A Work in Progress: Tighter Bounds on the Information Bottleneck for Deep Learning

Nir Z. Weingarten
Efi Arzi school of CS
Reichman University
nirzvi89@gmail.com

Moshe Butman
Efi Arzi school of CS
Reichman University
butman.moshe@post.runi.ac.il

Ran Gilad-Bachrach
BioMed Engineering Dept.
Tel-Aviv University
rgb@tau.ac.il

1 Introduction

The field of Deep Neural Nets (DNNs) is evolving rapidly with new architectures emerging to better extract information from available data. The Information Bottleneck, IB, (Tishby, Pereira, and Bialek 1999) offers an optimal information theoretic framework for data modeling. However, IB is intractable in most settings. In recent years attempts were made to combine deep learning with IB both for optimization and to explain the inner workings of deep neural nets. VAE (Kingma and Welling 2014) inspired variational approximations such as VIB (Alemi et al. 2017) became a popular method to approximate bounds on the required mutual information computations. This work continues this direction by introducing a new tractable variational upper bound on the IB functional which is empirically tighter than previous bounds. When used as an objective function it enhances the performance of previous IB-inspired DNNs in terms of test accuracy and robustness to adversarial attacks across several challenging tasks.

2 Related work

The Information Bottleneck method (Tishby, Pereira, and Bialek 1999) extends rate-distortion (Blahut 1972) by optimizing a signal's compression to a given downstream task. IB works by minimizing the functional $L_{P(Z|X)} = I(Z;X) - \beta I(Z;Y)$ for a chosen β. The IB functional is only tractable when mutual information can be computed. Tishby and Zaslavsky (2015) proposed an IB interpretation of DNNs regarding them as Markov cascades of intermediate representations between hidden layers. Shwartz-Ziv and Tishby (2017) visualized and analyzed the rate-distortion behavior of DNNs over a toy problem with a known distribution. Alemi et al. (2017) introduced the Variational Information Bottleneck (VIB) as a tractable variational approximation for an upper bound on the IB objective for classifier DNN optimization. To derive VIB an upper bound on the IB objective is first suggested using the identity $\forall P, Q : D_{KL}(P||Q) \geq 0$. Next, intractable distributions are replaced with variational approximations. Finally, a discretized empirical estimation of the approximation to the upper bound is devised using stochastic DNNs. This results in a loss function that is identical to VAE loss consisting of cross entropy and KL regularization. When applied to difficult tasks VIB was shown to cause a slight reduction in test-accuracy while generating substantial improvements in robustness to adversarial attacks.

37th Information-Theoretic Principles in Cognitive Systems Workshop at the 37th Conference on Neural Information Processing Systems (NeurIPS 2023).

3 From VIB to VUB

IB upper bound - Let $\mathcal{D}$ be our training data, $P^*(Y)$ be the true discrete categorical distribution for $y \in \mathcal{D}$, $C(Y)$ be the modeled distribution and CH be discrete cross entropy. From the Gibbs inequality we have that $CH\left(P^*(Y), C(Y)\right) \geq H(Y)$ and as for any discrete distribution $H(Y) \geq H(Y|Z) \geq 0$ we have that:

$$CH\left(P^*(Y), C(Y)\right) - H(Y|Z) \geq 0 \tag{1.0}$$

We use this inequality to derive a new upper bound for the IB objective L_{IB}:

$$\begin{aligned}
L_{IB} =& I\left(Z;X\right) - I\left(Z;Y\right) && (2.0)\\
\overset{(1.0)}{\leq}& I\left(Z;X\right) + CH\left(P^*(Y), C(Y)\right) - H(Y|Z) && (2.1)\\
=& H(Z) - H(Z|X) + CH\left(P^*(Y), C(Y)\right) - H(Y|Z) && (2.2)
\end{aligned}$$

We proceed to develop an explicit bound L_{UB}:

$$\begin{aligned}
&L_{UB} \equiv && (2.3)\\
&-\int p^*(z) log\left(p^*(z)\right) dz\\
&+\int\int p^*(z,x) log\left(p^*(z|x)\right) dxdz\\
&-\sum_{x,y\in\mathcal{D}} P^*(Y=y|x) log\left(C(Y=y)\right)\\
&+\sum_{y\in\mathcal{D}}\int P^*(Y=y|z)p^*(z) log\left(P^*(Y=y|z)p^*(z)\right) dz
\end{aligned}$$

