Retrieval as a defense mechanism against adversarial examples in convolutional neural networks

Junbo Zhao
New York University
New York, NY
jakezhao@cs.nyu.edu

Jinyang Li
New York University
New York, NY
jinyang@cs.nyu.edu

Kyunghyun Cho
New York University
Facebook AI Research
New York, NY
kyunghyun.cho@nyu.edu

1 INTRODUCTION

Most, if not all, of the recently successful object recognition systems are built as an end-to-end, fully-differentiable convolutional neural networks [10, 12, 14]. Such an use of end-to-end differentiable models allows us to use an efficient gradient-based learning algorithm to train a large-scale object recognition system. One consequence of this paradigm is that the gradient of the network's output could be computed not only with respect to the network's parameters but also with respect to the input pixels. On the positive side, this provides us an efficient and effective way to visualize the internal working of such a deep neural network [see, e.g., 19]. On the other hand, it has been recently found that this gradient information could be used to create a so-called adversarial example that *fools* the network [see, e.g., 15, 17].

In this abstract, we propose a proactive defense mechanism against adversarial examples by incorporating an off-the-shelf non-differentiable retrieval engine as a part of a deep convolutional neural network, motivated by recently proposed retrieval-based approaches to machine translation [8], text classification [18] and language modeling [9]. Instead of feeding in a given input image x to a convolutional network as it is, we feed in a set of similar images $\{x_1, \ldots, x_N\}$, that collectively represent the original input image x, retrieved by an external retrieval engine. These retrieved images are combined by the way of attention mechanism [1] into a single vector before being classified. We conjecture that this use of non-differentiable engine renders the gradient of the network output with respect to the input pixels less useful for the purpose of generating an adversarial example.

A core challenge of the proposed approach is the efficiency of the external retrieval mechanism. We use locality sensitive hashing [LSH; see, e.g., 6] with a reduced feature dimension using random projection [see, e.g., 3, and references therein]. Since the proposed framework works with any off-the-shelf retrieval mechanism, it is possible to scale up the entire system by using latest advances from distributed processing and information retrieval, which we leave as future work.

We evaluate the proposed framework on SVHN [16] against the fast gradient sign method [FGSM, 7] and its iterative variant [iFGSM, 13]. Our preliminary results indicate that the proposed approach of incorporating a non-differentiable retrieval engine is a promising step toward building a deep convolutional network more robust to adversarial examples.

SysML'18, Feb 2018, Stanford, CA
2018. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 RETRIEVAL-AUGMENTED CONVOLUTIONAL NETWORKS

Given an input image x, the proposed classifier goes through three stages; retrieval, attention and classification. Here, we give a brief description of each stage. In Fig. 1, we graphically illustrate this entire process.

Retrieval. It has now become standard to use a feature vector from a deep convolutional neural network for image retrieval, as it has been found to reflect similarities better than the pixel-level straight-forward Euclidean distance [see, e.g., 2, 11]. We thus transform the input image x into a feature vector h by using a pretrained, conventional convolutional network. Such a feature vector h often has thousands dimensions, and it becomes impractical to use the full feature vector to retrieve nearest neighboring examples. In order to reduce the computational overhead, we randomly project this feature vector into a lower-dimensional space, i.e., $h' = Wh$, where W is a random Gaussian matrix. We use LSH for efficient retrieval of similar images from an entire training set using this randomly projected feature vector h'.[1] This results in a set of N retrieved images $\{(x_1, h_1), \ldots, (x_N, h_N)\}$, where N is a meta-parameter that should be decided based on generalization error as well as computational overhead.

