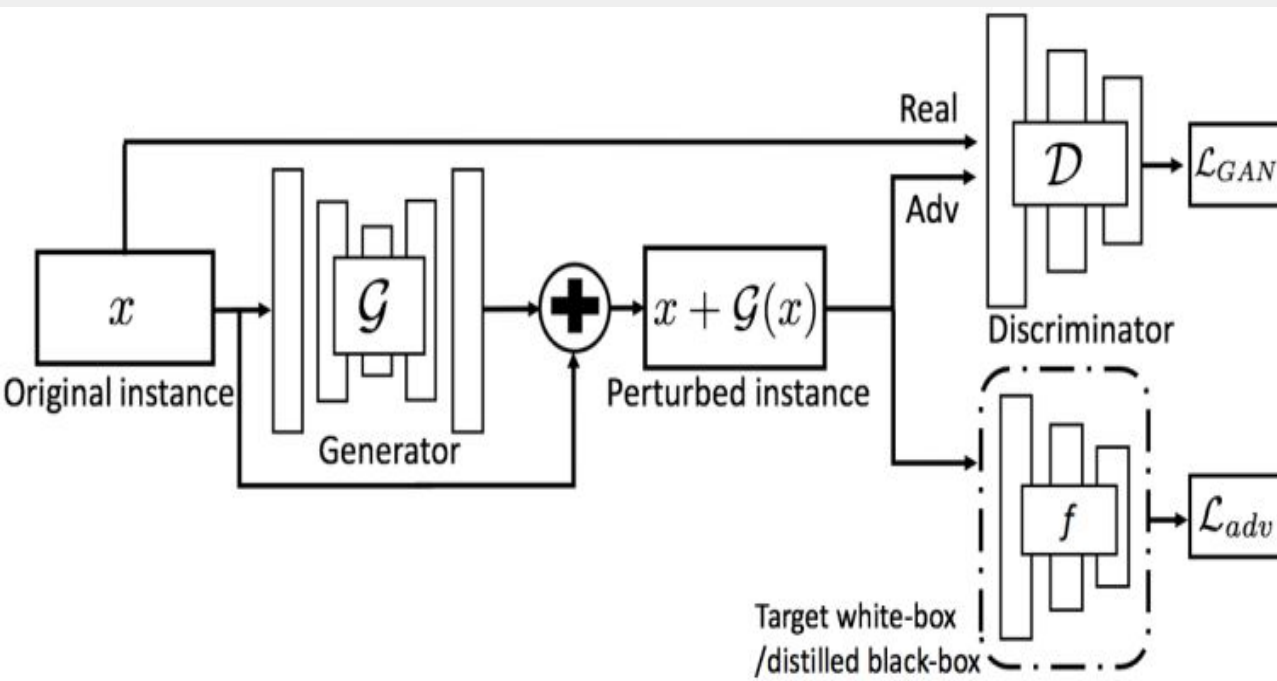
الگوریتم علامت گرادیان سریع

- حمله جعبه سفید زیرا حمله‌کننده به معماری و پارامترهای مدل در تمام زمان نیاز دارد.

	$+ .007 \times$		$=$	
x		$\text{sign}(\nabla_x J(\theta, x, y))$		$x + \epsilon \text{sign}(\nabla_x J(\theta, x, y))$
“panda”		“nematode”		“gibbon”
57.7% confidence		8.2% confidence		99.3 % confidence

4. روش حل مسئله

شبکه Adv-GAN



● مولد

○ تولید تغییرات جزئی

○ شبکه عصبی پیش‌خور

(Feed-Forward Neural Network)

● تمییز دهنده

○ بررسی نزدیک به واقعیت بودن نمونه

های تولید شده

● شبکه عصبی هدف

شبکه Adv-GAN

- مقایسه با روش علامت گرادیان سریع

○ پس از آموزش شبکه پیش‌خور، بی‌درنگ و فوراً می‌توان از آن برای تولید دستکاری‌های تقابلی برای هر نمونه‌ی ورودی، بدون نیاز به دسترسی به خود مدل استفاده کرد.

