Deep Learning & AI News #7

Interesting Things Related to Deep Learning

Mar 13th, 2021
https://github.com/NoAchache/TextBoxGan

* adapted from StyleGan2
M6: A Chinese Multimodal Pretrainer

Junyang Lin, Rui Men, An Yang, Chang Zhou, Ming Ding, Yichang Zhang, Peng Wang, Ang Wang, Le Jiang, Xianyan Jia, Jie Zhang, Jianwei Zhang, Xu Zou, Zhikang Li, Xiaodong Deng, Jie Liu, Jinbao Xue, Huiling Zhou, Jianxin Ma, Jin Yu, Yong Li, Wei Lin, Jingren Zhou, Jie Tang, Hongxia Yang


- Collaboration between Alibaba and Tsinghua University
- 1.9 Tb of images + 292 Gb of text
- Chinese, not English
- Trained 10 and 100 billion parameter transformers
- Pre-trained model can be used for many tasks: generating descriptions, image search, question answering, poem generation etc.

Table 1: Statistics of our pretraining dataset. We demonstrate the sources of our data, and we calculate the number of images, tokens, and passages, the average length, as well as the size of image and text.

<table>
<thead>
<tr>
<th>Source Modality</th>
<th>Images (M)</th>
<th>Tokens (B)</th>
<th>Passages (M)</th>
<th>Avg. Length</th>
<th>Image Size (TB)</th>
<th>Text Size (GB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Encyclopedia</td>
<td>31.4</td>
<td>34.0</td>
<td>923.5</td>
<td></td>
<td></td>
<td>65.1</td>
</tr>
<tr>
<td>Community QA</td>
<td>13.9</td>
<td>113.0</td>
<td>123.0</td>
<td></td>
<td></td>
<td>28.8</td>
</tr>
<tr>
<td>Forum discussion</td>
<td>8.7</td>
<td>39.0</td>
<td>223.1</td>
<td></td>
<td></td>
<td>18.0</td>
</tr>
<tr>
<td>Common Crawl</td>
<td>40.3</td>
<td>108.7</td>
<td>370.7</td>
<td></td>
<td></td>
<td>83.3</td>
</tr>
<tr>
<td>Encyclopedia &amp; Text</td>
<td>6.5</td>
<td>7.9</td>
<td>10.4</td>
<td>759.6</td>
<td>0.1</td>
<td>15.0</td>
</tr>
<tr>
<td>Crawled Webpages</td>
<td>46.0</td>
<td>9.1</td>
<td>106.0</td>
<td>85.8</td>
<td>1.5</td>
<td>70.0</td>
</tr>
<tr>
<td>E-commerce</td>
<td>8.0</td>
<td>0.5</td>
<td>8.5</td>
<td>62.1</td>
<td>0.3</td>
<td>12.2</td>
</tr>
<tr>
<td>Total</td>
<td>60.5</td>
<td>111.8</td>
<td>419.6</td>
<td>266.4</td>
<td>1.9</td>
<td>292.4</td>
</tr>
</tbody>
</table>

Table 2: Comparison with the existing large-scale Chinese corpora for pretraining. Our dataset is the largest dataset for Chinese pretraining. The size of texts is larger than that of the existing datasets, and the size of images is even larger than that of ImageNet.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Text Size (GB)</th>
<th>Image Size (GB)</th>
<th>Multidomain</th>
</tr>
</thead>
<tbody>
<tr>
<td>CN-Wikipedia</td>
<td>1.6</td>
<td>✖</td>
<td>✖</td>
</tr>
<tr>
<td>THUCTC</td>
<td>2.2</td>
<td>✖</td>
<td>✖</td>
</tr>
<tr>
<td>HFL</td>
<td>21.6</td>
<td>✖</td>
<td>✖</td>
</tr>
<tr>
<td>CLUE Corpus</td>
<td>100.0</td>
<td>✖</td>
<td>✖</td>
</tr>
<tr>
<td>ImageNet</td>
<td>✖</td>
<td>✖</td>
<td>✔</td>
</tr>
<tr>
<td>M6-Corpus</td>
<td>292.4</td>
<td>1900</td>
<td>✔</td>
</tr>
</tbody>
</table>

