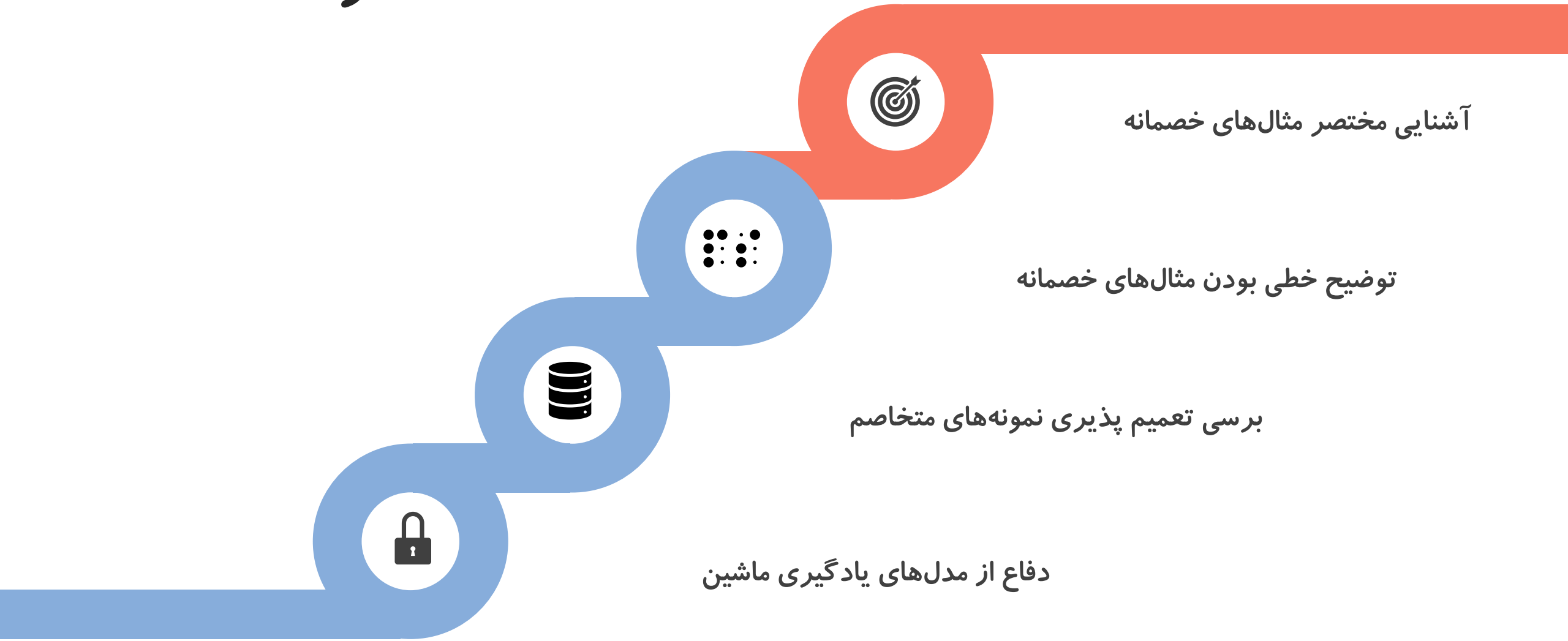
دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

