Combating Extreme Adversarial Vulnerabilities in WSI Classification

Gary Yao [1]

Abstract

The recent years' outstanding advancement in deep learning has created a desire to apply these algorithms ubiquitously, including partaking in clinical decisions. However, recent discoveries in adversarial vulnerabilities in neural network models are troubling. In this work, we examine adversarial attacks against a key AI pathology task, whole slide imaging classification. We demonstrate TopK Tiles Attack, a novel form of adversarial attack that leverages the attention mechanism in a modern deep learning WSI classification pipeline to cause massive classification error with extremely minor and unnoticeable changes to this input ($\leq$.2% pixels perturbed by 3 RGB values). We note the computational difficulties in adversarial training under weakly-supervised learning settings, proposing and experimenting workarounds. We also observe phenomena such as attention hijacking that call for more theoretical work to bring WSI classifiers to clinical settings under potential adversarial threats.

1. Introduction

Whole Slide Imaging (WSI) has grown into a very important tool for clinical diagnosis (Melo et al.). Under powerful digital microscopes, a slice of a tissue is digitized at a ultra-high resolution, resulting in whole slide images that capture detailed cell morphology of an entire slice of tissue. Pathologists utilize the fine details of whole slide imaging to perform diagnosis, teaching and telepathology (Pantanowitz et al., 2011). Those diagnostic works are often labor intensive and allude to certain specialized recognition of visual patterns. In this regard, the rapid advancements in deep learning models have reached human-level performance in many visual pattern recognition tasks (He et al.). Naturally, researchers have sought deep learning solutions to WSI-based diagnostic tasks.

Several deep learning pipelines have approach expert-level accuracy (Campanella et al.; Lu et al.). Paige.AI's model of WSI deep learning assistance has received FDA approval for clinical use (Ai & Gulturk). Without a doubt, deep learning holds great promises, but to be trusted in high-stakes situations requires demonstrated robustness against exploits such as adversarial attacks. In an adversarial attack, an instance of input is perturbed by a designed noise by a small amount, with the aim of shifting model's output by an unreasonably large degree (Goodfellow et al.). Adversarial attacks have been demonstrated to be a prevalent weakness among landmark deep learning models (Madry et al., 2017).

A number of works have demonstrated adversarial attacks on deep learning models for medical tasks (Hirano et al.). Past works have also analyzed economic interests in adversarial attacks upon medical AI (Finlayson et al.). Although, adversarial attack does not happen in natural datasets, widespread of AI pathology implies making expensive decisions, paving motivations for the 'hacking' of trusted AI systems. It is in the interest of medical professionals, patients as well as insurance providers to deploy deep learning models that are reliably robust against such adversarial attacks.

There are two main branches of methods to tiling-based WSI classifiers, weakly supervised learning (Lu et al.) and traditional multiple-instance learning (Kather et al.; Campanella et al.). Both branches are based on the paradigm of divide and conquer. First, a WSI is divided and filtered into smaller tiles containing tissue imaging of a standard dimensions. Then, from each tile is extracted a representation vector or tile-level score. This step is necessary because the ultra-high dimension of whole slide imaging renders performing an entire forwarded pass, while maintaining gradient information, computationally infeasible. Via some multiple instance learning method, this collection of extracted features or scores are combined into a slide-level representation or a slide-level label. While the exact implementation varies from model to model, we observed that a prevailing shift in method from a traditional multiple-instance learning implementation such as majority voting or one-cross-threshold classifiers (Kather et al.), to weakly supervised learning (Lu et al.; Shao et al., 2021) (Chen et al., 2022), where the model learns to identify the most important tiles, and

[1]Department of Artificial Intelligence Informatics, Mayo Clinic, Florida, USA. Correspondence to: Gary Yao <yao.gary@mayo.edu>.

The 2022 ICML Workshop on Computational Biology. Baltimore, Maryland, USA, 2022. Copyright 2022 by the author(s).

heavily leverage such tiles to make its decision.