Attention. The retrieved images are summarized, or combined, into a single vector with respect to the input image using the attention mechanism [1]. The resulting vector is the concatenation of a convex sum of the feature vectors of the retrieved images, i.e., $\bar{h} = \sum_{n=1}^{N} \alpha_n h_n$, and a convex sum of their label vectors (one-hot vectors), i.e., $\bar{y} = \sum_{n=1}^{N} \alpha_n y_n$. The coefficients are computed by

$$\alpha_n = \frac{\exp(\beta_n)}{\sum_{n=1}^{N} \exp(\beta_n)},$$

and $\beta_n = h^\top U h_n$ is an unnormalized score of the n-th retrieved example against the original input image. This process allows the network to put more emphasis on retrieved examples that are more similar or relevant to the original input while ignoring some that are irrelevant.

Classification. The resulting vector $\bar{h}$ from the attention mechanism is then fed through several fully-connected layers followed by a softmax [4] layer to arrive at the final predictive distribution.

[1] We note however that this specific choice of retrieval mechanism is not necessary and that better computing infrastructure and underlying system may allow us to use a full feature with a growing candidate set.

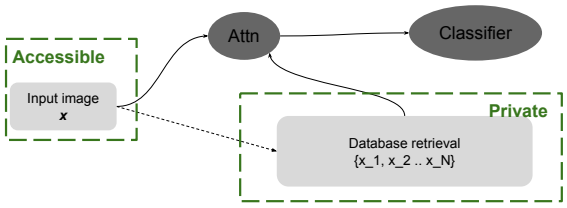

Figure 1: The proposed framework consists of three stages. The first stage is retrieval during which the private database is used to efficiently retrieve nearest representative images. The second and third stages are attention and classification. See text for more details.

3 PROACTIVELY ADDRESSING ADVERSARIAL EXAMPLES

Fast Gradient Sign Method (FGSM). Here, we consider a white-box setting in which an attacker has access to both the network architecture and parameters. In this case, Goodfellow et al. [7] proposed an efficient algorithm for generating an adversarial example that exploits the availability of the gradient w.r.t. the input, called the fast gradient sign method (FGSM). It modifies any valid input image by

$$x' = x + \epsilon \cdot \text{sign}(\nabla_x L(x, y)),$$

where L is the loss function, and ϵ controls the difference between the original input and the adversarial example. Kurakin et al. [13] improved this approach by iteratively manipulating the original image:

$$x^{(s)} = x^{(s-1)} + \frac{\epsilon}{S}\text{sign}(\nabla_x L(x^{(s-1)}, y)),$$

where S is the number of iterations, and s is the iteration index.

These algorithms find a nearby image, potentially imperceptible in the input space, that results in a wrong decision by the classifier. Intuitively, such an adversarial example lies in a subset of the input space that deviates from the (smooth) manifold defined by a set of training examples, in which the behavior of the classifier trained solely on these training examples is not well defined and potentially arbitrary.

One can view the proposed approach as a way to approximately project an input example onto the manifold defined by the training examples. The retrieval stage defines a local region of the highly nonlinear manifold near the original image, and the attention stage projects the original image onto this manifold by replacing it with a convex sum of retrieved examples. As long as the subsequent classifier well-behaves on any point between training examples,[2] that is, any convex sum of nearby training examples, these two first stages of the proposed framework effectively mitigates the issue of adversarial examples. In other words, the proposed approach proactively addresses adversarial examples. This is unlike some of the existing reactive solutions that either relies on retraining (or jointly training) with (potential) adversarial examples [7] or constraining the network configuration (parameters) [5].

4 PRELIMINARY EXPERIMENTS

Setup. We evaluate the proposed approach on SVHN [16] with 73,257 training (10,000 examples held-out for validation), 26,032 test and 100,000 extra examples. This set of 100k extra examples are assumed to be hidden (private) from the attacker and used by the retrieval engine. We train a conventional convolutional network as a baseline and as a feature extractor for the proposed framework. We test the proposed classifier with $N = 5$ and 10 to see the effect of the size of the retrieved set, which provides us the insight on the trade-off between the computational overhead and accuracy.

[2] This is a strong assumption that our empirical evaluation suggests otherwise.