بررسی و مهار مثال‌های خصمانه در یادگیری ماشین

ارائه دهنده: بردیا اردکانیان
استاد راهنما: دکتر رضا صفابخش

اردیبهشت ۱۴۰۱

اهداف ارائه



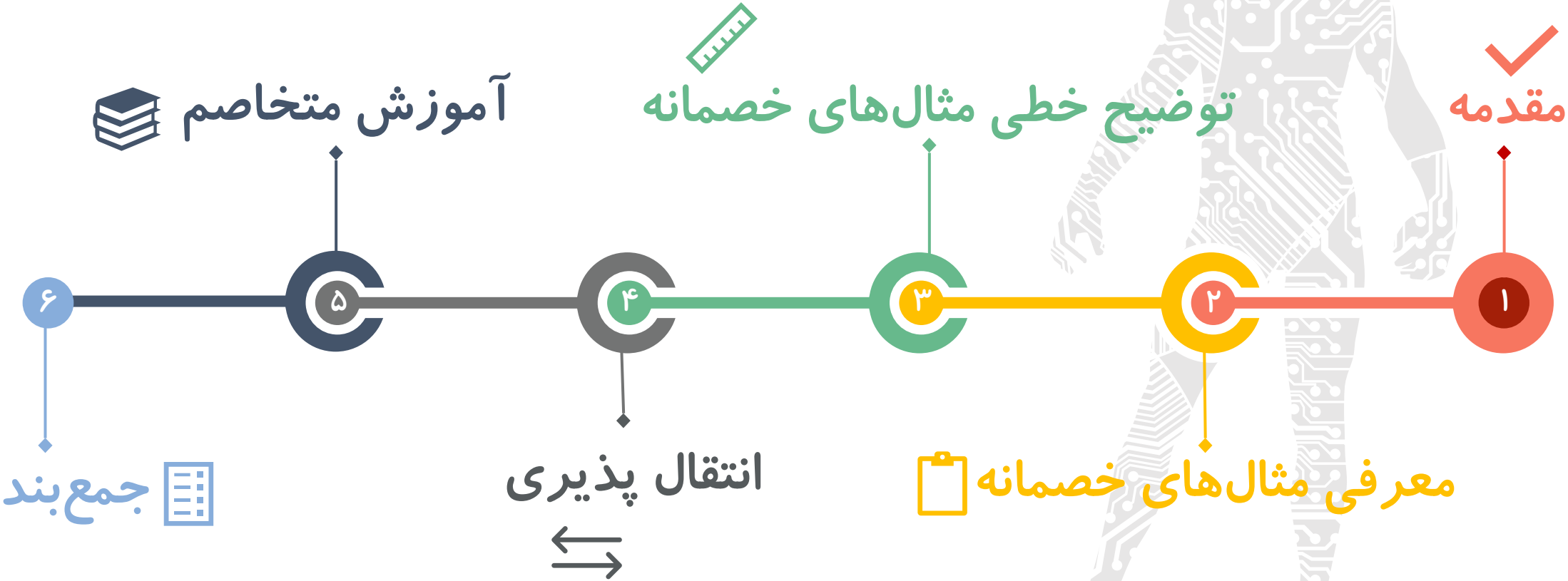
آشنایی مختصر مثال‌های خصمانه

توضیح خطی بودن مثال‌های خصمانه

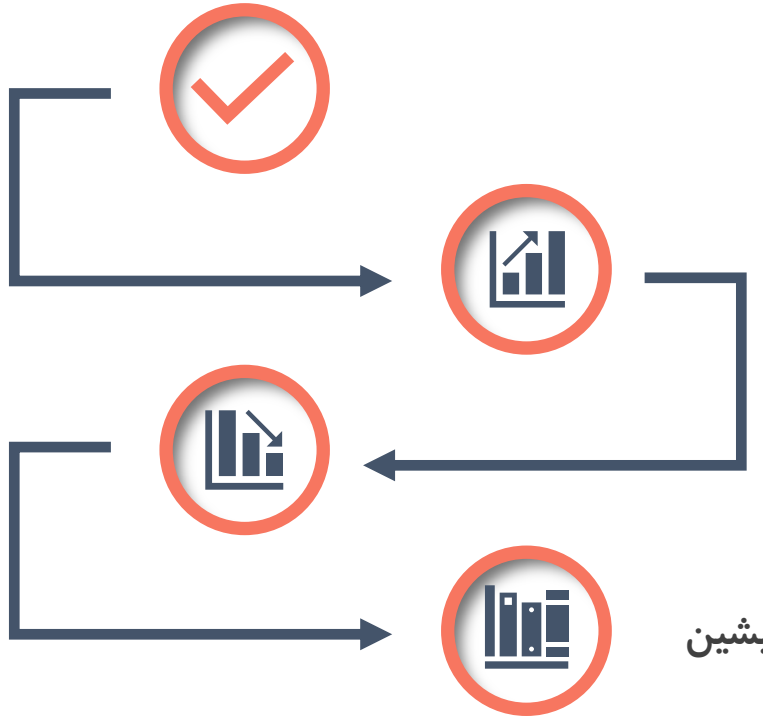
بررسی تعمیم پذیری نمونه‌های متخاصم

دفاع از مدل‌های یادگیری ماشین

سیر ارائه


مقدمه
۱
معرفی مثال‌های خصمانه
۲
توضیح خطی مثال‌های خصمانه
۳
انتقال پذیری
۴
آموزش متخاصم
۵
جمع‌بندی
۶

مقدمه



موفقیت‌های یادگیری ماشین

خطر شکست مدل‌های یادگیری ماشین

تحقیقات پیشین

موفقیت‌های یادگیری ماشین

Google Cloud Platform

یادگیری ماشین به
عنوان یک سرویس

FeatureSmith

تشخیص بدافزار

کشف کلاهبرداری

ماشین‌های خودران

خطر شکست مدل‌های یادگیری ماشین

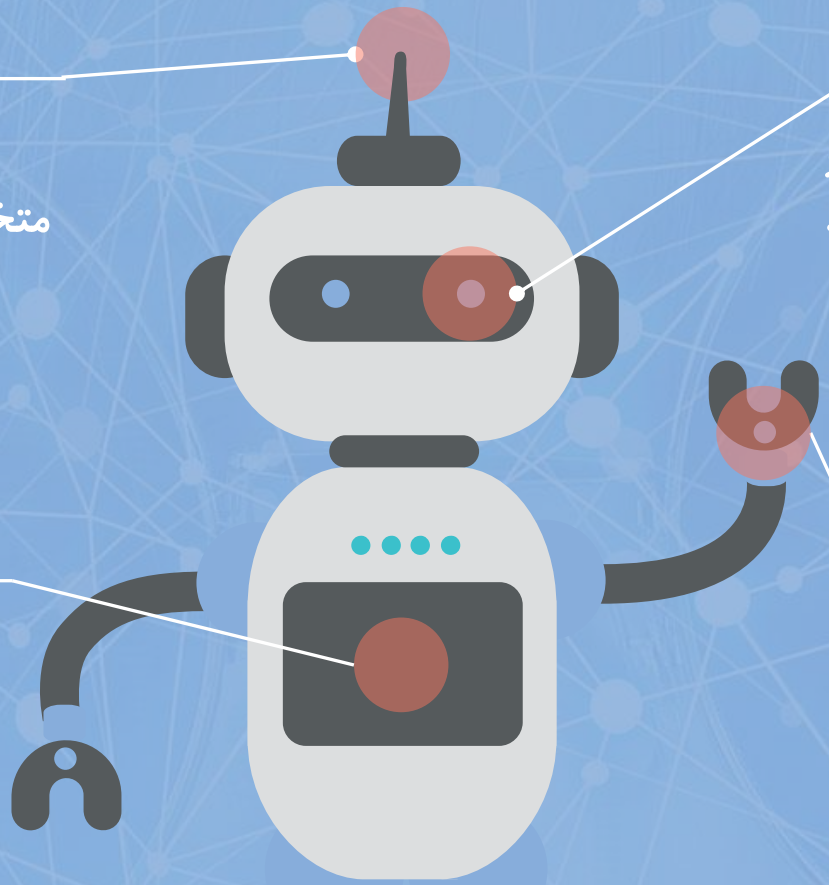
- یک دشمن مدلی که با یادگیری ماشینی آموزش دیده است را مجبور می‌کند پیش‌بینی اشتباه کند

تشخیص دهنده بدافزار	بینایی کامپیوتر
• در دسترس بودن مدل برای پرس و جوی فشرده • دسترسی به برچسب و امتیاز مدل • طبقه بندی کننده باینری (دو خروجی: بدافزار / بی‌خطر)	• در دسترس بودن مدل از راه دور از طریق یک API • فقط به برچسب مدل دسترسی داشته باشید • یک طبقه بندی کننده چند کلاسه (تا ۱۰۰۰ خروجی)

تحقیقات پیشین

بیش از ۱۰۰ مقاله در زمینه مثال‌های متخاصم و آموزش متخاصم مورد مطالعه قرار گرفته است.

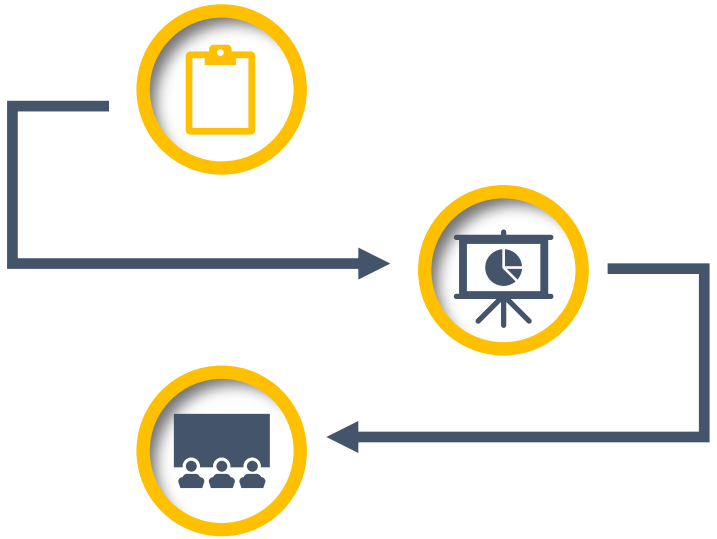
فرضیه‌ها نشان می‌دادند که دلیل مثال‌های متخاصم غیرخطی بودن شدید شبکه های عصبی عمیق است.



مطالعات صورت گرفته نشان‌داده‌اند که طبقه‌بندی کننده‌های مبتنی بر یادگیری ماشین مدرن در برابر مثال‌های متخاصم اسیب‌پذیر هستند.