- حمله جعبه نیمه‌سفید

توابع ضرر

- ضرر شبکه مولد تقابلی Adv-GAN (مولد و تمییزدهنده)

$$L_{GAN} = \mathbb{E}_x \log D(x) + \mathbb{E}_x \log(1 - D(x + G(x)))$$

- ضرر تقابلی

$$L_{adv}^f = \mathbb{E}_x l_f(x + G(x), t)$$

- ضرر کلی

$$L = L_{adv}^f + \alpha L_{GAN} + \beta L_{hinge}$$

- بدست آوردن D و G با حل بازی کمین بیش

$$\arg \min_G \max_D L$$

5. آزمایش‌ها و نتیجه‌ها

مجموعه داده

MNIST ●

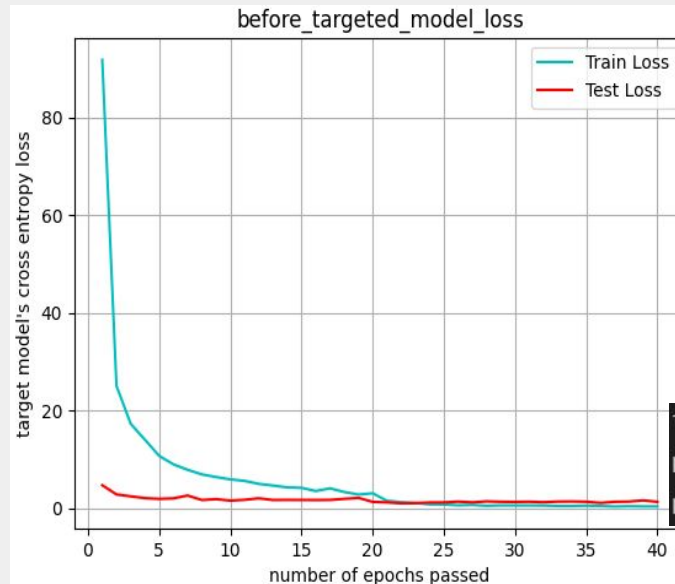
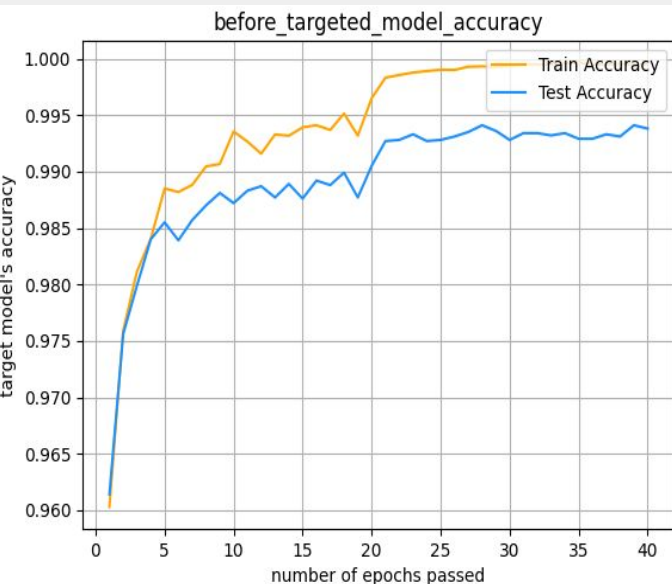
○ ۶۰۰۰۰ عکس 28×28 پیکسلی برای آموزش شبکه

○ ۱۰۰۰۰ عکس 28×28 پیکسلی برای تست شبکه

مراحل انجام آزمایش

1. آموزش شبکه هدف
2. آموزش شبکه Adv-GAN
3. ارزیابی نمونه‌های تقابلی تولید شده

1. آموزش شبکه هدف



○ معماری شبکه:

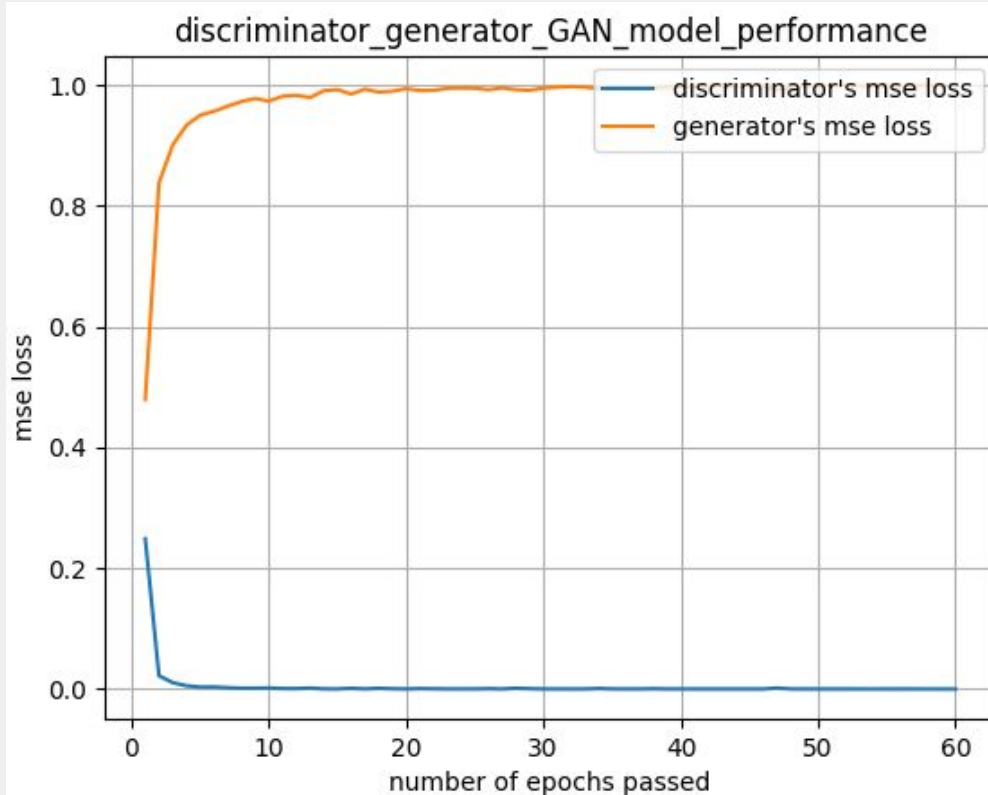
■ ۴ لایه کانولوشن ۲ بعدی

■ ۳ لایه کاملاً متصل

○ عملکرد

```
test_num_correct: 9930 total test data: 10000  
model loss on testing set: 1.504620  
model accuracy on testing set: 0.993000
```