3.1 Visual and Linguistic Inputs

The mainstream multimodal pretraining methods transform images to feature sequences via object detection. However, the performance of the object detectors as well as the expressivity of their backbones strongly impact the final performance of the pretrained models in the downstream tasks. We observe that a large proportion of the images contain only a few objects. Take the images of the data of e-commerce as an example. We randomly sample 1M images and perform object detection on the images. The results show that over 90% of the images contain fewer than 5 objects. Also, the objects have high overlapping with each other. To alleviate such influence, we turn to a simple but effective solution following Gao et al. [11] and Dosovitskiy et al. [8]. In general, we split an image into patches and extract features of the 2D patches with a trained feature extractor, say ResNet-50. Then we line up the representations to a sequence by their positions. The processing of the input word sequence is much simpler. We follow the similar preprocessing procedures in the previous work [4, 10, 22]. We apply WordPiece [32, 43] and masking to the word sequence and embed them with an embedding layer, following BERT [6].
Figure 3: An overview of the pretraining tasks for M6. The design of masking strategies allows the learning of different tasks under the same framework. M6 is pretrained with image-based text denoising, image captioning, text denoising, and language modeling.
Phantom of the ADAS: Phantom Attacks on Driver-Assistance Systems

Ben Nassi¹, Dudi Nassi¹, Raz Ben-Netanel¹, Yisroel Mirsky¹,², Oleg Drokin³, Yuval Elovici¹

Video Demonstration - https://youtu.be/1cSw4fXYqWI
{nassib,nassid,razx,yisroel,elovici}@post.bgu.ac.il, green@linuxhacker.ru
¹ Ben-Gurion University of the Negev, ² Georgia Tech, ³ Independent Tesla Researcher

Abstract

The absence of deployed vehicular communication systems, which prevents the advanced driving assistance systems (ADASs) and autopilots of semi/fully autonomous cars to validate their virtual perception regarding the physical environment surrounding the car with a third party, has been exploited in various attacks suggested by researchers. Since the application of these attacks comes with a cost (exposure of the attacker's identity), the delicate exposure vs. application balance has held, and attacks of this kind have not yet been encountered in the wild. In this paper, we investigate a new perceptual challenge that causes the ADASs and autopilots of semi/fully autonomous to consider depthless objects (phantoms) as real. We show how attackers can exploit this perceptual challenge to apply phantom attacks and change the abovementioned balance, without the need to physically approach the attack scene, by projecting a phantom via a drone equipped with a portable projector or by presenting a phantom on a hacked digital billboard that faces the Internet and is located near roads. We show that the car industry has not considered this type of attack by demonstrating the attack on today's most advanced ADAS and autopilot technologies: Mobileye 630 PRO and the Tesla Model X, HW 2.5; our experiments show that when presented with various phantoms, a car's ADAS or autopilot considers the phantoms as real objects, causing these systems to trigger the brakes, steer into the lane of oncoming traffic, and issue notifications about fake road signs. In order to mitigate this attack, we present a model that analyzes a detected object's context, surface, and reflected light, which is capable of detecting phantoms with 0.99 AUC. Finally, we explain why the deployment of vehicular communication systems might reduce attackers' opportunities to apply phantom attacks but won't eliminate them.

I. Introduction

After years of research and development, automobile technology is rapidly approaching the point at which human drivers can be replaced, as cars are now capable of supporting semi/fully autonomous driving [1, 2]. While the deployment of semi/fully autonomous cars has already begun in many countries around the world, the deployment of vehicular communication systems [3], a set of protocols intended for exchanging information between vehicles and roadside units, has been delayed [4]. The eventual deployment of such systems, which include V2V (vehicle-to-vehicle), V2I (vehicle-to-infrastructure), V2P (vehicle-to-pedestrian), and V2X (vehicle-to-everything) communication systems, is intended to supply semi/fully autonomous cars with information and validation regarding lanes, road signs, and obstacles.