فرضیه‌ها نشان‌داده است این نتایج اغلب به عنوان یک نقص در شبکه های عمیق تفسیر می شوند

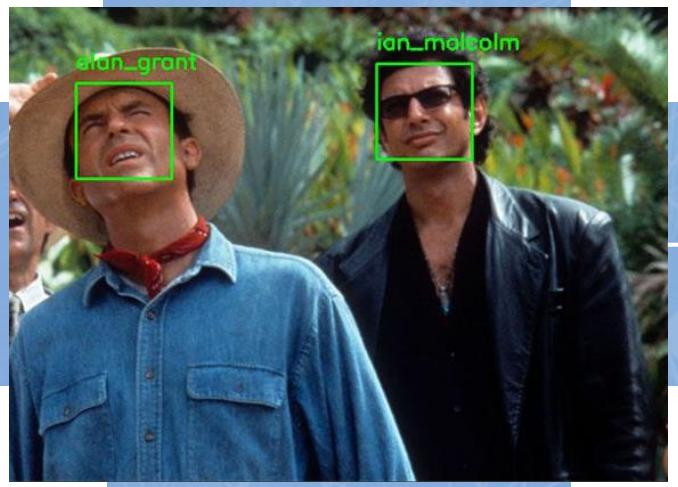
معرفی مثال‌های خصمانه

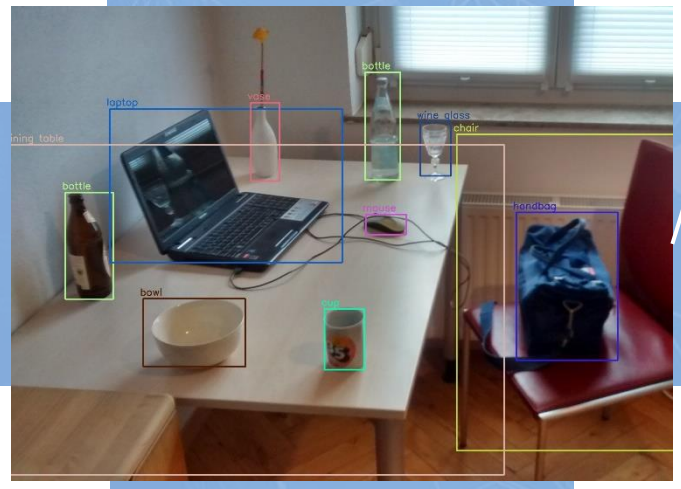


درک وظایف توسط مدل یادگیری ماشین

شناخت مثال‌های خصمانه

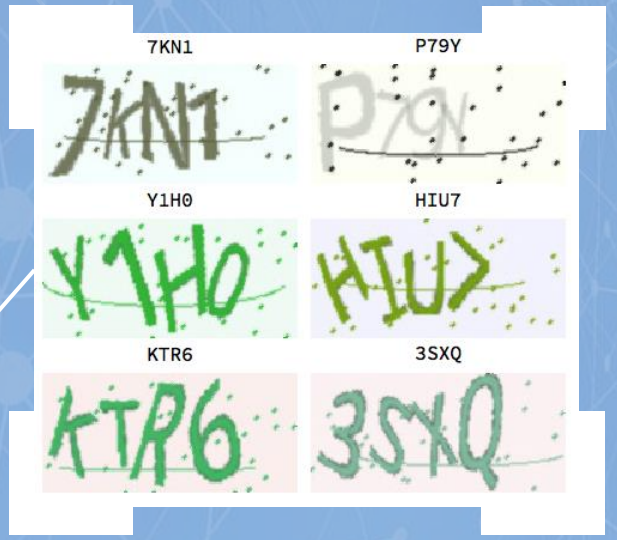
کارهایی که یادگیری ماشین بهتر از انسان‌ها می‌دهند