Variational approximation - We replace intractable distributions with variational approximations. Let $p^*(x,y,z)$ be the unknown joint distribution, $e(z|x)$ a variational encoder approximating $p^*(z|x)$, $c(y|z)$ a variational classifier approximating $p^*(y|z)$. Since $p^*(z)$ is intractable we use a variational approximation $r(z)$ as done in VIB and VAE. Using total probability and the Markov chain $Z \leftarrow X \leftarrow Y$ to derive $p^*(y,z)$ we define L_{VUB} as an approximation to L_{UB} as follows:

$$\begin{aligned}
&L_{VUB} \equiv && (3.0)\\
&-\int r(z) log\left(r(z)\right) dz && \text{(i)}\\
&+\int\int e(z|x)p^*(x) log\left(e(z|x)\right) dzdx && \text{(ii)}\\
&-\sum_{x,y\in\mathcal{D}} P^*(Y=y|x) log\left(C(Y=y)\right) && \text{(iii)}\\
&+\sum_{x,y\in\mathcal{D}}\int C(Y=y|z)e(z|x)p^*(x) log\left(C(Y=y|z)\right) dz && \text{(iv)}
\end{aligned}$$

Empirical estimation - We estimate the approximation to the upper bound using stochastic DNNs. Let e_ϕ be a stochastic DNN encoder with parameters ϕ applying the reparameterization trick (Kingma and Welling 2014) such that $e_\phi(x) \sim N(\mu, \Sigma)$. Let C_λ be a discrete classifier DNN parameterized by λ such that $C_\lambda(\hat{z}) \sim Multinomial$. As in VIB and VAE we chose a standard gaussian for $r(z)$, hence $r(z)$ is unparameterized and can be ignored in the optimization process. We can empirically estimate L_{VUB} over the training data $\mathcal{D}$ using Monte Carlo sampling over the following functional:

$$\hat{L}_{VUB} \equiv \frac{1}{N} \sum_{n=1}^{N} \left[-H\left(e_\phi(x_n)\right) - P^*(y_n) \cdot log\left(C_\lambda\left(e_\phi(x_n)\right)\right) - \beta \cdot H\left(C_\lambda\left(e_\phi(x_n)\right)\right) \right] \tag{4.0}$$

Notice that VUB is in fact a CH term regulated by two conditional entropy terms: $-H(Z|X)$ and $-H(Y|Z)$ similarly to Pereyra et al. (2017).

Comparison to VIB - VIB is empirically estimated as a CH term and a positive KL regularization term. VUB is estimated as a CH term and two negative conditional entropy terms. Hence VUB is a tighter empirical estimation of the IB functional than VIB. A deeper analysis of these differences is given in section 5.

4 Experiments

We follow the experimental setup used by Alemi et al. (2017) over ImageNet (Deng et al. 2009), which we have expanded to also accommodate NLP tasks. Due to space constraints, additional technical details, results and the specific NLP tasks are provided in appendix A. Models were evaluated using test set accuracy, estimated rate and distortion, and robustness to the Fast Gradient Sign (FGS) (Goodfellow, Shlens, and Szegedy 2015) and targeted L_2 (Carlini and Wagner 2017), (Kaiwen 2018) adversarial attacks. In addition we tracked IB metrics to provide quantitive assessment of the learning process. For image classification a pre-trained inceptionV3 (Szegedy et al. 2016) base model was used and achieved a 77.21% accuracy on the image-net 2012 validation set. Note that inceptionV3 yields a slightly worse single shot accuracy than inceptionV2 (80.4%) when run in a single model and single crop setting, however we've used InceptionV3 over V2 for simplicity. As shown in Table 1 image classification evaluation results confirm that while VIB reduces performance on the validation set, it significantly improves robustness to adversarial attacks. Moreover, these results demonstrate that VUB significantly outperforms VIB in terms of validation accuracy, resilience to untargeted attacks, and estimated rate-distortion. However, VIB outperforms VUB in the context of targeted attacks. A comparison of the best VIB and VUB models further substantiates these findings, with statistical significance confirmed by a p-value of less than 0.05 in a Wilcoxon rank sum test. VUB allows us to easily estimate rate and distortion in training and plot an estimated information curve for real world problems. Notably, DNNs trained with VUB reproduced the same information plane behavior previously demonstrated on toy problems by Shwartz-Ziv and Tishby (2017) for both error minimization and representation compression phases, as shown in Figure 1 in Appendix A. Code is available on `https://github.com/hopl1t/vub.git`.