2. آموزش شبکه Adv-GAN



معماری شبکه: ○

مولد ■

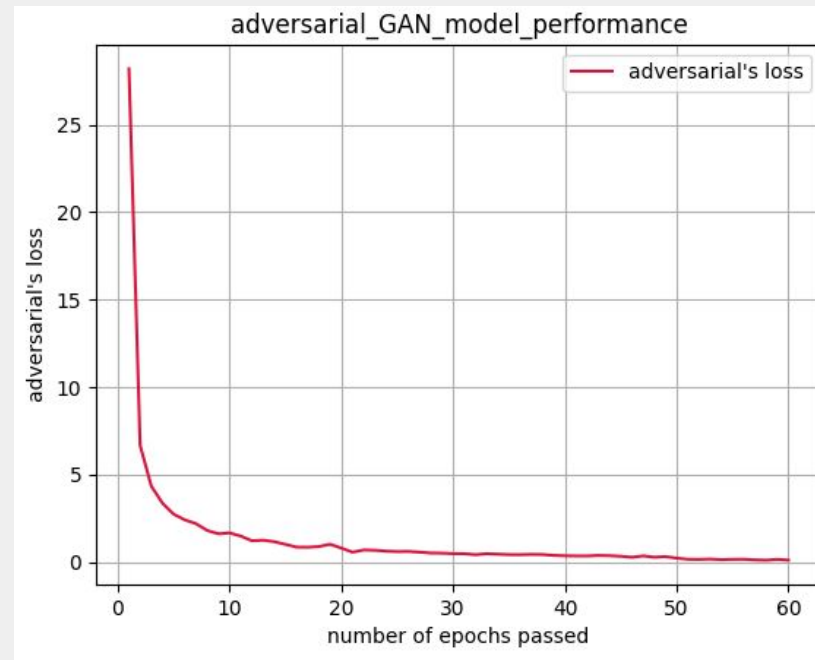
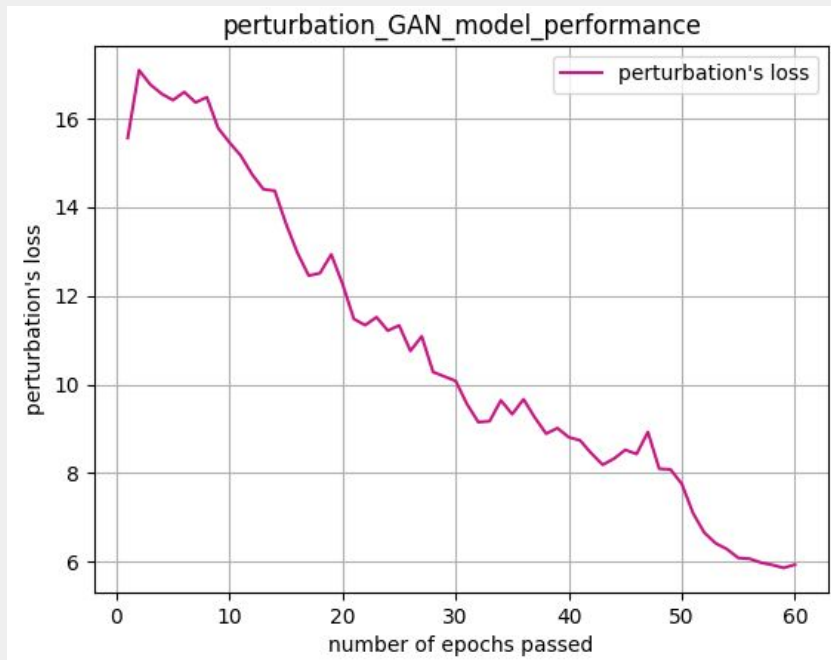
1. کدگذار

2. گلوگاه: ۴ بلوک Resnet

3. کدگشا

تمییزدهنده ■

2. آموزش شبکه Adv-GAN



3. ارزیابی نمونه‌های تقابلی تولید شده

```
Testing set Accuracy score: 0.430000  
Testing set attack success rate: 99.570000
```