تشخیص چهره و اشیاء

حل کردن کپچا



آیا شبکه های عصبی این وظایف را درک می کنند؟

- **آزمایش فکری اتاق چینی جان سرل**

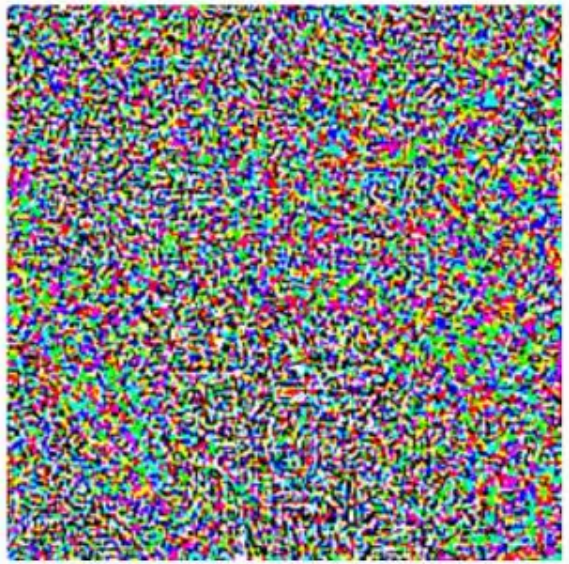
這是一個文本 → 我知道這是我可以閱讀的文本

- **اگر جمله در کتاب دستورعمل نباشد جه اتفاقی می‌افتد؟**

你知道這段文字是什麼嗎？→

معرفی مثال‌های خصمانه

$+ .007 \times$



$=$

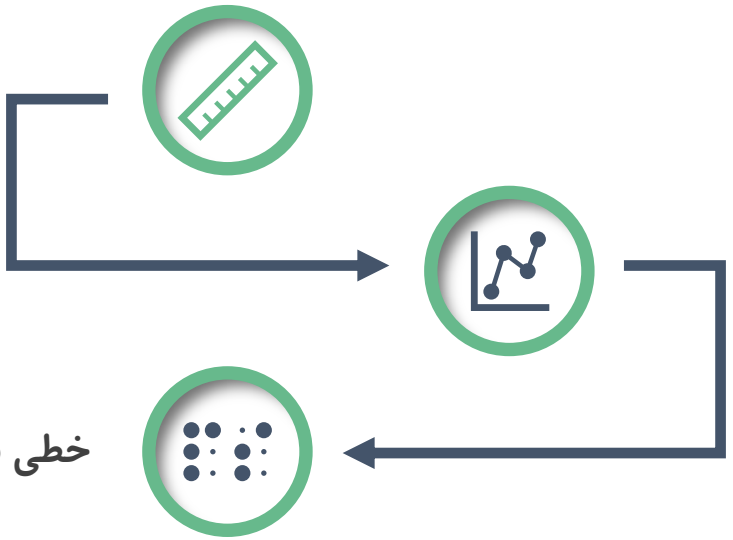
«پاندا»
اطمینان ۵۷.۷ درصد

«نماتد»
اطمینان ۸.۲ درصد

«گیبون»
اطمینان ۹۹.۳ درصد

مثال‌های خصمانه نشان‌دهنده کوچک ترین تغییرات دامنه هستند

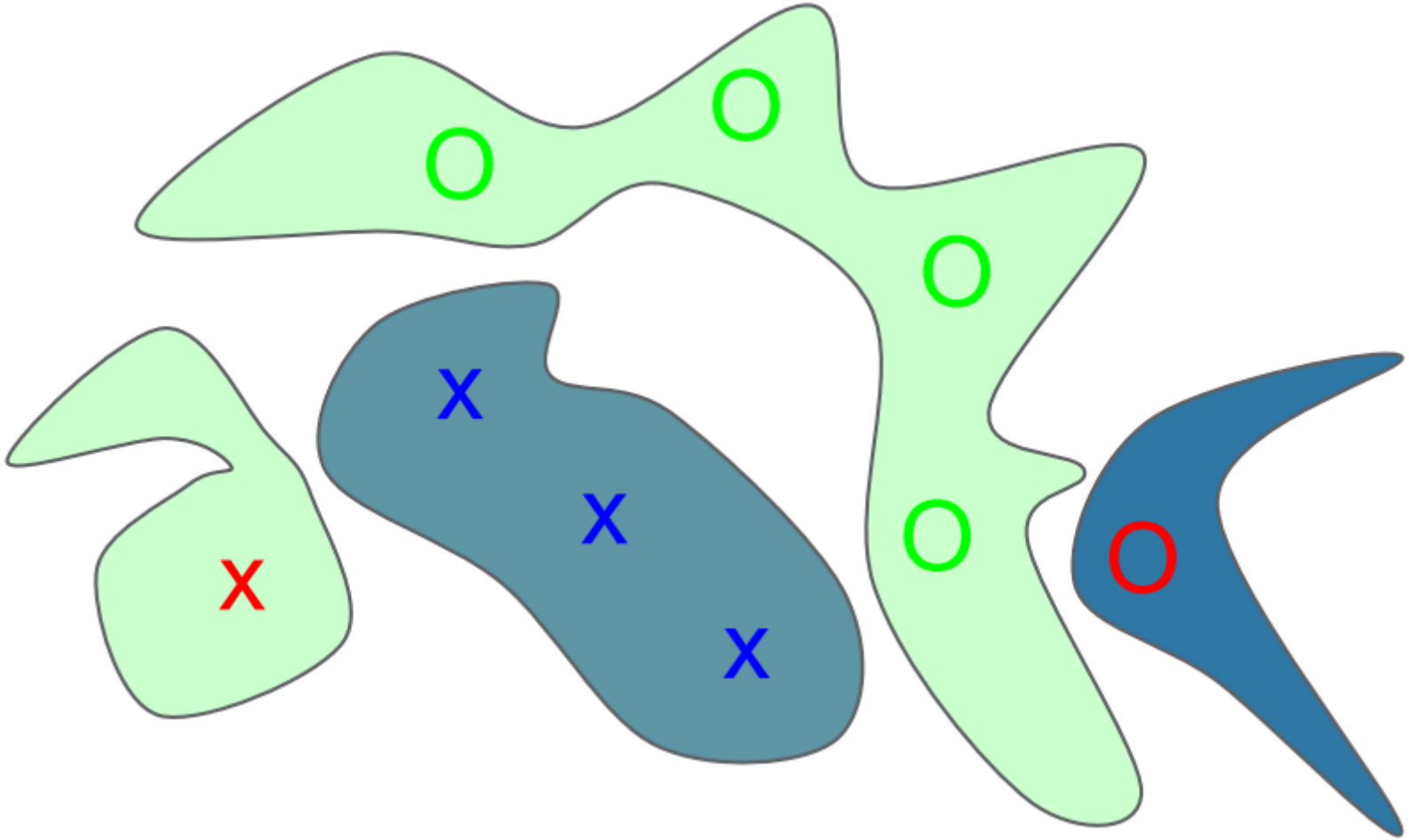