In a recent study on adversarial attacks of MIL WSI classifiers (Laleh et al., 2022), a successful slide-level attack against traditional MIL models requires perturbing a significant proportion of tiles. We demonstrate in this study that weakly-supervised models with learned tile importance allow for a far more extreme form of adversarial vulnerability. We then explore and experiment various methods to defend against such a style of attack, including computational tricks to bypass computational difficulties of end-to-end adversarial training with WSI data. Along the way, we also present the phenomenon of attention hijacking and motivations for partial adversarial training. Together, we demonstrate that a complex multi-model pipeline can be adversarially attacked attacked and defended.

2. Methodology

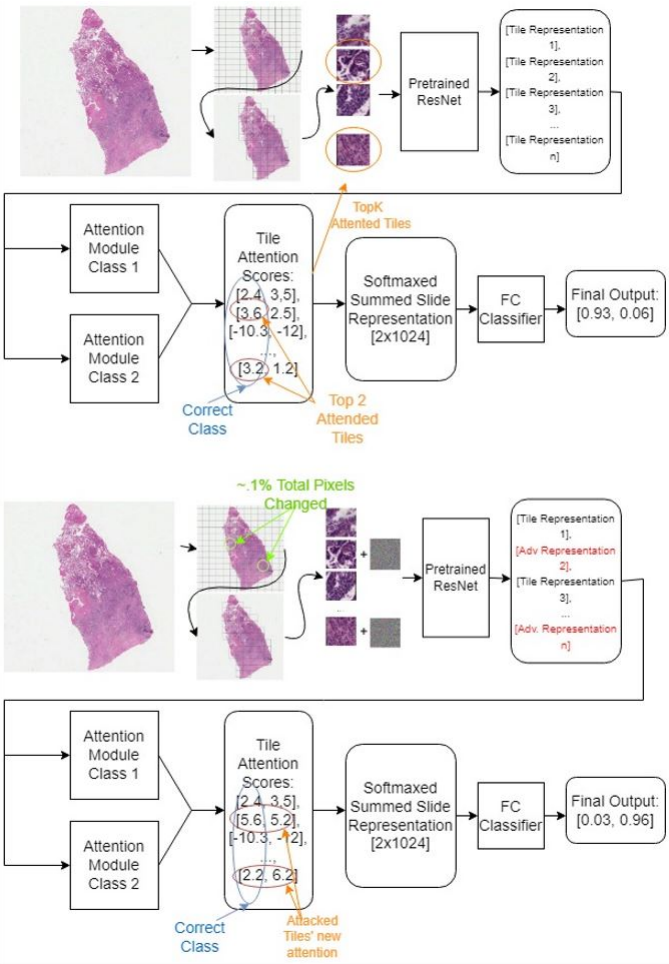

Figure 1. A: The original CLAM pipeline and the identification of top k attended tiles. B: TopK attack where top 2 tiles are replaced with adversarial tiles.

The primary attack target is the CLAM weakly supervised learning pipeline. This pipeline uses a pretrained ResNet50 to perform feature extraction on tiles containing foreground tissue. An attention module assigns an attention score to each tile-representation. Weighted by softmaxed attention scores, tile-level representations are combined into slide-level representation. One of the target tasks for this work is to perform lung cancer subtyping between slides of lung squamous cell carcinoma (LUSC) and adenocarcinoma (LUAD) using 1001 slides provided by The Cancer Genome Atlas Program (TCGA). Specifically, we utilized a CLAM author provided model for lung cancer subtyping as the main attack target. We also use an author provided train, validation, test split, such that all accuracy levels reported are of the combined validation and test splits.

