



دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)



بررسی و مهار مثال‌های خصمانه در یادگیری ماشین

ارائه دهنده: بردیا اردکانیان
استاد راهنما: دکتر رضا صفا بخش

اردیبهشت ۱۴۰۱

اهداف ارائه

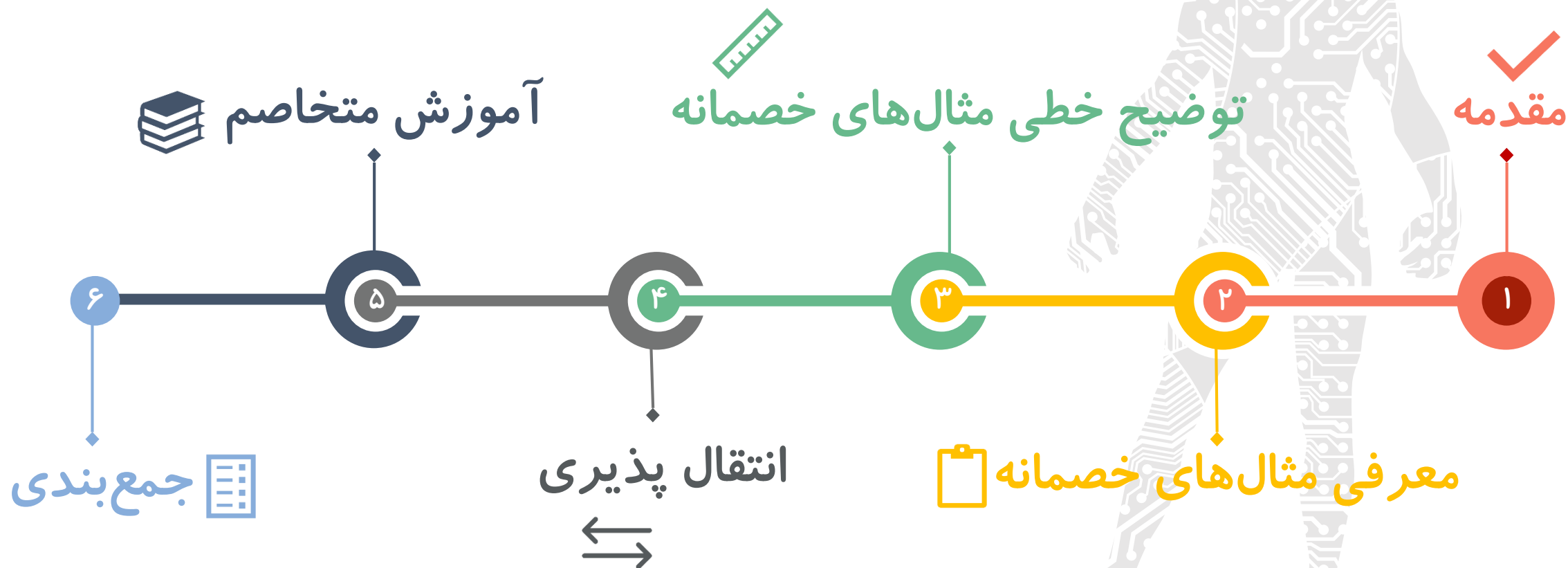
آشنایی مختصر مثال‌های خصمانه

توضیح خطی بودن مثال‌های خصمانه

بررسی تعمیم پذیری نمونه‌های متخاصم

دفاع از مدل‌های یادگیری ماشین

سیر ارائه



مقدمه



موفقیت‌های یادگیری ماشین



ماشین‌های خودران



کشف کلاهبرداری



FeatureSmith

تشخیص بدافزار



Google Cloud Platform

یادگیری ماشین به
عنوان یک سرویس

خطر شکست مدل‌های یادگیری ماشین

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

بینایی کامپیوتر

- 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)