5 Discussion

The case for the IB functional, and it's variational approximations, as a theoretical limit for optimal representation relies on three assumptions: (1) It suffices to optimize the mutual information metric to optimize a model's performance; (2) Forgetting more information about the input while keeping

β	**Val** ↑	**FGS** ↓ $\epsilon=0.1$	**FGS** ↓ $\epsilon=0.5$	**CW**↑	**R**↑	**D**↑
Vanilla model						
-	**77.2%**	**68.9%**	**67.7%**	**788**	-	-
VIB models						
10^{-3}	**72.0%** ±.2%	**62.6%** ±.4%	**66.0%** ±.2%	**3204** ±485	**45** ±.2	**7.99** ±.02
10^{-2}	**71.5%** ±.1%	**60.6%** ±.3%	**65.8%** ±.1%	**1761** ±235	**13** ±.04	**4.33** ±.02
10^{-1}	**71.4%** ±.01%	**72.3%** ±.6%	**71.3%** ±.2%	**1602** ±366	**6.3** ±.004	**1.56** ±.004
VUB models						
10^{-3}	**73.8%** ±.08%	**45.1%** ±.2%	**61.7%** ±.1%	**788** ±.32	**594** ±.003	**16.68** ±.07
10^{-2}	**73.9%** ±.1%	**44.8%** ±.07%	**61.2%** ±.07%	**790** ±2.1	**594** ±.002	**16.47** ±.12
10^{-1}	**73.7%** ±0%	**45.6%** ±0%	**61.9%** ±0%	**790** ±4.6	**594** ±0	**16.77** ±.11

Table 1: Image-net evaluation scores for vanilla, VIB and VUB models, average over 3 runs with standard deviation. First column is performance on the image-net validation set. Second and third columns are the % of successful FGS attacks at $\epsilon = 0.1, 0.5$ (lower is better ↓). Fourth column is the average L_2 distance for a successful Carlini Wagner L_2 targeted attack (higher is better ↑), forth and fifth columns are final estimated rate (higher is better ↑) and distortion (higher is better ↑).

the same information about the output induces better generalization over unseen data; (3) Mutual information between the input, output and latent representation can be either computed or approximated to a desired level of accuracy. Our study strengthens the argument for using the Information Bottleneck combined with variational approximations to obtain robust models that can withstand adversarial attacks. The VUB objective penalizes high classifier entropy by the conditional entropy term $H(Y|Z)$ which follows from it's derivation. The same term was shown by Pereyra et al. (2017) to increase classification performance, arguably increasing classifier calibration, and might account for VUB's superior test set accuracy and resilience to untargeted attacks. VUB's inferior performance on targeted CW attacks might be explained by the limitations of our variational approximations, by the limitations of a mutual information objective in deep learning as suggested by Amjad and Geiger (2020) or by the difference between encoder regularizations used. Both methods promote a disentangled and smoother latent space by using a stochastic factorized prior (Chen et al. 2018), arguably making it harder for perturbations to alter it's latent semantics. In contrast to conditional entropy, KL regularization also enforces clustering around a 0 mean which might increase smoothness. Differences in latent smoothness and entropy might account for differences in resilience to attacks. It is possible that a smoother latent space with higher entropy will be less susceptible to targeted attacks as there is a bigger difference between two distinct classes in the latent space. On the other hand a less smooth space with lower entropy might require a bigger change in the latent space to transition to adjacent classes, making it more resilient to untargeted attacks. In light of these findings, we suggest that practitioners monitor the rate distortion ratio and validation set accuracy during training. These metrics may be more informative indicators of model performance, as validation set cross entropy could increase as the model becomes more calibrated.

Acknowledgements. No external funding received in the last 36 months for this paper's work.

References

Alemi, A. A.; Fischer, I.; Dillon, J. V.; and Murphy, K. 2017. Deep Variational Information Bottleneck. In *Proceedings of the International Conference on Learning Representations (ICLR)*. Google Research.

Amjad, R. A.; and Geiger, B. C. 2020. Learning Representations for Neural Network-Based Classification Using the Information Bottleneck Principle. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(9): 2225–2239.

Blahut, R. E. 1972. Computation of channel capacity and rate distortion function. *IEEE Transactions on Information Theory*, IT-18: 460–473.

Carlini, N.; and Wagner, D. A. 2017. Towards Evaluating the Robustness of Neural Networks. In *IEEE Symposium on Security and Privacy*, 39–57. IEEE Computer Society.