Given the delayed deployment of vehicular communication systems in most places around the world, autonomous driving largely relies on sensor fusion to replace human drivers. Passive and active sensors are used in order to create 360 3D virtual perception of the physical environment surrounding the car. However, the lack of vehicular communication system deployment has created a validation gap which limits the ability of semi/fully autonomous cars to validate their virtual perception of obstacles and lane markings with a third party, requiring them to rely solely on their sensors and validate one sensor's measurements with another. Given that the exploitation of this gap threatens the security of semi/fully autonomous cars, we ask the following question: Why haven’t attacks against semi/fully autonomous cars exploiting this validation gap been encountered in the wild?

Various attacks have already been demonstrated by researchers [5–14], causing cars to misclassify road signs [5–10], misperceive objects [11, 12], deviate to the lane of oncoming traffic [13], and navigate in the wrong direction [14]. These attacks can only be applied by skilled attackers (e.g., an expert

Adversarial Laser Beam: Effective Physical-World Attack to DNNs in a Blink

Ranjie Duan†† Xiaofeng Mao‡ A. K. Qin† Yun Yang† Yuefeng Chen‡ Shaokai Ye‡‡ Yuan He‡
†Swinburne University of Technology ‡Alibaba Group ‡‡EPFL

Abstract

Though it is well known that the performance of deep neural networks (DNNs) degrades under certain light conditions, there exists no study on the threats of light beams emitted from some physical source as adversarial attacker on DNNs in a real-world scenario. In this work, we show by simply using a laser beam that DNNs are easily fooled. To this end, we propose a novel attack method called Adversarial Laser Beam (AdvLB), which enables manipulation of laser beam’s physical parameters to perform adversarial attack. Experiments demonstrate the effectiveness of our proposed approach in both digital- and physical-settings. We further empirically analyze the evaluation results and reveal that the proposed laser beam attack may lead to some interesting prediction errors of the state-of-the-art DNNs. We envisage that the proposed AdvLB method enriches the current family of adversarial attacks and builds the foundation for future robustness studies for light.

1. Introduction

Natural phenomena may play the role of adversarial attackers, e.g., a blinding glare results in a fatal crash of a Tesla self-driving car. What if a beam of light can adversarially attack a DNN? Further, how about using a beam of light, specifically the laser beam, as the weapon to perform attacks. If we can do that, with the fastest speed in the world, the laser beam could achieve the fastest attack with no doubts. As shown in Figure 1, by using an off-the-shelf lighting device such as a laser pointer, the attacker can maliciously shoot a laser beam onto the target object to make the self-driving car fail to recognize target objects correctly.

We regard the attack illustrated in Figure 1 as a new type of adversarial attack, which is crucial but not yet exploited. Up to now, most researchers study the security of DNNs by exploring various adversarial attacks in digital-settings, where input images are added with deliberately crafted perturbations and then fed to the target DNN model [23, 10, 6, 3, 18]. However, in physical-world scenarios, images are typically captured by cameras and then directly fed to the target models, where attackers cannot directly manipulate the input image. Some recent efforts in developing physical-world attacks are addressed in [21, 8, 2, 7, 14].

The physical-world adversarial examples typically require large perturbations, because small perturbations are hard to be captured by cameras. In addition, the attacking effects of adversarial examples of small perturbations can be easily mitigated in complex physical-world environments [21, 9, 7]. Meanwhile, physical-world adversarial examples require high stealthiness to avoid being discovered by either the victim or defender before performing an attack successfully. Thus for creating physical-world adversarial examples, there is always a compromise between stealthiness and adversarial strength.

Most existing physical-world attacks adopt a “sticker-pasting” setting, i.e., the attacker prints adversarial perturbation as a sticker and then pastes it onto the target object [16, 2, 7, 8]. These attacks achieve the stealthiness of adversaries with extra efforts of designing adversarial perturbation or camouflaging adversarial images and finding

Figure 1: An example. When the camera of self-driving car captures object shot by the laser beam, it recognizes “trolleybus” as “amphibian” and “street sign” as “soap dispenser”.
5.2 Label fairness

As with most approaches to fair supervised learning, the approach described in the previous section assumes the outcome being predicted \( Y \) is measured accurately in the data used to assess the system. There are some cases where this assumption is reasonable—for example, websites can perfectly measure whether users click on a given button. However, in many cases, such as identifying bullying, the labels themselves are generated through human judgment, and may thus embed human biases. This is of concern for at least three reasons. First, accurate labels are needed to compute most model fairness metrics, including the metric in Section 5.1. Second, supervised learning systems trained on biased labels will learn those biases. Finally, labelers’ decisions might be used to directly intervene in the system. In this section, we describe how our high level fairness approach can be applied to assess human decision making, in the case where decisions can be compared to a ground truth.