توضیح خطی مثال‌های خصمانه



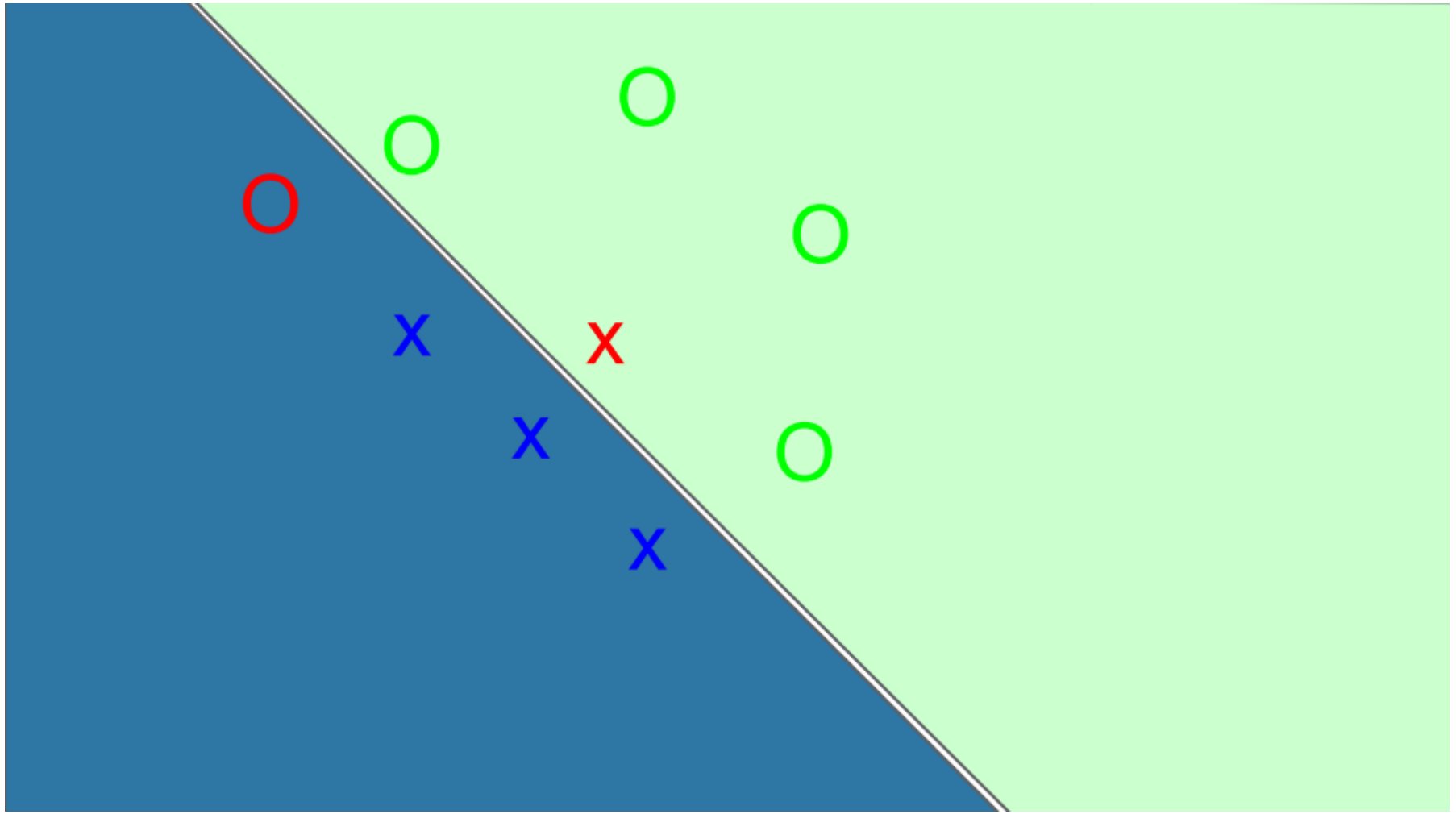
نمونه‌های خصمانه بیش پرازش و کم پرازش

خطی بودن تکه‌ای مدل‌های شبکه‌های عمیق

نمونه‌های خصمانه از بیش برازش



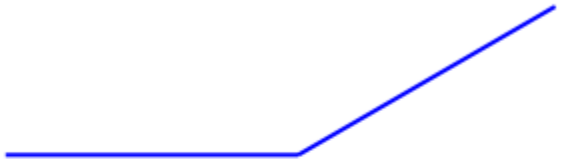
نمونه‌های خصمانه از کم برازش



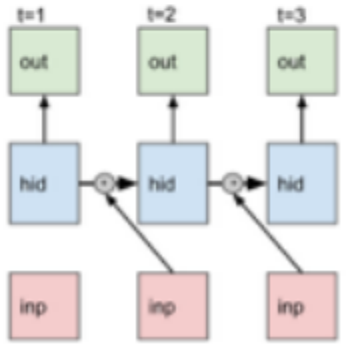
خطی بودن تکه‌ای مدل‌های شبکه‌های عمیق

ماکس اوت

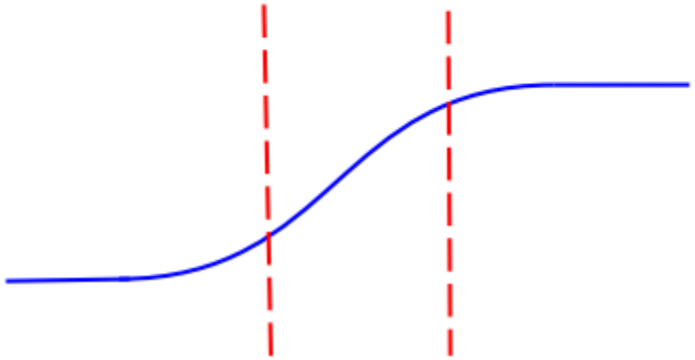
واحد یکسو شده خطی



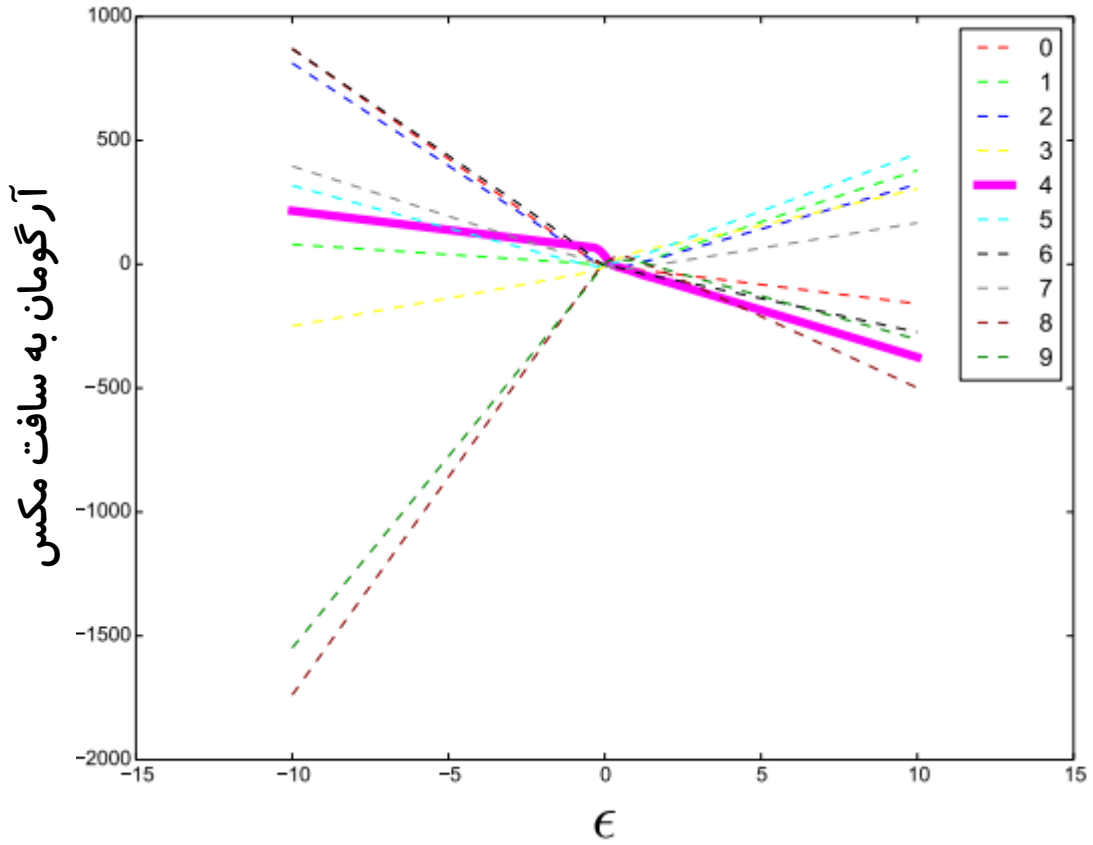
شبکه عصبی بازگشتی

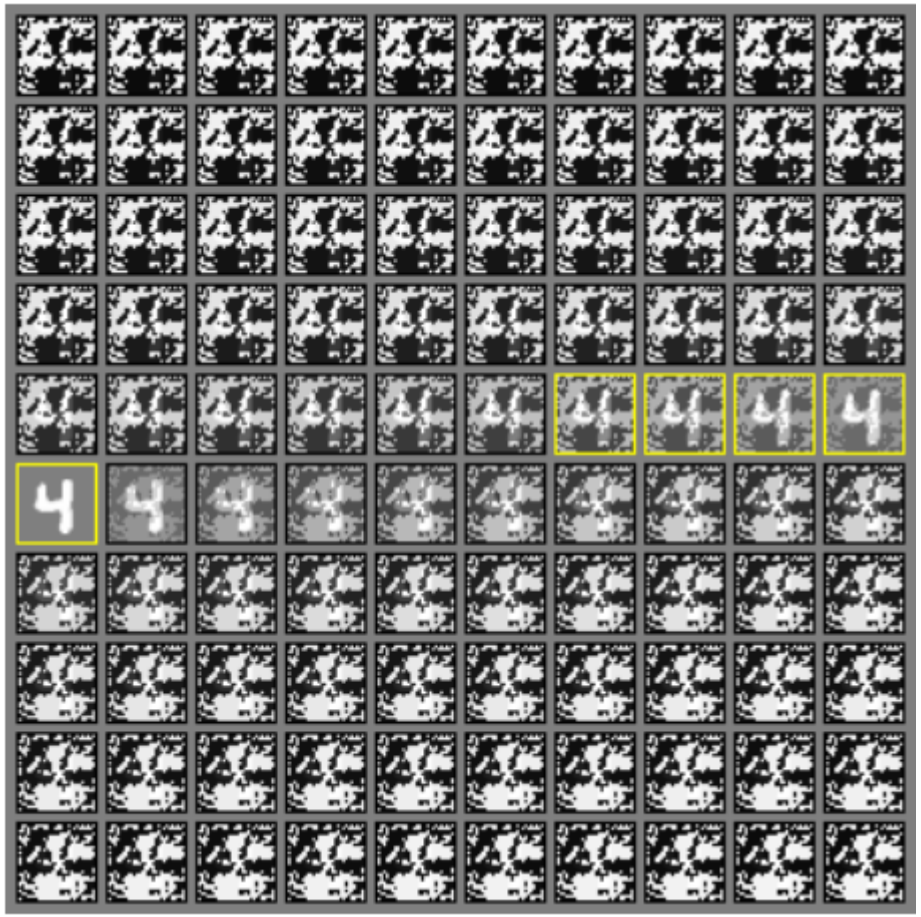