Chen, T. Q.; Li, X.; Grosse, R. B.; and Duvenaud, D. 2018. Isolating Sources of Disentanglement in Variational Autoencoders. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2615–2625.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. In *ICLR (Poster)*.

Kaiwen. 2018. pytorch-cw2. GitHub repository.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Pereyra, G.; Tucker, G.; Chorowski, J.; Kaiser, L.; and Hinton, G. E. 2017. Regularizing Neural Networks by Penalizing Confident Output Distributions. In *Proceedings of the International Conference on Learning Representations*. OpenReview.net.

Shwartz-Ziv, R.; and Tishby, N. 2017. Opening the Black Box of Deep Neural Networks via Information. 19 pages, 8 figures, arXiv:arXiv:1703.00810.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Tishby, N.; Pereira, F. C.; and Bialek, W. 1999. The Information Bottleneck Method. In *The 37th annual Allerton Conference on Communication, Control, and Computing.* Hebrew University, Jerusalem 91904, Israel.

Tishby, N.; and Zaslavsky, N. 2015. Deep Learning and the Information Bottleneck Principle. arXiv:1503.02406.

A Elaborated experiments

We follow the experimental setup used by Alemi et al. (2017). In addition we tracked IB metrics to provide quantitive assessment of the learning process. Image classification models were trained on the first 100,000 samples of the ImageNet 2012 dataset (Deng et al. 2009) and text classification over IMDB sentiment analysis dataset (Maas et al. 2011). For each dataset, a competitive pre-trained model (Vanilla model) was evaluated and then used to encode embeddings. These embeddings were then used as a dataset for a new stochastic classifier net with either a VIB or a VUB loss function. Stochastic classifiers consisted of two ReLU activated linear layers of the same dimensions as the pre-trained model's logits (2048 for image and 768 for text classification), followed by reparameterization and a final softmax activated FC layer. Learning rate was 10^{-4} and regularization terms were clipped not to surpass the CH term to stabilize the learning process. Batch sizes were 32 for ImageNet and 16 for IMDB. We used a single forward pass per sample for inference. Each model was trained and evaluated 3 times per β value with consistent performance and we display the mean and standard deviation per β setting. We tried $\beta = 10^{-i}$ for $i \in \{0, 1, 2, 3\}$ since previous studies indicated this is the best range for VIB (Alemi et al. 2017, 2018). We've also tried a few $\beta > 1$, $\beta < 10^{-3}$ values to verify that indeed these regions produce non-competitive results. Each model was evaluated using test set accuracy, estimated rate and distortion, and robustness to adversarial attacks over the test set. We utilized several adversarial attacks in our study. For image classification, we employed untargeted Fast Gradient Sign (FGS) attacks (Goodfellow, Shlens, and Szegedy 2015) as well as targeted L_2 optimization attacks (Carlini and Wagner 2017), (Kaiwen 2018). For text classification, we used the Deep Word Bug method (Gao et al. 2018), (Morris et al. 2020). Code to reconstruct the experiments is provided in the code & data appendix. All models were trained using an Nvidia RTX3080 GPU.

A.1 Image classification

A pre-trained inceptionV3 (Szegedy et al. 2016) base model was used and achieved a 77.21% accuracy on the image-net 2012 validation set (Test set for image-net is unavailable). Note that inceptionV3 yields a slightly worse single shot accuracy than inceptionV2 (80.4%) when run in a single model and single crop setting, however we've used InceptionV3 over V2 for simplicity. Each model was trained for 100 epochs.

A.1.1 Evaluation and analysis

Image classification evaluation results are shown in Table 1, examples of successful attacks are shown in Figures 2, 3. The empirical results presented in Table 1 confirm that while VIB reduces performance on the validation set, it significantly improves robustness to adversarial attacks. Moreover, these results demonstrate that VUB significantly outperforms VIB in terms of validation accuracy, resilience to untargeted attacks, and estimated rate-distortion. However, VIB outperforms VUB in the context of targeted attacks. A comparison of the best VIB and VUB models further substantiates these findings, with statistical significance confirmed by a p-value of less than 0.05 in a Wilcoxon rank sum test.

VUB allows us to easily estimate rate and distortion in training and plot an estimated information curve for real world problems. DNNs trained with VUB reproduced the same information plane behavior previously demonstrated on toy problems by Shwartz-Ziv and Tishby (2017) for both error minimization and representation compression phases, as shown in Figure 1.

