# Supplementary material for Defensive Tensorization

Adrian Bulat*,[1]
adrian@adrianbulat.com
Jean Kossaifi*,[2]
jean.kossaifi@gmail.com
Sourav Bhattacharya[1]
sourav.b1@samsung.com
Yannis Panagakis[3]
yannisp@di.uoa.gr
Timothy Hospedales[1,4]
t.hospedales@ed.ac.uk
Georgios Tzimiropoulos[1,5]
g.tzimiropoulos@qmul.ac.uk
Nicholas D Lane[1,6]
nic.lane@samsung.com
Maja Pantic[7]
m.pantic@imperial.ac.uk

[1] Samsung AI Centre
Cambridge, UK
[2] NVIDIA
New York, USA
[3] University of Athens
Athens, Greece
[4] University of Edinburgh
Edinburgh, UK
[5] Queen Mary University London
London, UK
[6] University of Cambridge
Cambridge, UK
[7] Imperial College London
London, UK

## A Effect on the optimization landscape

To visually assess the impact of both our randomization scheme on the optimization landscape, we visualise the evaluation of the loss in a fixed neighbourhood around an unseen data point. Specifically, we visualize the loss function learned by the model in the neighbourhood of new, unseen data point $x$. For clarity, we visualise the loss in a 2–dimensional space, along direction of the gradient at $x$ (x-axis) and a randomly chosen direction, orthogonal to the direction of gradient (y-axis). Next, a mesh-grid is constructed by sampling uniformly points along these two directions for the range $[-0.5, 0.5]$. Then a contour plot is constructed by evaluating the losses for all points on the mesh-grid. The result for our best model can be seen in figure 1.

Intuitively, the randomization (which is done in the *latent* subspace of the decomposition, not on the weights themselves), changes the loss function, at each pass, making it hard to converge to a fixed attack due to the presence of many spurious minimums. This can be seen by looking at the landscape of the loss function around an arbitrary sample in Fig. 1. The landscape is inline with the finding of Madry et al. [3], where the authors show the adversarial

* - denotes equal contribution

training smooths the loss space around 0. This is even more noticeable for the method that combines our approach with adversarial training (see Fig. 1).

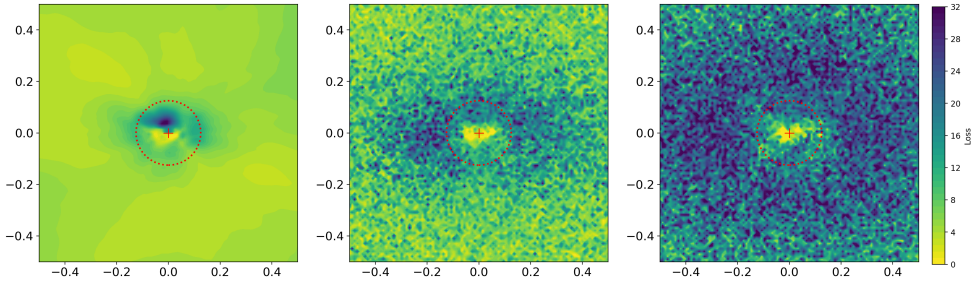


Figure 1: **Contour plot of the loss surface** of our model, adversarially trained model for various values of $\theta$ evaluated on the $l_\infty$ neighbourhood of an unseen CIFAR-10 image. The same direction was used for all three plots. The red circle denotes the $\varepsilon = 32$ neighbourhood.

# B Defensive tensorization for audio classification

To further demonstrate the generalizability of our approach, this section considers adversarial attacks on the audio domain, measuring the efficacy of our method.

## B.1 Experimental setting and implementation details

**Speech Command:** Speech Command [2] is an audio recognition dataset comprised of 105,000 1-second utterances of words from a large number of users spanning over a small vocabulary. The objective is to recognize among ten spoken words: *yes, no, up, down, left, right, on, off, stop, go*, in addition to recognizing words outside the vocabulary as *unknown*, and detecting *silence*. The dataset is balanced and all audio recordings are captured with a sampling frequency of 16 KHz. We use a 80%-10%-10% splits for training, validation and testing respectively.

**Implementation details:** For the experiments conducted on the Speech Command dataset we build on the SoundNet5 [1] architecture containing 5 convolutional layers *[in_channels, out_channels, kernel, stride, padding]*: $[1, 16, (1\times 64), (1\times 2), (0\times 32)]$, $[16, 32, (1\times 32), (1\times 2), (0\times 16)]$, $[32, 64, (1\times 16), (1\times 2), (0\times 8)]$, $[64, 128, (1\times 8), (1\times 2), (0\times 4)]$, $[128, 256, (1\times 4), (1\times 2), (0\times 2)]$ and 2 linear ones: $[512, 256]$ and $[256, 12]$. Each convolutional layer was followed by a max-pooling operation. We trained all of the audio models using Adam [2] for 50 epochs with an initial learning rate set to 0.01 that was dropped by $0.1\times$ at epoch 25 and 35.

**Attacking the model:** The attack model largely follows the procedure used for images, with a small adaptation: On the Speech Command dataset, since the raw data is in the range $[-1, 1]$, we scaled the value of $\varepsilon$ accordingly, running it for the following values $\varepsilon = \{0.008, 0.032, 0.063\}$.

Table 1: **Performance on Speech Command** for FGSM attacks using both binary and real-valued models. Notice that our approach is significantly more robust.

| Quant. | $\varepsilon$ | Baseline | Defensive Tensorization | | |
|---|---|---|---|---|---|
| | | | $\theta = 0.99$ | $\theta = 0.95$ | $\theta = 0.9$ |
| Real | **No attack** | **93.8** | 92.0 | 89.6 | 88.1 |
| | **0.008** | 33.0 | 49.6 | 58.2 | **61.0** |
| | **0.032** | 14.9 | 33.0 | 40.2 | **44.2** |
| | **0.063** | 7.6 | 23.8 | 31.8 | **35.7** |
| Binary | **No attack** | 88.0 | **89.0** | 83.5 | 83.2 |
| | **0.008** | 12.2 | 50.1 | 54.4 | **56.0** |
| | **0.032** | 3.0 | 31.5 | 35.6 | **40.2** |
| | **0.063** | 0.2 | 26.7 | 30.3 | **30.5** |

## B.2 Results

In-line with the latest success in audio recognition, we consider an end-to-end audio model following SoundNet [1] architecture, that operates directly on the raw audio signal, without requiring any feature extractions (e.g., MFCC or log mel-spectrogram). We found that the end-to-end models show higher degree of vulnerability to the adversarial attacks, e.g., around 6% absolute degradation compared to the model operating on log mel-spectrogram. In case of the small vocabulary audio recognition task, we only consider FGSM attack and summarize our findings in Table 1. With a higher degree of stochasticity (i.e. $\theta = 0.9$), both the real and the binarized model exhibit much higher resilience to the adversarial attacks.

# References

[1] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*, 2016.

[2] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[3] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[4] Pete Warden. Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition. *arXiv e-prints*, art. arXiv:1804.03209, Apr 2018.