تشخیص دهنده بدافزار

- در دسترس بودن مدل برای پرس و جوی فشرده
- دسترسی به برچسب و امتیاز مدل
- طبقه بندی کننده باینری (دو خروجی: بدافزار / بی خطر)

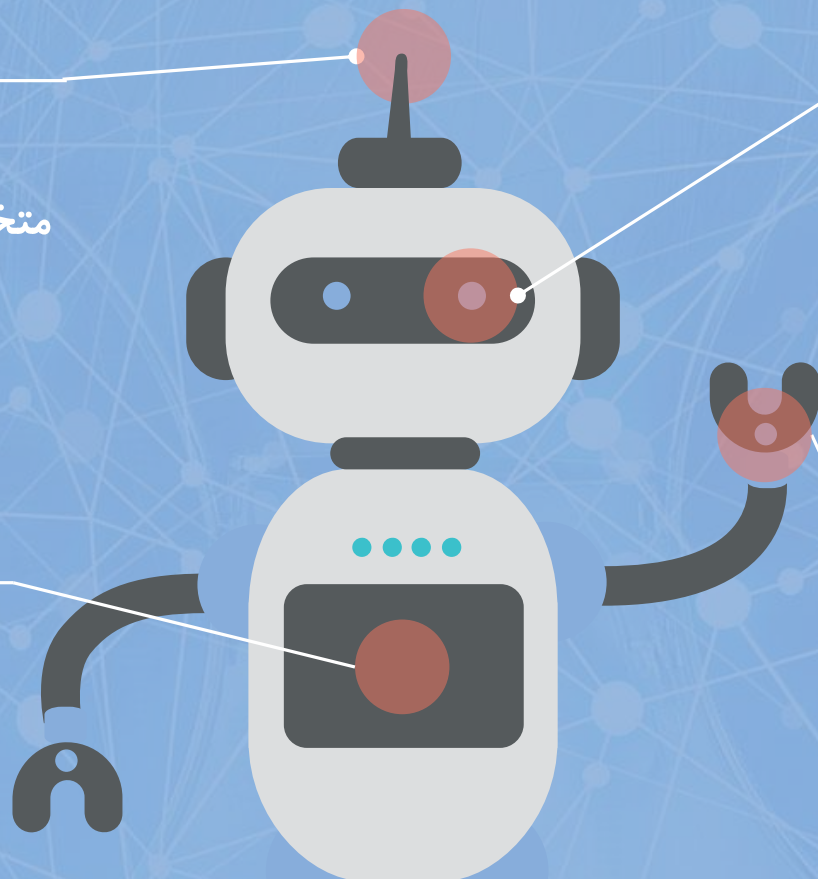
تحقیقات پیشین

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

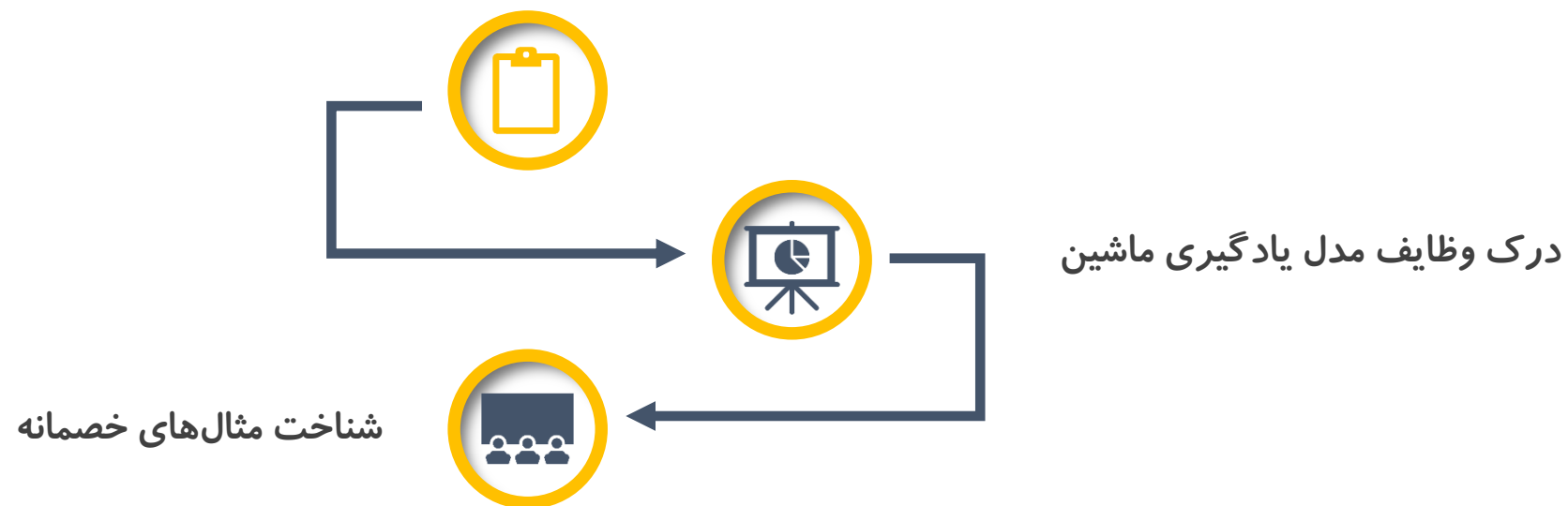
بیش از ۱۰۰ مقاله در زمینه مثال‌های متخاصم و آموزش متخاصم مورد مطالعه قرار گرفته است.

