RecViT: Enhancing Vision Transformer with Top-Down Information Flow

Štefan Pócoš[a], Iveta Bečková[b] and Igor Farkaš[c]
Faculty of Mathematics, Physics and Informatics, Comenius University Bratislava, Mlynská dolina F1, 842 48 Bratislava, Slovakia

Keywords: Attention, Transformer, Recurrence, Adversarial Examples, Robustness, Heatmap.

Abstract: We propose and analyse a novel neural network architecture — recurrent vision transformer (RecViT). Building upon the popular vision transformer (ViT), we add a biologically inspired top-down connection, letting the network 'reconsider' its initial prediction. Moreover, using a recurrent connection creates space for feeding multiple similar, yet slightly modified or augmented inputs into the network, in a single forward pass. As it has been shown that a top-down connection can increase accuracy in case of convolutional networks, we analyse our architecture, combined with multiple training strategies, in the adversarial examples (AEs) scenario. Our results show that some versions of RecViT indeed exhibit more robust behaviour than the baseline ViT, on the tested datasets yielding ≈18 % and ≈22 % absolute improvement in robustness while the accuracy drop was only ≈1 %. We also leverage the fact that transformer networks have certain level of inherent explainability. By visualising attention maps of various input images, we gain some insight into the inner workings of our network. Finally, using annotated segmentation masks, we numerically compare the quality of attention maps on original and adversarial images.

1 INTRODUCTION

Recent advances in deep learning provide remarkable accuracy in many fields. Unfortunately, these advances do not often include the robustness of the systems, i.e. the ability to correctly process out-of-distribution data, such as adversarial examples (AEs) i.e., inputs created by addition of a subtle, yet carefully crafted noise which corrupts the correct classification (Szegedy et al., 2014). This often leaves the networks unprotected and unsuitable for security-critical applications, unless we can ensure a fully controlled environment. Therefore, a need for transparent, explainable, and interpretable models is rising (Vilone and Longo, 2020). Moreover, most of the current neural network models are purely feed forward, thus they only provide a bottom-up approach. On the other hand, a biologically more plausible way, which occurs also in the human visual cortex, is to combine bottom-up approach with top-down connections (Kietzmann et al., 2019).

In our work, we build upon the vision transformer model (Dosovitskiy et al., 2021), which we augment with a top-down connection. The extra information flow is modelled by recurrently passing the activations of the class token from the output layer into the input layer. This allows for a repetition of the classification process, where in subsequent iterations the network can take into account its previous output as well.

In addition, we also suggest several ways of modifying the network inputs, to fully utilise its capabilities. These are later evaluated on adversarially-generated data. We discover that one variant of RecViT seems to be more robust, with only a slight drop in accuracy on clean data. This supports the theory of robustness–accuracy trade-off (Tsipras et al., 2019). On the other hand, we also discover positive correlation between clean and adversarial accuracy in multiple runs of that specific variant, meaning that those networks that perform well on original data, are also more accurate on AEs, which seems to contradict the aforementioned trade-off.

Moreover, we provide comparison of the activations of AEs with the activations of test-set examples by visualising the models' attention maps. The difference is also evaluated numerically, by

[a] https://orcid.org/0000-0003-3799-7038
[b] https://orcid.org/0000-0002-6396-9770
[c] https://orcid.org/0000-0003-3503-2080

Pócoš, Š., Bečková, I. and Farkaš, I.
RecViT: Enhancing Vision Transformer with Top-Down Information Flow.
DOI: 10.5220/0012464700003660

In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024) - Volume 3: VISAPP*, pages 749-756
ISBN: 978-989-758-679-8; ISSN: 2184-4321

computing the similarity between annotated segmentation masks and the attention maps. Our results confirm that even a slight adversarial modification of the input results in great changes of the network attention. For reproducibility and transparency, our source codes can be found on the address https://github.com/Stefan78/RecViT.

This paper is structured as follows: First, we discuss the related work and similar architectures in section 2. Then, in section 3, we describe the data (both clean and adversarial) that we use for our experiments. Section 4 follows with detailed description of the proposed model and used training procedure. Results of experiments regarding network robustness are summarised in section 5. Further analysis of AEs through attention maps is provided in section 6. We conclude the paper and list the ideas for future work in section 7.