Table 1: Results on SVHN.

ϵ	Baseline	Retrieval-based		Baseline	Retrieval-based	
		$N = 5$	$N = 10$		$N = 5$	$N = 10$
Clean						
0	**95.27** %	90.48 %	91.91 %			
FGSM				iFGSM		
0.01	75.05 %	80.53 %	**82.58** %	69.42 %	77.25 %	**80.30** %
0.02	54.55 %	72.17 %	**71.83** %	37.13 %	63.46 %	**64.97** %
0.04	31.92 %	**57.33** %	53.98 %	8.30 %	**44.28** %	40.37 %
0.06	22.22 %	**45.33** %	40.81 %	1.77 %	**33.08** %	26.34 %
0.08	17.24 %	**35.99** %	31.70 %	0.35 %	**26.41** %	18.40 %

Result and Analysis. We summarize the results in Table 1. From this table, we make two observations. First, on clean, non-adversarial images, the baseline–conventional convolutional network– clearly outperforms the proposed approach. This was expected, as projection by the attention mechanism inevitably results in information loss. It is however encouraging that this loss in accuracy decreases as we increase the number of retrieved examples, suggesting that with a more efficient retrieval engine, the proposed approach may largely recover the accuracy. Second, we find the proposed approach significantly more robust to adversarial examples, although we see that the proposed approach is not perfectly immune to adversarial examples. We find these observations make the proposed approach based on the off-the-shelf, non-differentiable retrieval engine a promising step toward building a robust object recognition system.

5 DISCUSSION

We proposed a retrieval-based approach to building a classifier robust to adversarial examples. How our approach mitigates the attack by adversarial examples could be understood from two perspectives. First, we reduce the attacker's access to the internal working of the network by introducing a non-differentiable retrieval engine and having a private set of retrieval database. Second, our approach efficiently project a new example onto the data manifold, avoiding those adversarial examples off the manifold.

Although promising, there are a number of issues that need to be addressed. First, our experiments have been limited to only two types of adversarial attacks and to a single dataset. Second, we currently rely on a feature vector extracted by another classifier, which may be a source of attack on its own. We plan to test an unsupervised feature extractor as an alternative. Last and perhaps most important, the proposed approach assumes the classifier works well on any convex sum of nearby training examples, which is not true (as our results demonstrate.) A learning algorithm that reflects this assumption must be investigated in the future.

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[2] Yoshua Bengio, Grégoire Mesnil, Yann Dauphin, and Salah Rifai. 2013. Better mixing via deep representations. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*. 552–560.

[3] Ella Bingham and Heikki Mannila. 2001. Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 245–250.

[4] John S Bridle. 1990. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In *Neurocomputing*. Springer, 227–236.

[5] Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. 2017. Parseval networks: Improving robustness to adversarial examples. In *International Conference on Machine Learning*. 854–863.

[6] Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab S Mirrokni. 2004. Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the twentieth annual symposium on Computational geometry*. ACM, 253–262.

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

[8] Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor OK Li. 2017. Search Engine Guided Non-Parametric Neural Machine Translation. *arXiv preprint arXiv:1705.07267* (2017).

[9] Kelvin Guu, Tatsunori B Hashimoto, Yonatan Oren, and Percy Liang. 2017. Generating Sentences by Editing Prototypes. *arXiv preprint arXiv:1709.08878* (2017).

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[11] Alex Krizhevsky and Geoffrey E Hinton. 2011. Using very deep autoencoders for content-based image retrieval.. In *ESANN*.

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[13] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. 2016. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533* (2016).

[14] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436–444.

[15] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2574–2582.

[16] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. 2011. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, Vol. 2011. 5.

[17] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).

[18] Zhiguo Wang, Wael Hamza, and Linfeng Song. 2017. k-Nearest Neighbor Augmented Neural Networks for Text Classification. *arXiv preprint arXiv:1708.07863* (2017).

[19] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. 2015. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579* (2015).