زیگموند



گوناگونی دقت





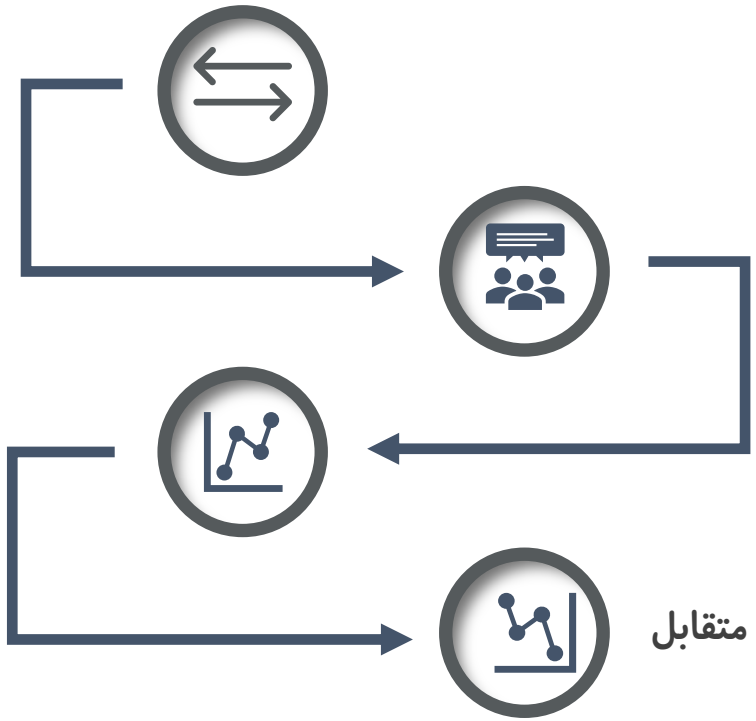
چرا با وجود سادگی در به غلط انداختن مدل‌های خطی، همچنان از آنها استفاده می‌کنیم؟

۱ بهینه‌سازی کم هزینه مدل‌های خطی

۲ وجود نداشتن روش دیگری برای آموزش

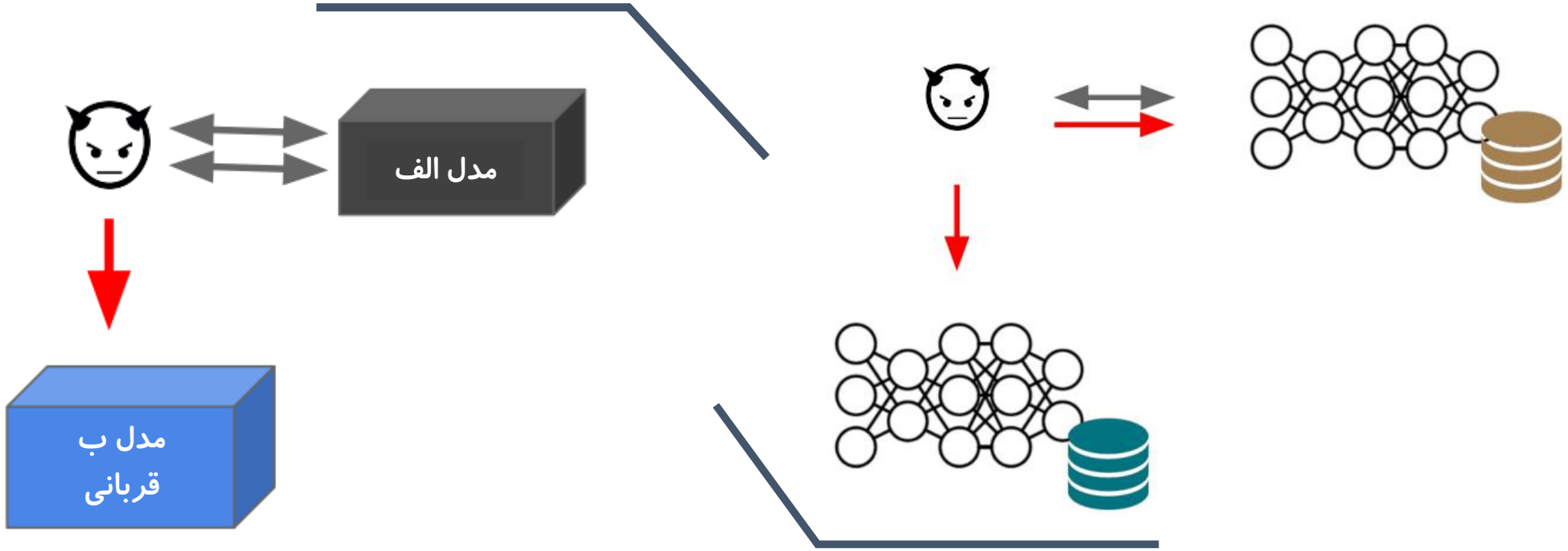
۳ آموزش اکثر مدل‌های خود را با استفاده از روش‌های مبتنی بر گرادیان

انتقال پذیری



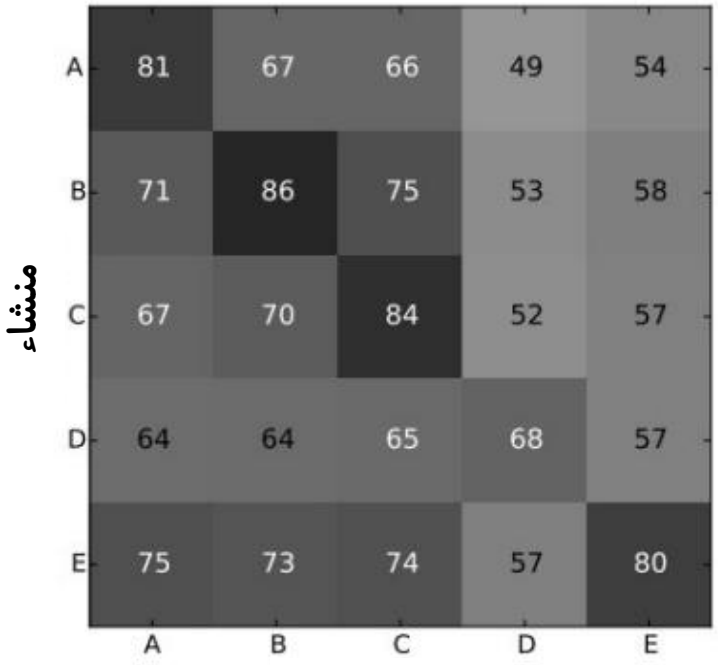
قابلیت انتقال مثال خصمانه

نمونه‌هایی که برای گمراه کردن مدل اول ساخته شده‌اند، احتمالاً مدل دوم را نیز گمراه می‌کنند.