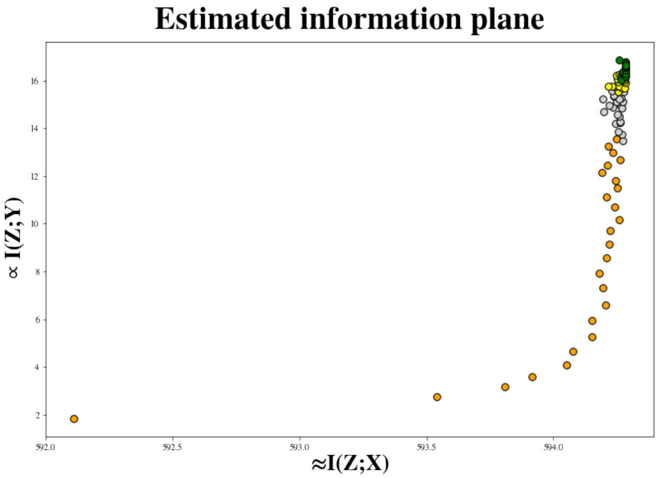

Figure 1: Estimated information plane metrics per epoch for VUB with $\beta = 0.01$. $I(Z;X)$ is approximated by $H(R) - H(Z|X)$ and $\frac{1}{CH(Y;\hat{Y})}$ is used as an analog for $I(Z;Y)$. The epochs have been grouped and color-coded in intervals of 25 epochs in the order: Orange (0-25), gray (26-50), red (51-75) and green (76-100).

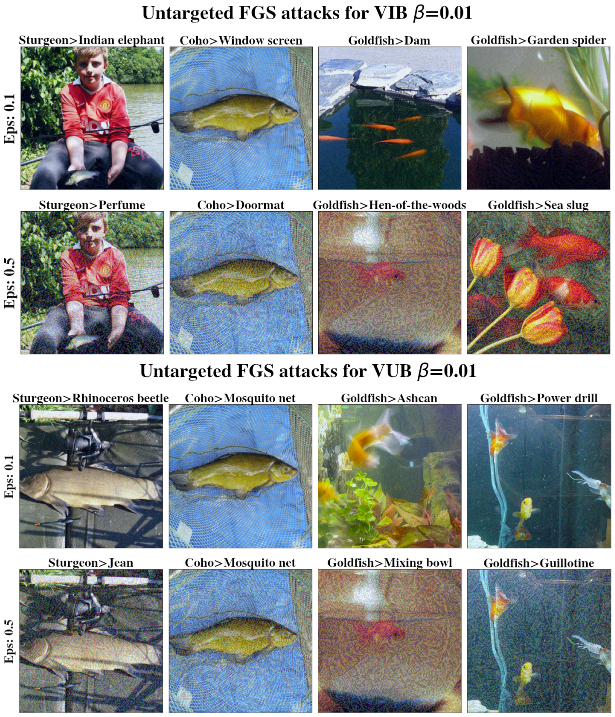

Figure 2: Successful untargeted FGS attack examples. Images are perturbations of previously successfully classified instances from the ImageNet validation set. Perturbation magnitude is determined by the parameter ϵ shown on the left, the higher the more perturbed. Notice the deterioration of image quality as ϵ increases. Original and wrongly assigned labels are listed at the top of each image.

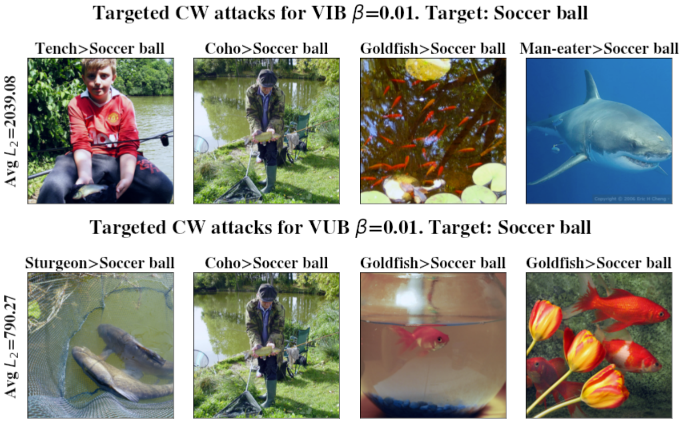

Figure 3: Successful targeted CW attack examples. Images are perturbations of previously successfully classified instances from the ImageNet validation set. The target label is 'Soccer ball'. Average L_2 distance required for a successful attack is shown on the left. The higher the required L_2 distance the greater the visible change required to fool the model. Original and wrongly assigned labels are listed at the top of each image. Mind the difference in noticeable change as compared to FGS perturbations and between VIB and VUB perturbations.