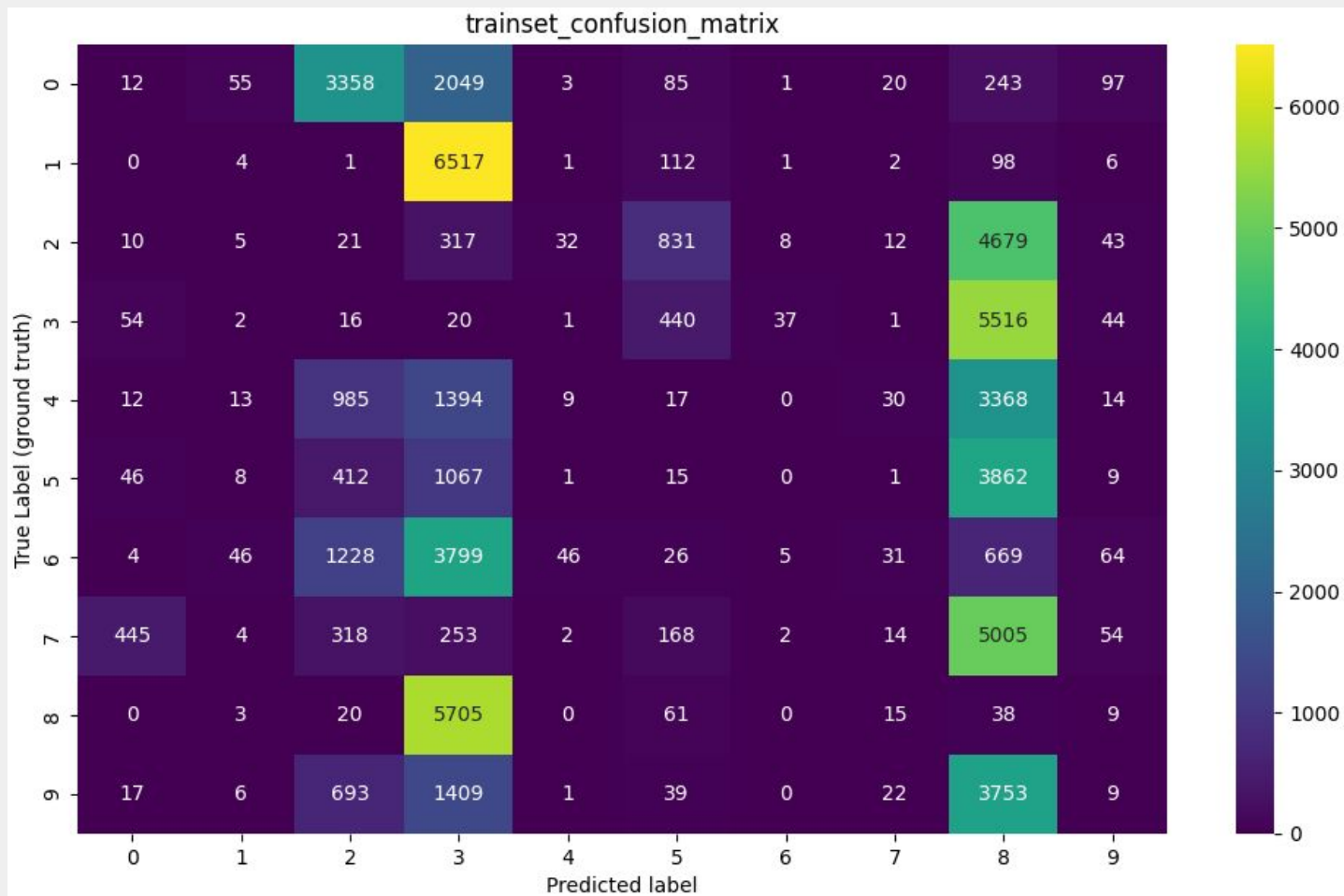
○ دقت

○ میزان موفقیت حمله

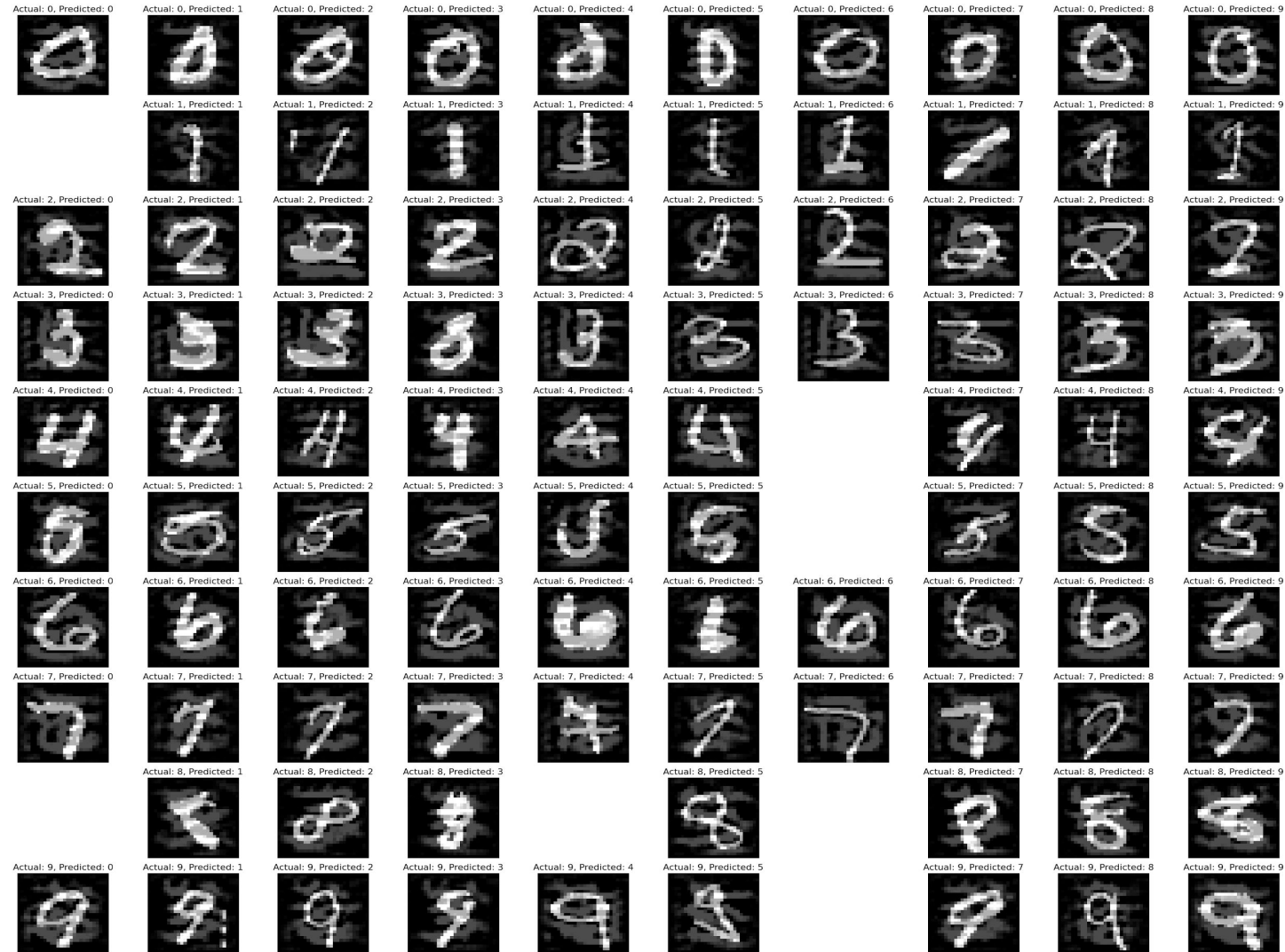
○ دقت هر کلاس

```
Testing set per-class accuracy:
```

```
[(0, 0.30612244897959184), (1, 0.1762114537444934), (2, 1.3565891472868217), (3, 0.297029702970297), (4, 0.10183299389002037), (5, 0.336322869955157), (6, 0.10438413361169102), (7, 0.0), (8, 1.2320328542094456), (9, 0.39643211100099107)]
```