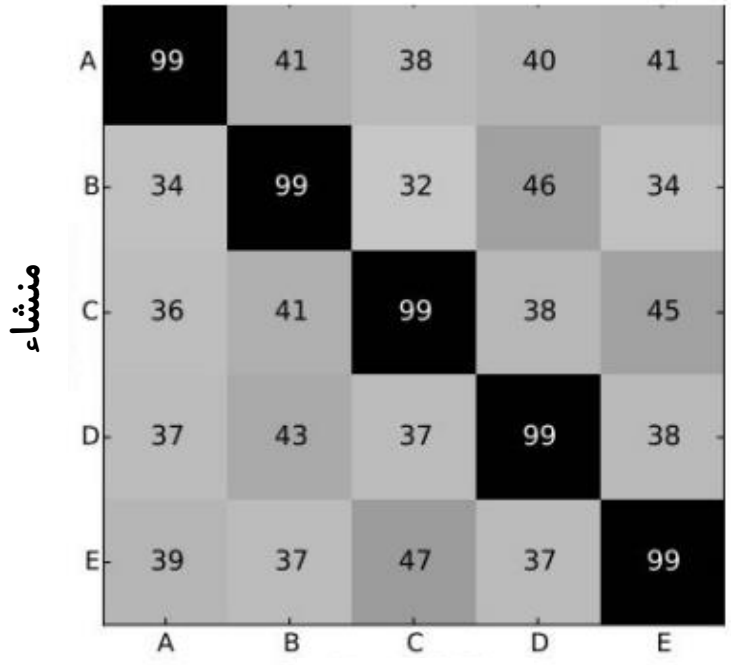
داده آموزشی متقابل

شبکه‌های عمیق



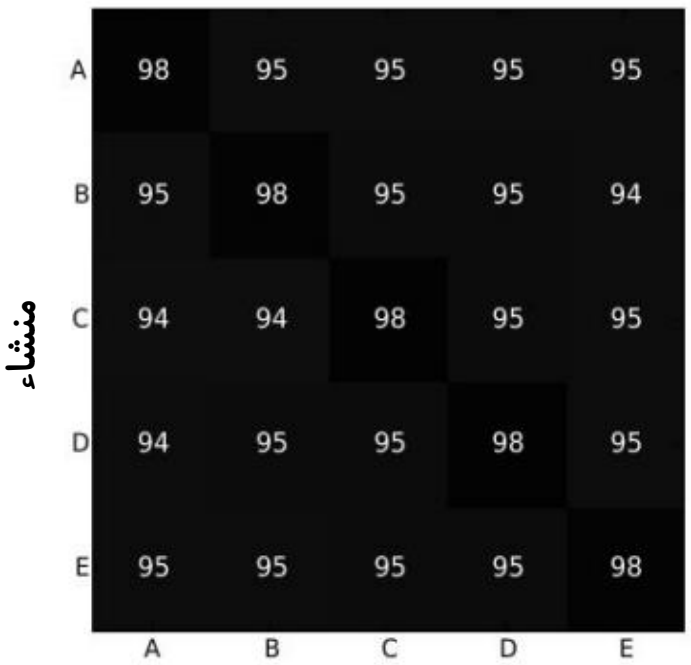
متوسط

ماشین بردار پشتیبان



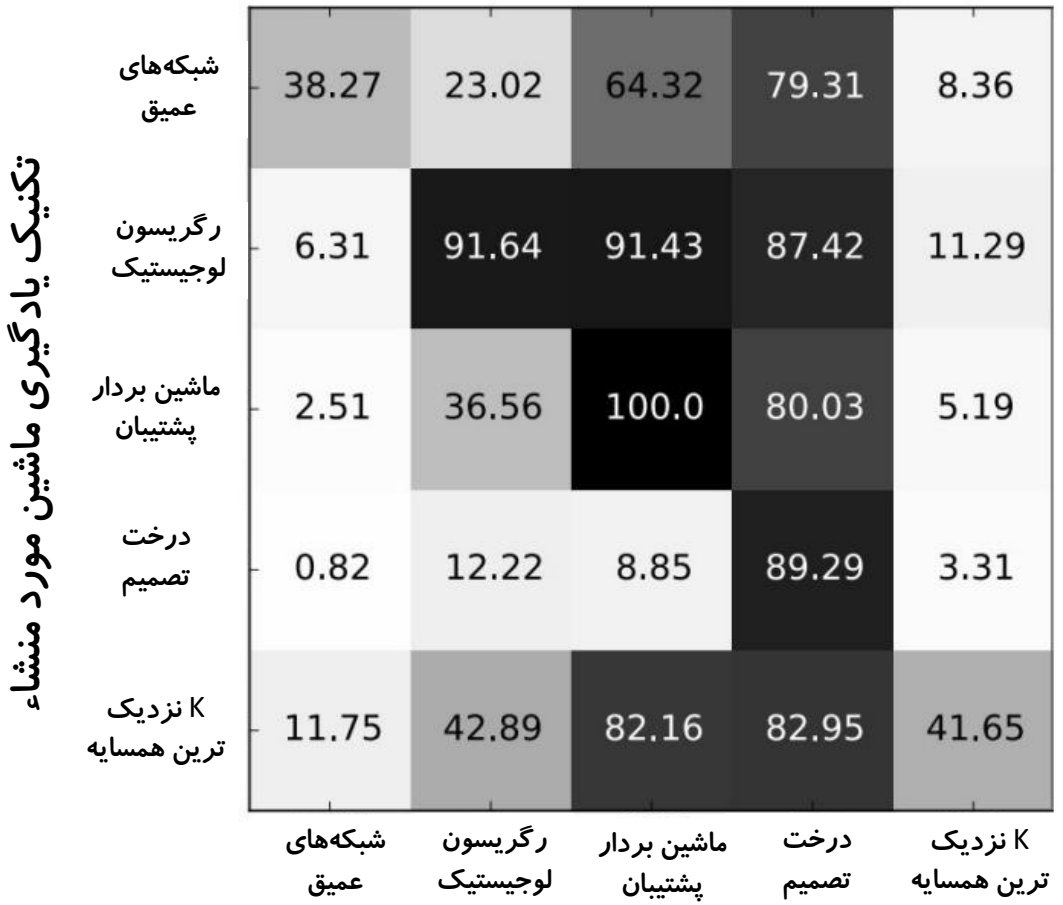
ضعیف

رگریسون لوجیستیک



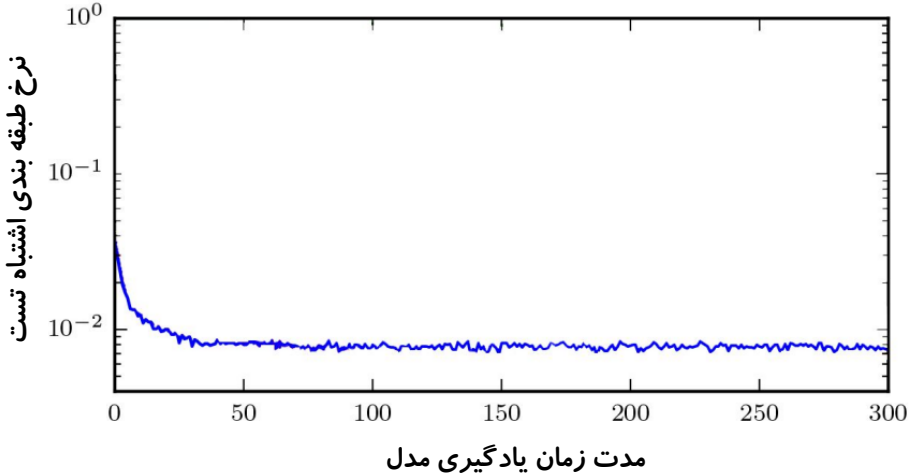
قوی

روش آموزشی متقابل

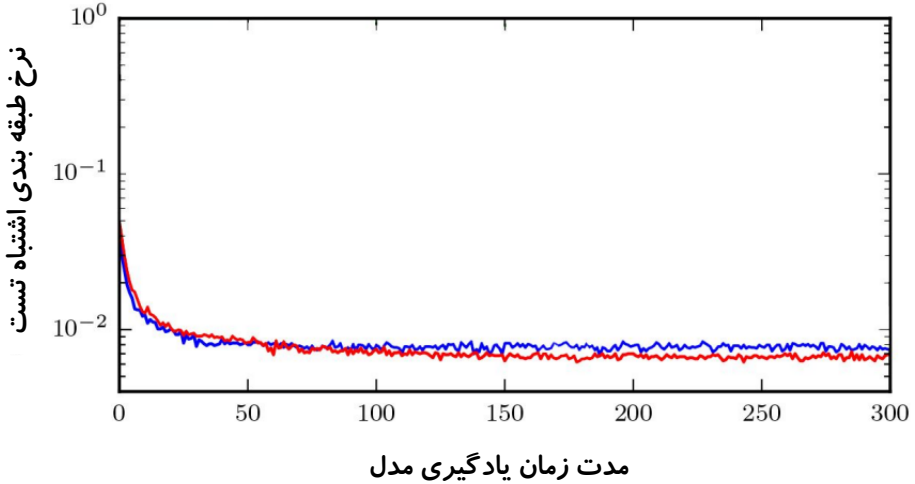


آموزش با نمونه‌های متخاصم

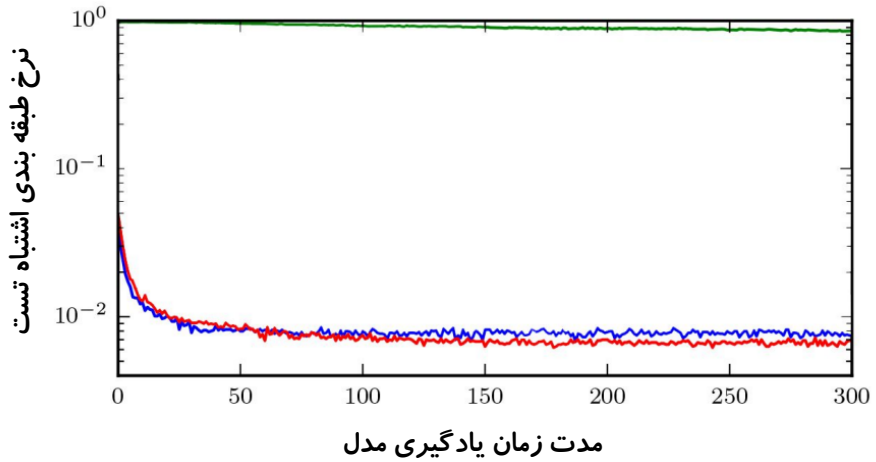
- یک مکانیزم دفاعی در برابر نمونه‌های متخاصم.
- مثال‌های خصمانه با برچسب درست را وارد داده آموزشی شود.
- با کمک این روش مدل یادگیری ماشین را تقویت می‌شود.
- در برابر حملات احتمالی مقاومت بهتری نشان دهد.



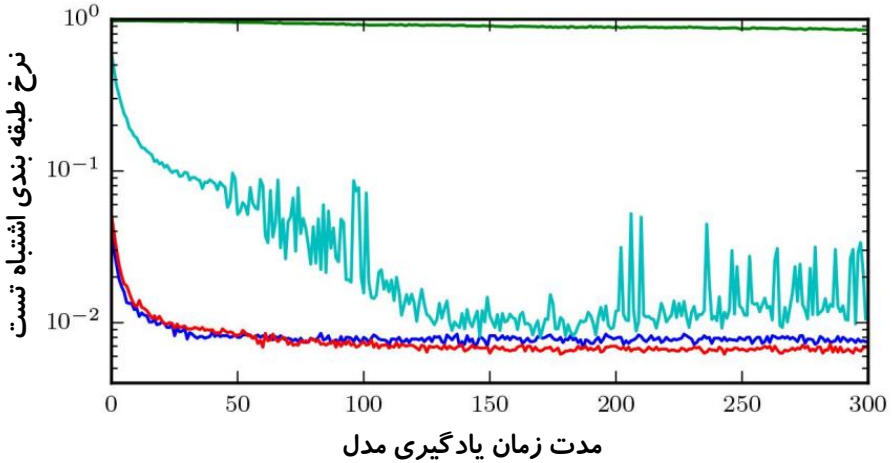
مدل آموزش دیده شده و تست شده با داده تمیز



مدل آموزش دیده شده با ترکیب داده تمیز و متخاصم
و تست شده با داده تمیز



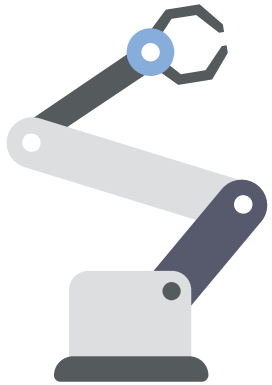
مدل آموزش دیده شده و تست شده با داده متخاصم



مدل آموزش دیده شده با ترکیب داده تمیز و متخاصم
و تست شده با داده متخاصم

آموزش با نمونه‌های متخاصم

- در اینجا به خوبی کار می‌کند زیرا همان حمله توسط مهاجم و طبقه‌بندی کننده استفاده می‌شود.
- تعمیم استحکام مدل به حملات تطبیقی دشوارتر است.
- طبقه بندی کننده باید از همه استراتژی های مهاجم آگاه باشد.
- آیا آموزش خصمانه فقط کتاب دستورالعمل را گسترش می دهد؟
- آیا مدل سازی مولد می تواند به درک بهتر منجر شود؟
- آیا یادگیری از طریق تعامل با یک محیط می تواند منجر به درک بهتر شود؟



جمع‌بندی

نتیجه‌گیری