Algorithm 1 TopK Tiles Attack

$y, a \leftarrow clam(ResNet(tiles))$
$idx \leftarrow topk(a[c], k)$
$\nabla tiles[idx] \leftarrow back\_propagation(y[target], tiles[idx])$
$t_{attcked} \leftarrow tiles[idx] + sign(\nabla tiles[idx]) * e$
return $t_{attcked}$

We propose Topk Tiles Attack (algo. 1). The general framework of this attack is to target the most influential k tiles according to some influential metric inherent to the target model. In CLAM, the influence metric is tile attention scores. Specifically, this attack finds the top-k attended tiles according to attentions scores of the target class. The framework can be specifically implemented with a wide choice of attacks on the by treating top-k tiles as variable inputs and non-topk tile inputs as constants. We utilized Fast Signed Gradient Attack in our implementation. This algorithm takes two parameters, e for attack strength and k for topk tile selection. We experimented this attack by validating against the adversarial TCGA lung subtyping dataset, using various combinations of e and k.

It is desirable to minimize changes in attention scores caused by adversarial attacks for several reasons. One reason is maintaining the high influence of the perturbed top k tiles on model decision. Another reason is to avoid the extreme changes in attention scores we observed in Topk Tiles Attack that can make an attack obvious to detection algorithms, to be discussed in a later section. We implemented Attention Compensation Attack, a secondary attack on the discrepancy between adversarial attention and natural attention.

$$\mathrm{t}_{ac} \leftarrow t_{attcked} + \mathrm{sign}(\nabla_{t_{attcked}} L_a(\theta, t_{attcked}, y)) * f$$

where $\mathrm{L}_a(\theta, tiles[idx], y) = (a_{attacked} - a)^2$

In implementation, the gradients upon attacked tiles were calculated via a forward pass of adversarial tiles (augmented with natural non-topk tile representation) and attention loss calculated against the natural accuracy. We experimented

with TopK Tiles attack with attention compensation where e = f and e = f/2.

We also experimented with adversarial training as a defense mechanism. Adversarial training is usually performed on a single model using adversarial samples as data augmentation. However, the CLAM pipeline is a combination of two deep learning models, ResNet50 and weakly supervised CLAM. Since both models are end-to-end differentiable, we treat these as a combined model = CLAM(ResNet([tiles])) that can be end-to-end trained via gradient descent. However, the reoccurring issue in performing WSI classification arise where the repeated use of feature extractor makes infeasible fitting this entire forward propagation on any standard GPU. We circumvent this challenge using the two augmented versions of adversarial training below.

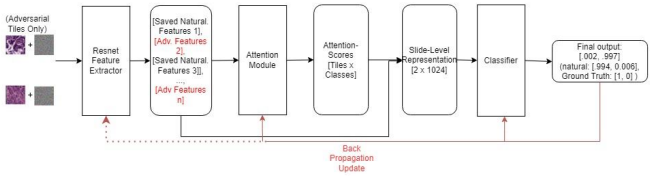

Figure 2. two modes of adversarial training

The first is a partial adversarial training, where only the later CLAM portion of the entire model is trained against a dataset of tile representations with TopK Attack adversarial augmentation. For each slide, we replace among the pre-processed tile features the TopK tiles' features with features extracted from adversarial versions of these tiles.

The second is a full pipeline adversarial training where we input only adversarial tiles to the ResNet feature extractor. For each adversarial tile representation vector, this vector replaces the representation of the natural tile. Under this training setup, both the feature extractor and the CLAM weakly supervised model are adversarial trained in an end-to-end fashion, while avoiding the computational difficulties arisen from repeated usage of the feature extractor.

3. Result and Discussion

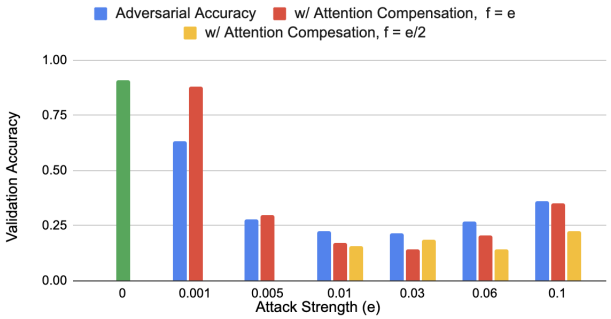

Figure 3. Accuracy of CLAM against adversarial input of various strengths, k = 10, lung cancer subtyping. Green bar is baseline