2 RELATED WORK

The idea of incorporating a top-down mechanism in convolutional neural networks for image classification led to improvement upon previous state-of-the-art models (Stollenga et al., 2014). The authors designed an adaptive weighting of convolutional kernels, which in subsequent iterations helped the network to focus on more specific parts of the image, instead of all image parts at once. In other lines of work, vision transformers (ViTs) (Dosovitskiy et al., 2021) have been proposed as the natural adaptation of the transformer architecture (Vaswani et al., 2017) for visual input. Currently, ViTs dominate the field of computer vision, thus their robustness against AEs is a hot topic. Recent research shows that they exhibit similar robustness as convolutional networks, albeit more specialised attacks still need to be considered (Bai et al., 2021).

Since the ViTs play such an important role in modern vision tasks, countless variations have been proposed, a few of them already including some form of recurrent connections. Perceiver (Jaegle et al., 2021) was designed to be able to scale to high dimensional inputs, by progressively reducing the dimensionality using attention modules with (potentially) shared weights. Another work (Gehrig and Scaramuzza, 2023) uses recurrent blocks composed of multiple various parts including convolutions, attention modules, LSTM modules, and more. Possibly the most similar to our architecture is the RViT (Messina et al., 2022). The key difference is that our model has a recurrent connection only in the class token, patch tokens are computed from the input in each iteration anew (thus allowing for an input sequence of slightly varied images), while RViT uses recurrence also in the patch tokens.

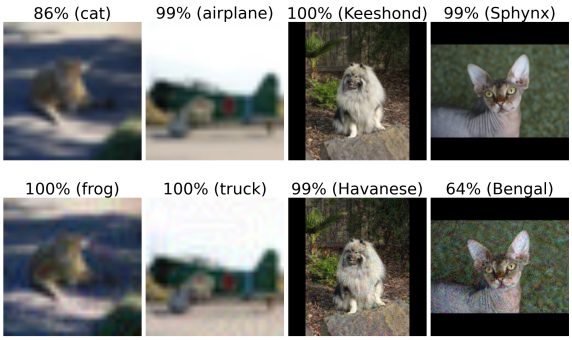

Figure 1: Two pairs of input images (top) with their corresponding adversarial examples (bottom) on CIFAR-10 (left half) and PET (right half) datasets.

To the best of our knowledge, ours is the first model to directly bind only the class token in ViTs, and send the information for further processing, without the need of having additional parameters.

3 DATA

3.1 Datasets

In this study, we analyse networks on two benchmark datasets for image classification: CIFAR-10 (Krizhevsky, 2009) and Oxford-IIIT Pet (PET) (Parkhi et al., 2012). CIFAR-10 consists of 32×32 pixel images, each belonging to one of ten classes of animals or vehicles. On the other hand, the PET dataset is much more diverse and complex, as there are 37 classes altogether with variable image resolutions, much higher than that of CIFAR-10. To achieve consistent representations and evaluation, we transform all the PET inputs to 224×224 pixel images. A major advantage of the PET dataset is the availability of pixel level trimap segmentations, distinguishing the object of interest, the background, and the area in between.

3.2 Generating Corrupted Input

To produce out-of-distribution images which are later used for exploring the network robustness (and providing explanations), we follow the well researched area of adversarial examples (AEs). AEs are such inputs to machine learning models, which cause intentional misclassification, even though they only slightly differ from the original, correctly classified

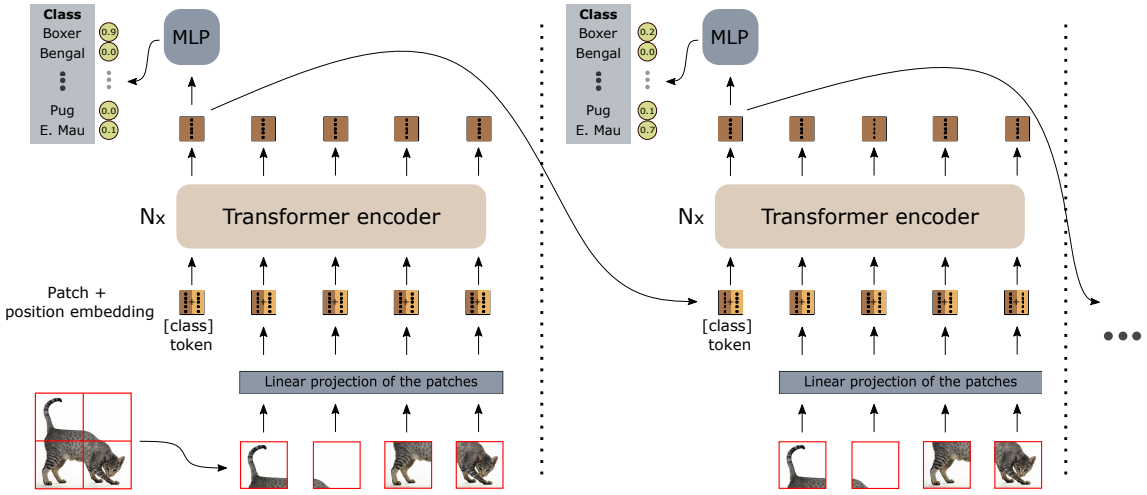

Figure 2: Two iterations of the RecViT architecture. After the first pass is computed, the resulting class token activation is used in the second pass, simulating the top-down information flow.