۱ وجود مثال های متخاصم نشان می‌دهد که توانایی توضیح داده های آموزشی یا حتی توانایی برچسب گذاری صحیح داده های آزمایشی به این معنا نیست که مدل‌های ما واقعاً وظایفی را که از آنها خواسته‌ایم درک می‌کنند.

۲ پاسخ مدل در نقاطی که در توزیع داده‌ها رخ نمی‌دهند بیش از حد مطمئن است و این پیش‌بینی‌های مطمئن اغلب بسیار نادرست هستند.

۳ خانواده‌ای از روش های سریع برای تولید نمونه‌های متخاصم وجود دارد.

۴ می‌توان به کمک آموزش متخاصم تا حدی در برابر نمونه‌های متخاصم مقاومت کنیم و برای حملات احتمالی آمادگی بیشتری داشته باشیم.

۵ با اینکه می‌توان تا حدی در برابر حملات متخاصم مقاومت کرد؛ وجود نمونه‌های متخاصم بیان می‌کند خانواده‌های مدلی که ما استفاده می‌کنیم، ذاتاً ناقص هستند.

پیشنهادات

۱ ایده توسعه رویه‌های بهینه سازی که قادر به آموزش مدل‌هایی که رفتار آنها به صورت محلی پایدارتر است.

۲ مطالعات بیشتر به منظور شناخت روش‌های بیشتر برای تولید مثال‌های متخاصم

۳ مطالعات به منظور درک بهتر شبکه‌های عمیق به کمک مثال‌های متخاصم.

۴ استفاده از نمونه‌های متخاصم برای بهبود عملکرد مدل‌های مختلف در برابر حملات احتمالی.

منابع

[1] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever Joan Bruna Dumitru Erhan Ian Goodfellow Rob Fergus. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199, 2014b.

[2] Alexey Kurakin, Ian Goodfellow, Samy Bengio. Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533, 2016.

[3] Papernot, N., Adversarial Examples in Machine Learning, 2017.

[4] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).

[5] Goodfellow, Adversarial Examples. Re-Work Deep Learning Summit, 2015.

با تشکر از توجه شما