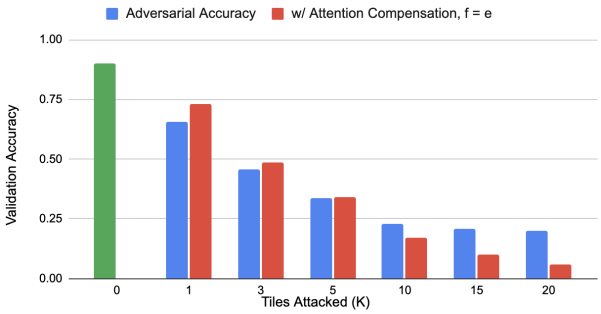

Figure 4. Accuracy of CLAM against adversarial input with various number of tiles replaced per slide, e, f = 0.01

We demonstrate that with minimal changes in input data, fewer than 20 tiles of on average 13255 tissue tiles per slide, perturbed by 3 units in RGB values, the model loses significant accuracy. One noteworthy observation is that increased attack strength eventually causes decreased attack effectiveness, because of loss in attention score, suggesting some innate robustness in weakly-supervised learning. As can be seen in Fig. 5 Fig. 6, When a single attacked tile is isolated, the attack is difficult to recognize by human eyes. Embedding such an attack among all features of a slide would be extremely laborious for pathologists to recognize, even at high attack strengths that would be easily recognized in other attack settings. The visual effect of adversarial attack in Fig. 5 is made easier to recognize than in practice, because the attack is centered and the crops are of high magnification. We must consider that experienced pathologists operate at low magnifications for a majority of work routines.

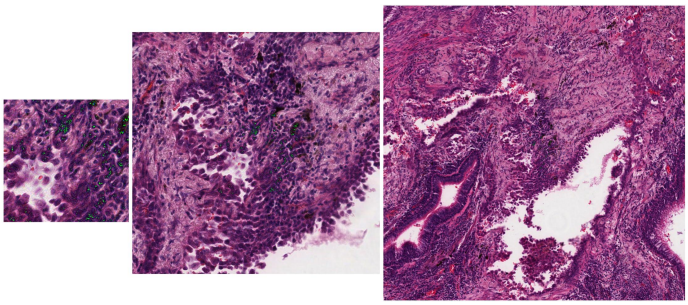

Figure 5. Example adversarial slide embedded among natural slides, e = 0.1 (25 RGB values)

We observed that in a majority of successful attacks, one or several attacked tiles exhibit extremely high attention scores (Fig. 7). We call this phenomenon attention hijacking. While this phenomenon can explain the very high effectiveness of Topk Tiles attack, it be a giveaway to adversarial detection algorithms, due to out-of-distribution behaviors. With a secondary attack in attention hijacking, however, those extreme changes in attention scores are much alleviated. In this, we demonstrated that adversarial attacks

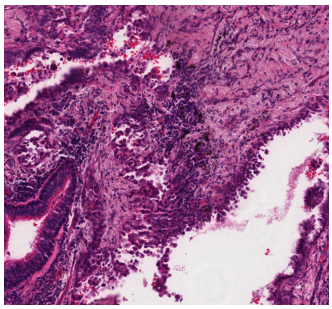

Figure 6. Adversarial tile is embedded among natural tiles, e = 0.05 (12 RGB values)

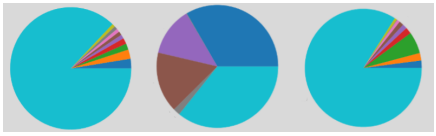

Figure 7. Changes in attention share of top 10 tiles of a sample slide under Topk Tiles Attack. Left: Natural. Middle: Adv. Right: Adv. with Attention Compensation.

can be effectively executed on key latent factors such as attention scores, and that multiple attacks can be executed in parallel to achieve multiple objectives. We demonstrate that Topk Attack with Attention Compensation eludes not only human-eye detection, but also interpretability tools popular to deep learning-based AI pathology.