فرضیه‌ها نشان داده است این نتایج اغلب به عنوان یک نقص در شبکه‌های عمیق تفسیر می‌شوند

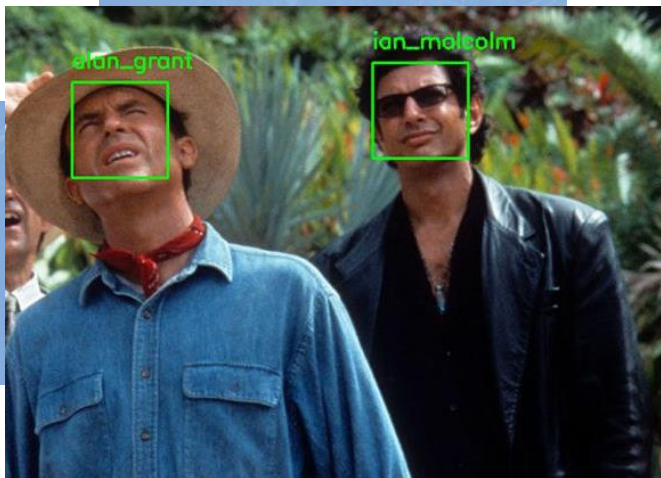
مطالعات صورت گرفته نشان داده‌اند که طبقه‌بندی کننده‌های مبتنی بر یادگیری ماشین مدرن در برابر مثال‌های متخاصم آسیب‌پذیر هستند.



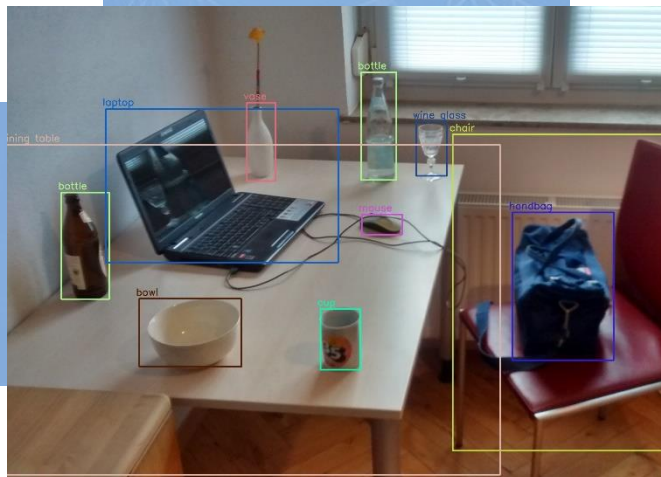
معرفی مثال‌های خصمانه



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 چه اتفاقی می افتد؟

你知道這段文字是什麼嗎？→

معرفی مثال‌های خصمانه



$\times 0.007 +$



$=$



«پاندا»