inputs (Szegedy et al., 2014). AEs can be crafted in numerous ways, however, when comparing the defences it is often tricky to generate AEs in an unbiased manner towards a certain model (Carlini et al., 2019). To avoid bias, we opt to leverage the transferability property of AEs (Liu et al., 2016), which basically means that an AE fooling a specific network might also fool other networks trained to perform the same task.

For a fair comparison of ViT vs RecViT, we do not generate the adversarial data on any transformer-type network. Therefore, we begin by training 3 convolutional networks with differing architectures on each of the two datasets. To ensure high classification quality, we fine-tune AlexNet, ResNet, and VGG, after which we achieve 85.03 % (AlexNet), 91.52 % (ResNet) and 90.75 % (VGG) accuracy on CIFAR-10 and 74.27 %(AlexNet), 87.22 %(ResNet) and 89.15 %(VGG) on PET dataset. The next step is to run the projected gradient descent attack (PGD) (Madry et al., 2018) individually on each network, resulting in AEs corresponding to the networks.[1] Since the PGD attack belongs to a class of white-box adversarial attacks with high transferability properties (Tramèr et al., 2018), we can use it to test the robustness of other networks. PGD attack utilises the gradient of the network w.r.t. the input image x, slightly modifying it over multiple iterations, according to the formula:

$$x(t+1) = x(t) + \alpha \text{sign}(\nabla_x L(\theta, x, y)), \qquad (1)$$

where α is the step size and L is the loss computed from θ, representing the network parameters, and y denotes the correct class. After each iteration, the input is projected onto an L_∞ ball of radius ε centred in the original input x. For the initial point $x(0)$ a random point within this ball is chosen.

In order to produce as diverse AEs as possible (using the PGD attack), we gradually increase the ε value, resulting in AEs with perturbation magnitude of $\varepsilon \in \{0.01, 0.02, ..., 0.2\}$ for CIFAR-10 and $\varepsilon \in \{0.07, 0.22, ..., 0.202\}$ for PET. To also evaluate the robustness on a set of particularly strong AEs, we specifically distinguish a set of 'cross-validated' (C-V) AEs i.e. those, which fool all of the three networks used for their generation. It can be assumed that these AEs are the most transferable, so they will have the highest success for a random, unprotected network. Altogether for each dataset we construct 4 groups of AEs: the first three are disjoint, they are AEs that were generated on AlexNet, ResNet, and VGG respectively. The fourth group are the cross-validated AEs. A sample of AEs for both of the datasets is shown in Fig. 1.

4 RECURRENT ViT

Building upon the Vision Transformer model (Dosovitskiy et al., 2021), we design a network enhanced with a recurrent connection. During image classification using ViT, one has the option of using the so-called class token, which serves as an extra representation (the rest are created by processing the image patches). The class token at the top of the network represents accumulated data about the image class, which is further inserted to a relatively simple MLP to produce the final classification. In our case, af-

[1]To generate AEs, we use the adversarial-robustness-toolbox (Nicolae et al., 2018).

ter the network forward pass, the class token is sent to a second iteration, in which the image patches are computed in the same manner, but the class token already contains relevant information about the image content.[2] Thus, we expect the network to focus on particular image regions, which coincide with the representations of the class token. A detailed scheme of the architecture can be found in Fig. 2.

4.1 Training

For weight initialisation we use the ViT-Ti pretrained model, which is a relatively small setup of a vision transformer, having $\approx 5.8M$ parameters (Gani et al., 2022). Next, we set a fixed number of iterations k after which we expect to have the final prediction. During the training, we use backpropagation through time (BPTT) to fully adjust the shared weights through the whole computational graph. Furthermore, we suggest two training modes, differing in the way the loss function is defined.

The **method 1 (M1)** minimises the prediction error only between the last prediction and the desired output. This emphasises the fact that only the last prediction matters, allowing the network to actually 'reconsider' and change its prediction from iteration to iteration:

$$Loss = L_{CE}(\theta_1; d, f_{\theta_2}(c_{k-1})), \tag{2}$$

where L_{CE} is the cross-entropy loss, θ_1 denotes the model parameters, d is the target class and f_{θ_2} is the MLP directing the classification of the class token c_{k-1} in the final iteration.