A.2 Text classification

A fine tuned BERT uncased (Devlin et al. 2019) base model was used and achieved a 95.5% accuracy on the IMDB sentiment analysis test set. Each model was trained for 150 epochs and the first 200 entries in the test set used for evaluation and adversarial attacks.

A.2.1 Evaluation and analysis

Text classification evaluation results are shown in Table 2, examples of successful attacks are shown in Figure 3. The best VUB and VIB models reach similar results in terms of test set evaluation and resilience to the Deep Word Bug attack. Interestingly, VUB achieves better rate distortion ratio without a noticeable improvement in performance. This is further discussed in Section 5.

β	**Test↑**	**DWB↓**	**R↑**	**D↑**
		Vanilla model		
-	**95.5%**	**75.9%**	-	-
		VIB models		
10^{-3}	**95.0%** ±.4%	**33.6%** ±3.8%	**4.17** ±.07	**50.35** ±1.78
10^{-2}	**95.0%** ±1%	**35.3%** ±3.2%	**1.23** ±.1	**35.16** ±.68
10^{-1}	**94.2%** ±.2%	**87.9%** ±4.2	**0.58** ±.005	**21.16** ±1.56
		VUB models		
10^{-3}	**95.0%** ±.5%	**32.7%** ±3.2%	**220.4** ±1.3	**64.28** ±10.7
10^{-2}	**94.0%** ±.2%	**40.7%** ±2.7%	**220.2** ±.45	**71.04** ±5.88
10^{-1}	**94.83%** ±0.6%	**46.2%** ±5.9%	**220.8** ±0.2	**38.51** ±4.96

Table 2: Evaluation for vanilla, VIB and VUB models, average over 3 runs with standard deviation over the IMDB dataset. First column is performance on the test set, second is % of successful Deep Word Bug attacks (lower is better ↓), third and fourth columns are final rate (higher is better ↑) and distortion (lower is better ↑).

Original text
great historical movie, will not allow a viewer to leave once you begin to watch. View is presented differently than displayed by most school books on this *subject*. My only fault for this movie is it was photographed in black and white; wished it had been in color ... wow !
Perturbed text
gnreat historical movie, will not allow a viewer to leave once you begin to watch. View is presented differently than displayed by most school books on this *sSbject*. My only fault for this movie is it was photographed in black and white; wished it had been in color ... wow !

Table 3: Example of a successful Deep Word Bug attack on a vanilla Bert model fine tuned over the IMDB dataset. The original label is ’Positive sentiment’. Perturbations, marked in italic font, change the classification to ’Negative sentiment’. VUB and VIB classifiers are far less susceptible to these perturbations as shown in Table 2.

References (Appendix)

Alemi, A. A.; Fischer, I.; Dillon, J. V.; and Murphy, K. 2017. Deep Variational Information Bottleneck. In *Proceedings of the International Conference on Learning Representations (ICLR)*. Google Research.

Alemi, A. A.; Poole, B.; Fischer, I.; Dillon, J. V.; Saurous, R. A.; and Murphy, K. 2018. Fixing a Broken ELBO. In *Proceedings of Machine Learning Research*, volume 80, 159–168. PMLR.

Carlini, N.; and Wagner, D. A. 2017. Towards Evaluating the Robustness of Neural Networks. In *IEEE Symposium on Security and Privacy*, 39–57. IEEE Computer Society.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota.

Gao, J.; Lanchantin, J.; Soffa, M. L.; and Qi, Y. 2018. Black-box generation of adversarial text sequences to evade deep learning classifiers. In *2018 IEEE Security and Privacy Workshops (SPW)*, 50–56. IEEE.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. In *ICLR (Poster)*.

Kaiwen. 2018. pytorch-cw2. GitHub repository.

Maas, A.; Daly, R. E.; Pham, P. T.; Huang, D.; Ng, A. Y.; and Potts, C. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, 142–150.

Morris, J.; Lifland, E.; Yoo, J. Y.; Grigsby, J.; Jin, D.; and Qi, Y. 2020. TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 119–126.

Shwartz-Ziv, R.; and Tishby, N. 2017. Opening the Black Box of Deep Neural Networks via Information. 19 pages, 8 figures, arXiv:arXiv:1703.00810.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.