اطمینان ۵۷.۷ درصد

«نماد»

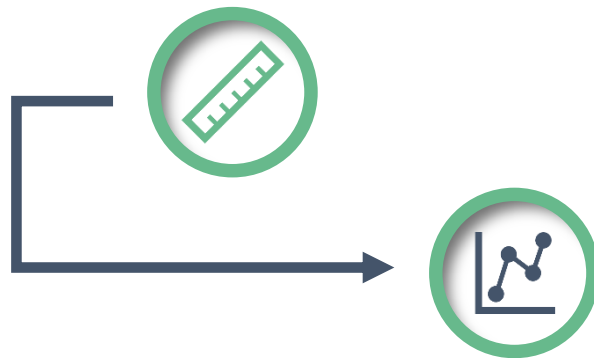
اطمینان ۸.۲ درصد

«گیبون»

اطمینان ۹۹.۳ درصد

مثال‌های خصمانه نشان‌دهنده کوچک‌ترین تغییرات دامنه هستند

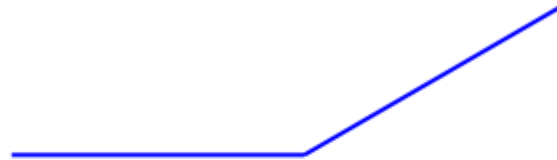
توضیح خطی مثال‌های خصمانه



خطی بودن تکه‌ای مدل‌های شبکه‌های عمیق

خطی بودن تکه‌ای مدل‌های شبکه‌های عمیق

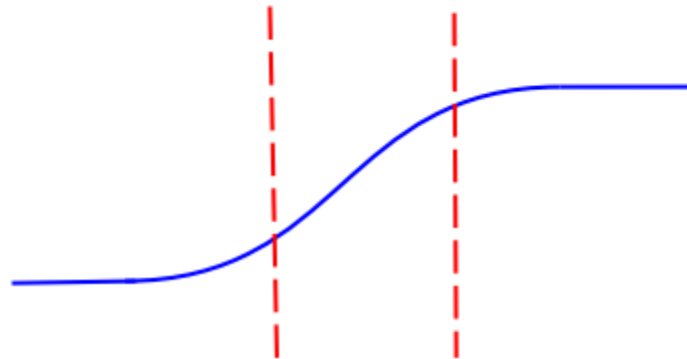
واحد یکسو شده خطی



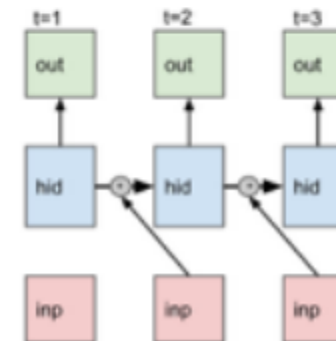
ماکس اوت



سیگموئید



شبکه عصبی بازگشتی



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