In our bullying and harassment example, the decision being made by human labelers is whether a given post violates a bullying policy as written. These decisions won’t always be correct—labelers may misunderstand the policy or the post, make a mistake, or be misled by implicit or explicit biases. To track these errors, we also collect (for a subset of posts) the judgement of an expert in applying the written policy, whose decisions provide the ground truth for each
The Deep Bootstrap Framework: Good Online Learners are Good Offline Generalizers: https://arxiv.org/abs/2010.08127

- **Real World** \((N, T)\): Train a model on \(N\) train samples from a distribution, for \(T\) minibatch stochastic gradient descent (SGD) steps, re-using the same \(N\) samples in multiple epochs, as usual. This corresponds to running SGD on the empirical loss (loss on training data), and is the standard training procedure in supervised learning.

- **Ideal World** \((T)\): Train the same model for \(T\) steps, but use fresh samples from the distribution in each SGD step. That is, we run the exact same training code (same optimizer, learning-rates, batch-size, etc.), but sample a fresh train set in each epoch instead of reusing samples. In this ideal world setting, with an effectively infinite "train set", there is no difference between train error and test error.

**CIFAR-5m.** We construct a dataset of 6 million synthetic CIFAR-10-like images, by sampling from the CIFAR-10 Denoising Diffusion generative model of **Ho et al. (2020)**, and labeling the unconditional samples by a 98.5% accurate Big-Transfer model (**Kolesnikov et al., 2019**). These are

We now claim that for all \(t\) until the Real World converges, these two models \(f_t, f^\text{iid}_t\) have similar test performance. In our main claims, we differ slightly from the presentation in the Introduction in that we consider the “soft-error” of classifiers instead of their hard-errors. The soft-accuracy of classifiers is defined as the softmax probability on the correct label, and (soft-error) := 1 − (soft-accuracy).
In order to quantify this observation, we simulated an ideal world setting by creating a new dataset, which we call CIFAR-5m. We trained a generative model on CIFAR-10, which we then used to generate ~6 million images. The scale of the dataset was chosen to ensure that it is “virtually infinite” from the model’s perspective, so that the model never resamples the same data. That is, in the ideal world, the model sees an entirely fresh set of samples.

The real world model is trained on 50K samples for 100 epochs, and the ideal world model is trained on 5M samples for a single epoch. The lines show the test error vs. the number of SGD steps.
Effect of pre-training — pre-trained ViTs optimize faster in the ideal world.
SEER: The start of a more powerful, flexible, and accessible era for computer vision

March 4, 2021

https://ai.facebook.com/blog/seer-the-start-of-a-more-powerful-flexible-and-accessible-era-for-computer-vision


- SEER = SEIf-suppERvised
- new billion-parameter self-supervised computer vision model
- pretraining on a billion random, unlabeled and uncurated public Instagram images
- self-supervised SOTA: reaching 84.2 percent top-1 accuracy on ImageNet
- SwAV (https://arxiv.org/abs/2006.09882) uses online clustering to rapidly group images with similar visual concepts and leverage their similarities (doesn't need pair-wise comparisons; fast)

A LIBRARY FOR STATE-OF-THE-ART SELF-SUPERVISED LEARNING

GET STARTED  TUTORIALS  DOCS  GITHUB

Powered by PyTorch
Built on top of PyTorch which allows using all

SOTA Self-Supervision methods

Benchmark tasks
Variety of benchmarks tasks (linear image

Scalable
Easy to train model on 1-gpu, multi-gpu and

https://vissl.ai