ماتریس درهم ریختگی ○



○ نمونه‌های تقابلی
تولیدشده

6. نتیجه گیری و پیشنهادها

نتیجه گیری

- شبکه Adv-GAN یک روش حمله نوین و قدرتمند برای حمله به شبکه‌های عصبی عمیق
- ایده‌ی اصلی این شبکه برگرفته از شبکه‌های مولد تقابلی
- قابلیت استفاده در حمله‌های جعبه نیمه‌سفید، جعبه سیاه، هدفمند و بدون هدف
- پس از آموزش بخش مولد، تولید دستکاری‌های تقابلی‌ای به صورت مستقل و بهینه
- رساندن عملکرد ۹۹/۳٪ شبکه هدف به ۰/۴۳٪

پیشنهاها و کارهای آینده

- تکنیک‌هایی برای دفاع با استفاده از شبکه‌های مولد تقابلی
- استفاده از مجموعه داده‌های پیچیده و پرده‌تر مانند CIFAR-100 و ImageNet
- حمله به شبکه‌های Recurrent و مجموعه داده‌های متنی و غیر تصویری با استفاده از شبکه‌های مولد تقابلی

با تشکر از توجه شما

تمامی فایل‌های پروژه در https://github.com/yegmor/Final_Project قابل دسترسی هستند.

yegmor@gmail.com

منابع

- ABADI, M., AGARWAL, A., BARHAM, P., BREVDO, E., CHEN, Z., CITRO, C., CORRADO, G., DAVIS, A., DEAN, J., DEVIN, M., GHEMAWAT, S., GOODFELLOW, I., HARP, A., IRVING, G., ISARD, M., JIA, Y., JÓZEFOWICZ, R., KAISER, L., KUDLUR, M., LEVENBERG, J., MANÉ, D., MONGA, R., MOORE, S., MURRAY, D., OLAH, C., SCHUSTER, M., SHLENS, J., STEINER, B., SUTSKEVER, I., TALWAR, K., TUCKER, P., VANHOUCHE, V., VASUDEVAN, V., VIÉGAS, F., VINYALS, O., WARDEN, P., WATTENBERG, M., WICKE, M., YU, Y., AND ZHENG, X. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. ArXiv abs/1603.04467 (2016).
- AKHTAR, N., AND MIAN, A. Threat of adversarial attacks on deep learning in computer vision: A survey. IEEE Access 6 (2018), 14410–14430.
- CARLINI, N., AND WAGNER, D. A. Towards evaluating the robustness of neural networks. 2017 IEEE Symposium on Security and Privacy (SP) (2017), 39–57.
- EVTIMOV, I., EYKHOLT, K., FERNANDES, E., KOHNO, T., LI, B., PRAKASH, A., RAHMATI, A., AND SONG, D. Robust physical-world attacks on deep learning models. arXiv: Cryptography and Security (2017).
- GIUSTI, A., GUZZI, J., CIRESAN, D., HE, F., RODRÍGUEZ, J. P., FONTANA, F., FAESSLER, M., FORSTER, C., SCHMIDHUBER, J., CARO, G. D., SCARAMUZZA, D., AND GAMBARDELLA, L. A machine learning approach to visual perception of forest trails for mobile robots. IEEE Robotics and Automation Letters 1 (2016), 661–667.
- GOODFELLOW, I., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDE-FARLEY, D., OZAI, S., COURVILLE, A. C., AND BENGIO, Y. Generative adversarial nets. in NIPS (2014).
- GOODFELLOW, I., SHLENS, J., AND SZEGEDY, C. Explaining and harnessing adversarial examples. CoRR abs/1412.6572 (2015).
- GROSSE, K., PAPERNOT, N., MANOHARAN, P., BACKES, M., AND MCDANIEL, P. Adversarial examples for malware detection. in ESORICS (2017).
- HE, K., ZHANG, X., REN, S., AND SUN, J. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016), 770–778.