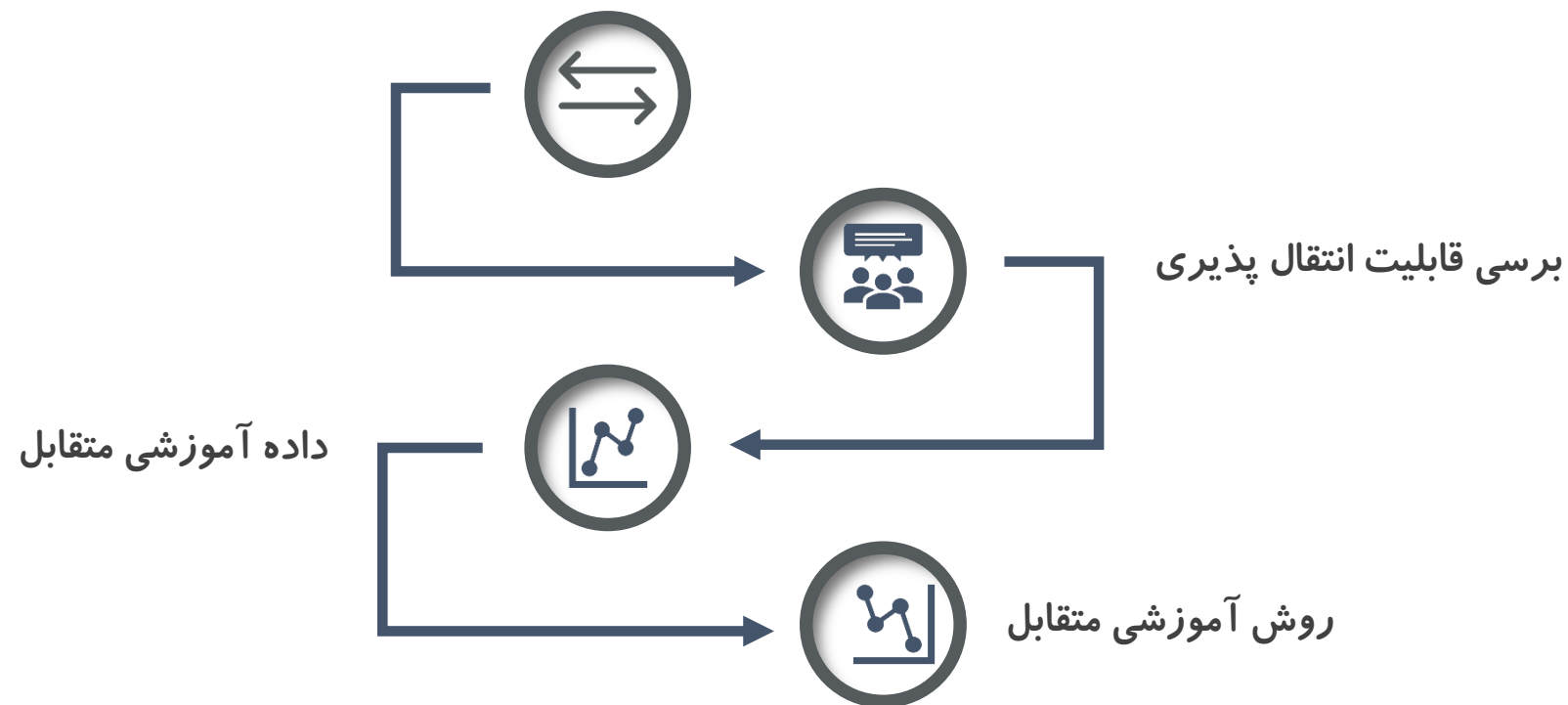
۱

بهینه‌سازی کم هزینه مدل‌های خطی

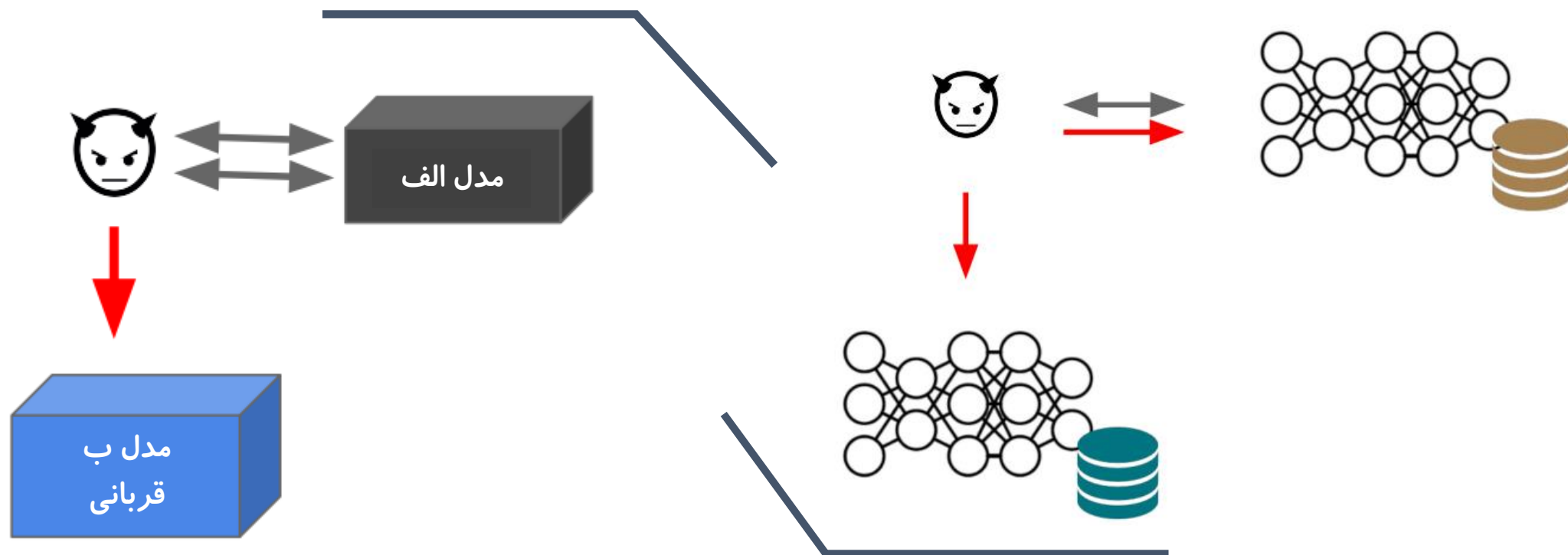
۲

وجود نداشتن روش دیگری برای آموزش

انتقال پذیری



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



داده آموزشی متقابل

رگريسون لوجیستیک

منشاء	A	98	95	95	95	95
	B	95	98	95	95	94
	C	94	94	98	95	95
	D	94	95	95	98	95
	E	95	95	95	95	98
		A	B	C	D	E

هدف

قوی

ماشین بردار پشتیبان

منشاء	A	99	41	38	40	41
	B	34	99	32	46	34
	C	36	41	99	38	45
	D	37	43	37	99	38
	E	39	37	47	37	99
		A	B	C	D	E

هدف

ضعیف

شبکه‌های عمیق