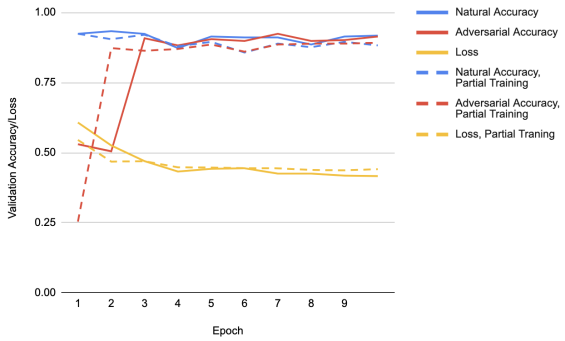

Figure 8. Adversarial Training Validation Stats with e = 0.01, k = 10

In adversarial training, both training modes succeeded in recovering adversarial accuracy to natural accuracy (fig. 8). After whole pipeline adversarial training, CLAM demonstrated to withstand adversarial attacks containing more adversarial tiles than it was exposed to during training (Fig.9). This result may pave way for general low-cost adversarial defense training for all weakly-supervised learning models that utilizes repeated feature extractor.

Despite good defensive results without performing adversarial training on the feature extractor, we find this practice unsubstantiated under theoretical frameworks. This issue

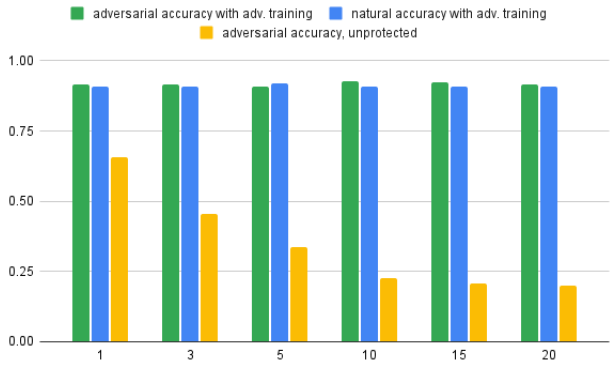

Figure 9. Whole Pipeline Adversarial Training with k = 5 against adversarial dataset with various k values

can be seen in the well-known saddle point adversarial attack formulation (Madry et al., 2017). $\min_\theta$, where $\mathrm{p}(\theta) = \mathbb{E}_{(x,y)\sim D}\left[\max_{\sigma\in S} L(\theta, x+\sigma, y)\right]$. Where L is the loss function according to a model with θ as parameters, x as input set, y as label set. The inner maximization refers to an attack's objective with σ picked from allowed set S.

Under this framework, a defensive method such as adversarial training is a heuristic minimizer whose empirical backings are observed against allowed sets S commonly defined by close samples around natural inputs, via L-norm bounds(Madry et al., 2017). Here, the difference between adversarial features extracted and natural features extracted: ResNet(natural tiles) - ResNet(adversarial tiles), is no longer bounded by the some reasonable L-norm bound (Fig 9). In undertaking partial training, the model may need to make associations between distant natural and adversarial sample pairs, an unexpected condition in current defensive schemes.

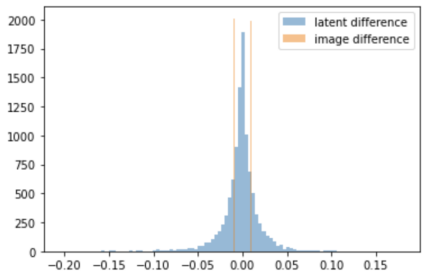

Figure 10. Shift in feature difference, latent vs input

To our knowledge, behaviors of partially robust-trained models have not been studied, as this concept is strange in single model setting, but is very important to tile-based WSI classifiers and other feature-extractor reliant pipelines for hyper-dimensional medical imaging. Here we leave several open problems, and point out significant motivations for future work into adversarial defense in multiple instance learning and hyper-dimensional data in guaranteeing safe and trustworthy clinical AI pathology.

References

Ai, P. and Gulturk, E. Re: DEN200080 trade/device name:. pp. 6.

Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K. J., Brogi, E., Reuter, V. E., Klimstra, D. S., and Fuchs, T. J. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. 25(8): 1301–1309. ISSN 1078-8956, 1546-170X. doi: 10.1038/s41591-019-0508-1. URL http://www.nature.com/articles/s41591-019-0508-1.

Chen, R. J., Chen, C., Li, Y., Chen, T. Y., Trister, A. D., Krishnan, R. G., and Mahmood, F. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16144–16155, June 2022.

Finlayson, S. G., Chung, H. W., Kohane, I. S., and Beam, A. L. Adversarial attacks against medical deep learning systems. URL https://arxiv.org/abs/1804.05296.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. URL http://arxiv.org/abs/1412.6572.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. URL http://arxiv.org/abs/1512.03385.

Hirano, H., Minagi, A., and Takemoto, K. Universal adversarial attacks on deep neural networks for medical image classification. 21(1):9. ISSN 1471-2342. doi: 10.1186/s12880-020-00530-y. URL https://bmcmedimaging.biomedcentral.com/articles/10.1186/s12880-020-00530-y.

Kather, J. N., Pearson, A. T., Halama, N., Jäger, D., Krause, J., Loosen, S. H., Marx, A., Boor, P., Tacke, F., Neumann, U. P., Grabsch, H. I., Yoshikawa, T., Brenner, H., Chang-Claude, J., Hoffmeister, M., Trautwein, C., and Luedde, T. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. 25(7):1054–1056. ISSN 1546-170X. doi: 10.1038/s41591-019-0462-y. URL https://doi.org/10.1038/s41591-019-0462-y.

Laleh, N. G., Truhn, D., Veldhuizen, G. P., Han, T., van Treeck, M., Buelow, R., Langer, R., Dislich, B., Boor, P., Schulz, V., and Kather, J. N. Adversarial attacks and adversarial robustness in computational pathology, 2022. URL https://doi.org/10.1101/2022.03.15.484515.

Lu, M. Y., Williamson, D. F. K., Chen, T. Y., Chen, R. J., Barbieri, M., and Mahmood, F. Data-efficient and weakly supervised computational pathology on whole-slide images. 5(6):555–570. ISSN 2157-846X. doi: 10.1038/s41551-020-00682-w. URL http://www.nature.com/articles/s41551-020-00682-w.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks, 2017. URL https://arxiv.org/abs/1706.06083.

Melo, R. C. N., Raas, M. W. D., Palazzi, C., Neves, V. H., Malta, K. K., and Silva, T. P. Whole slide imaging and its applications to histopathological studies of liver disorders. 6. ISSN 2296-858X. URL https://www.frontiersin.org/article/10.3389/fmed.2019.00310.

Pantanowitz, L., Valenstein, P. N., Evans, A. J., Kaplan, K. J., Pfeifer, J. D., Wilbur, D. C., Collins, L. C., and Colgan, T. J. Review of the current state of whole slide imaging in pathology. *Journal of Pathology Informatics*, 2(1):36, 2011. ISSN 2153-3539. doi: https://doi.org/10.4103/2153-3539.83746. URL https://www.sciencedirect.com/science/article/pii/S2153353922002127.

Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., and Zhang, Y. Transmil: Transformer based correlated multiple instance learning for whole slide image classication. *CoRR*, abs/2106.00908, 2021. URL https://arxiv.org/abs/2106.00908.

A. Code and Data Availability

Code for all experiment and analysis have been made available at: https://github.com/garyyjn/topkattack

Attacked model parameters, adversarial trained models, adversarial dataset, and CLAM extracted features are available at: https://drive.google.com/drive/folders/1OV5yQF9O0kK8NdXbygeSxtcMY9OEyx_C?usp=sharing

Diagnostic Slides can be accessed via https://portal.gdc.cancer.gov/

B. Acknowledgement

I would like to express my gratitude to members of Hwang Lab for their guidance, and the Mahmood Lab for their great help in reproducing CLAM.