- HEAVEN, D. Why deep-learning ais are so easy to fool.
- HELMSTAEDTER, M., BRIGGMAN, K., TURAGA, S. C., JAIN, V., SEUNG, H., AND DENK, W. Connectomic reconstruction of the inner plexiform layer in the mouse retina. *Nature* 500 (2013), 168–174.
- HINTON, G. E., DENG, L., YU, D., DAHL, G. E., RAHMAN MOHAMED, A., JAITLY, N., SENIOR, A., VANHOUCKE, V., NGUYEN, P., SAINATH, T., AND KINGSBURY, B. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine* 29 (2012), 82.
- JIA, Y., SHELHAMER, E., DONAHUE, J., KARAYEV, S., LONG, J., GIRSHICK, R. B., GUADARRAMA, S., AND DARRELL, T. Caffe: Convolutional architecture for fast feature embedding. *Proceedings of the 22nd ACM international conference on Multimedia* (2014).
- KURAKIN, A., GOODFELLOW, I., AND BENGIO, S. Adversarial examples in the physical world. *ArXiv abs/1607.02533* (2017).
- KURAKIN, A., GOODFELLOW, I., AND BENGIO, S. Adversarial machine learning at scale. *ArXiv abs/1611.01236* (2017).
- MADRY, A., MAKELOV, A., SCHMIDT, L., TSIPRAS, D., AND VLADU, A. Towards deep learning models resistant to adversarial attacks. *ArXiv abs/1706.06083* (2018).
- MNIH, V., KAVUKCUOGLU, K., SILVER, D., RUSU, A. A., VENESS, J., BELLEMARE, M. G., GRAVES, A., RIEDMILLER, M. A., FIDJELAND, A., OSTROVSKI, G., PETERSEN, S., BEATTIE, C., SADIK, A., ANTONOGLOU, I., KING, H., KUMARAN, D., WIERSTRA, D., LEGG, S., AND HASSABIS, D. Human-level control through deep reinforcement learning. *Nature* 518 (2015), 529–533.
- MOOSAVI-DEZFOOLI, S.-M., FAWZI, A., AND FROSSARD, P. Deepfool: A simple and accurate method to fool deep neural networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), 2574–2582.
- NAJAFABADI, M. M., VILLANUSTRE, F., KHOSHGOFTAAR, T., SELIYA, N., WALD, R., AND MUHAREMAGIC, E. A. Deep learning applications and challenges in big data analytics. *Journal of Big Data* 2 (2014), 1–21.
- PAPERNOT, N., MCDANIEL, P., SINHA, A., AND WELLMAN, M. P. Towards the science of security and privacy in machine learning. *ArXiv abs/1611.03814* (2016)
- SU, J., VARGAS, D. V., AND SAKURAI, K. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation* 23 (2019), 828–841.
- SUTSKEVER, I., VINYALS, O., AND LE, Q. V. Sequence to sequence learning with neural networks. in *NIPS* (2014).
- SZEGEDY, C., VANHOUCKE, V., IOFFE, S., SHLENS, J., AND WOJNA, Z. Rethinking the inception architecture for computer vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), 2818–2826.
- VEDALDI, A., AND LENC, K. Matconvnet: Convolutional neural networks for matlab. *Proceedings of the 23rd ACM international conference on Multimedia* (2015).
- XIAO, C., LI, B., ZHU, J.-Y., HE, W., LIU, M., AND SONG, D. Generating adversarial examples with adversarial networks. *ArXiv abs/1801.02610* (2018).
- XIONG, H. Y., ALIPANAHI, B., LEE, L. J., BRETSCHNEIDER, H., MERICO, D., YUEN, R., HUA, Y., GUEROUSSOV, S., NAJAFABADI, H., HUGHES, T., MORRIS, Q., BARASH, Y., KRAINER, A., JOJIC, N., SCHERER, S., BLENCOWE, B., AND FREY, B. The human splicing code reveals new insights into the genetic determinants of disease. *Science* 347 (2015).