منشاء	A	81	67	66	49	54
	B	71	86	75	53	58
	C	67	70	84	52	57
	D	64	64	65	68	57
	E	75	73	74	57	80
		A	B	C	D	E

هدف

متوسط

روش آموزشی متقابل

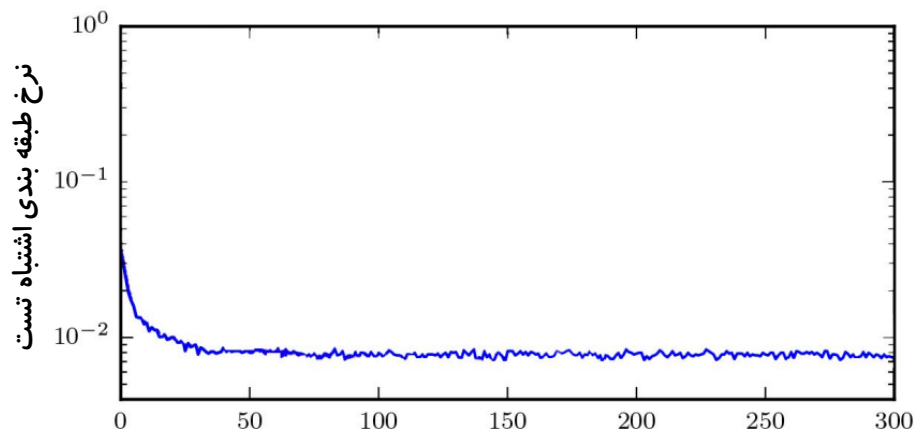
تکنیک یادگیری ماشین مورد منشاء	شبکه‌های عمیق	38.27	23.02	64.32	79.31	8.36
	رگرسیون لوجیستیک	6.31	91.64	91.43	87.42	11.29
	ماشین بردار پشتیبان	2.51	36.56	100.0	80.03	5.19
	درخت تصمیم	0.82	12.22	8.85	89.29	3.31
	K نزدیک ترین همسایه	11.75	42.89	82.16	82.95	41.65
	شبکه‌های عمیق	رگرسیون لوجیستیک	ماشین بردار پشتیبان	درخت تصمیم	K نزدیک ترین همسایه	

تکنیک یادگیری ماشین مورد هدف واقع شده

آموزش با نمونه‌های متخاصم

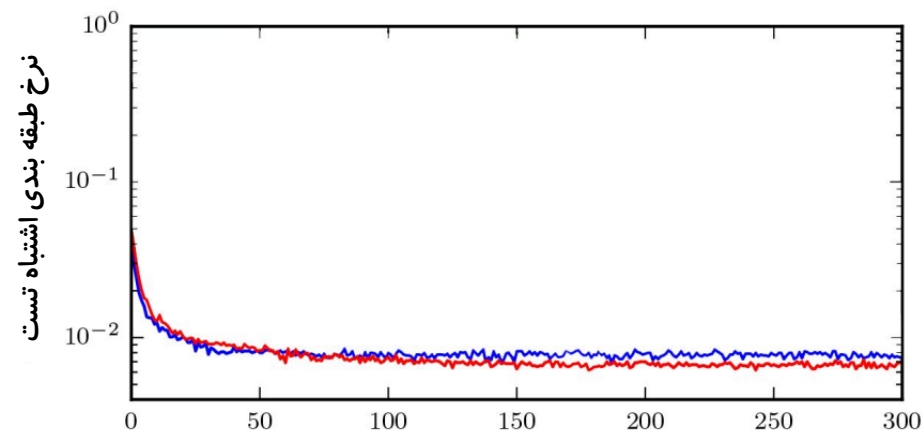


- یک مکانیزم دفاعی در برابر نمونه‌های متخاصم.
- مثال‌های خصمانه با برچسب درست را وارد داده آموزشی شود.
- با کمک این روش مدل یادگیری ماشین را تقویت می‌شود.
- در برابر حملات احتمالی مقاومت بهتری نشان دهد.



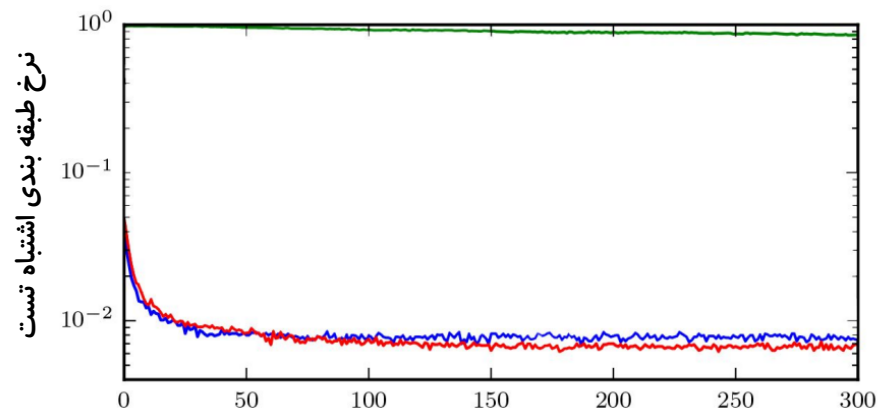
مدت زمان یادگیری مدل

مدل آموزش دیده شده و تست شده با داده تمیز



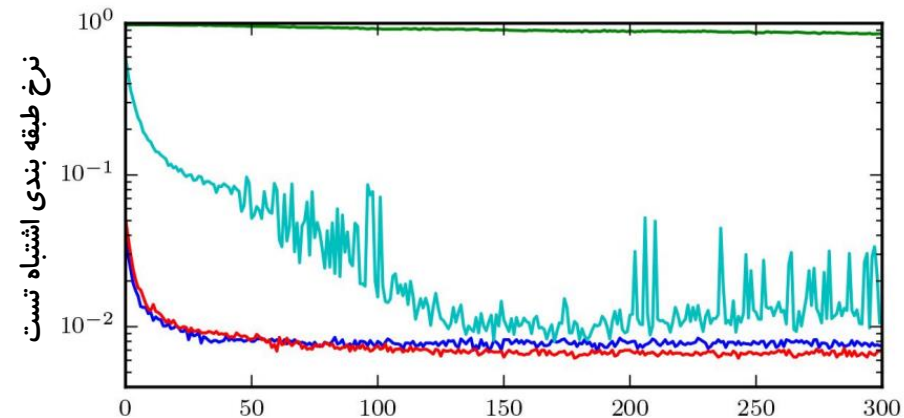
مدت زمان یادگیری مدل

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



مدت زمان یادگیری مدل

مدل آموزش دیده شده و تست شده با داده متخاصم

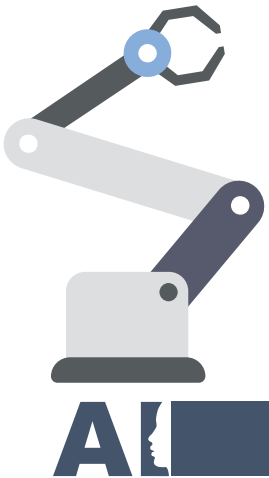


مدت زمان یادگیری مدل

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

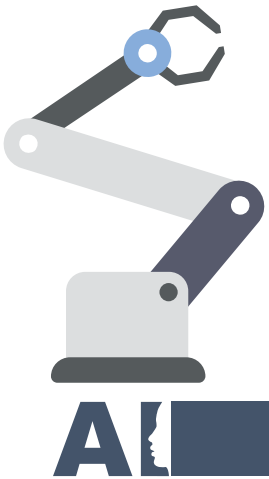
آموزش با نمونه‌های متخاصم

- در اینجا به خوبی کار می‌کند زیرا همان حمله توسط مهاجم و طبقه‌بندی کننده استفاده می‌شود.
- طبقه بندی کننده باید از همه استراتژی های مهاجم آگاه باشد.



آموزش با نمونه‌های متخصص

- آیا آموزش خصمانه فقط کتاب دستورالعمل را گسترش می‌دهد؟
- آیا مدل سازی مولد می‌تواند به درک بهتر منجر شود؟



جمع‌بندی



نتیجه گیری

۱

وجود مثال های متخاصم نشان می دهد که توانایی توضیح داده های آموزشی یا حتی توانایی برچسب گذاری صحیح داده های آزمایشی به این معنا نیست که مدل های ما واقعاً وظایفی را که از آنها خواسته ایم درک می کنند.

۲

پاسخ مدل در نقاطی که در توزیع داده ها رخ نمی دهند بیش از حد مطمئن است و این پیش بینی های مطمئن اغلب بسیار نادرست هستند.

۳

می توان به کمک آموزش متخاصم تا حدی در برابر نمونه های متخاصم مقاومت کنیم و برای حملات احتمالی آمادگی بیشتری داشته باشیم.

۴

با اینکه می توان تا حدی در برابر حملات متخاصم مقاومت کرد؛ وجود نمونه های متخاصم بیان می کند خانواده های مدلی که ما استفاده می کنیم، ذاتاً ناقص هستند.

پیشنهادهات

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

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

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

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