The **method 2 (M2)** optimises predictions in each time step. This promotes the idea that even the first 'guess' of the network should be valid (the subsequent iterations could then be interpreted as asking the network 'Are you sure that this image belongs to this class?'), and in case of different inputs across recurrence (which we elaborate on further in the text), that all predictions matter equally. The loss is constructed as follows (using the same nomenclature as in M1):

$$Loss = \sum_{t=0}^{k-1} L_{CE}(\theta_1; d, f_{\theta_2}(c_t)). \tag{3}$$

4.2 Input Modification Strategies

Besides providing options for the network to reconsider its initial prediction, recurrent connection also creates space to compute a single prediction from multiple different inputs. We take advantage of this property and provide a form of additional data into the network, making sure the original image is always included. Below, we describe the tested options, also shown in Fig. 3.

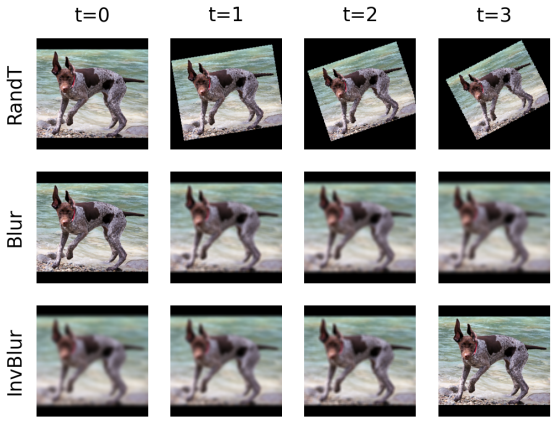

Figure 3: Visualisation of three different methods of input augmentation for RecViT within 4 iterations.

Vanilla. This version provides no augmentation. All inputs are the same picture.

Random Transform (RandT). Here we try to simulate the way people perceive an object. The object might be moving, or the person might be tilting their head, so that the object is relatively not at the same place for the duration of the object recognition. Therefore, we augment the original image by transforming it randomly, given translation, rotation, and scaling values. The transform parameters are computed beforehand and the transformation is repeated in each iteration (applied to the previous input), progressively altering the original input.

Blur. This strategy implements Gaussian blur of the input. By using blur of various strengths, we prompt the network to focus on different levels of detail. First, the input is unchanged, but as we proceed to further iterations the image gets more and more blurred.

Inverse Blur (InvBlur). The idea is based on blurring the images as well, however with inverted order of inputs. The first input is blurred the most and by the time we reach the final iteration, the image is perfectly clear. This aligns better with human perception, as people also see the 'big picture' first, and only notice the tiny details later.

[2]For the implementation we extend the timm library (Wightman, 2019).

Table 1: Average accuracy (in %) and standard deviation on PET dataset and generated AEs. For Baseline ViT we trained 10 runs, all other variants were 5 runs. Experiments were run for $k \in \{2,3,4\}$. For brevity, we only include results of the best performing k for each RecViT variant and training method.

	Test-set	AlexNet AEs	ResNet AEs	VGG AEs	C-V AEs
RecViT variant	Mean±Std	Mean±Std	Mean±Std	Mean±Std	Mean±Std
Baseline ViT	88.80±0.57	9.26±1.37	14.67±1.90	8.96±1.87	10.66±1.42
Vanilla M1, k=3	89.60±0.54	9.52±1.14	17.19±1.66	10.10±0.97	11.83±1.02
Vanilla M2, k=2	88.67±0.31	9.68±0.90	19.62±1.21	11.49±0.97	13.06±0.81
RandT M1, k=2	89.02±0.24	11.24±0.89	17.58±2.35	12.65±1.56	13.58±1.22
RandT M2, k=2	88.05±0.99	11.17±0.98	18.18±3.94	12.22±0.57	13.28±1.49
Blur M1, k=3	85.23±5.66	12.75±3.34	16.89±3.79	11.36±3.75	13.43±3.56
Blur M2, k=4	87.42±1.56	13.43±2.81	18.83±3.58	12.54±3.14	14.62±3.09
InvBlur M1, k=2	**86.99±0.98**	**15.57±3.60**	**24.91±10.70**	**17.83±6.95**	**18.93±6.63**
InvBlur M2, k=2	85.27±1.28	17.04±1.21	28.65±3.79	17.87±2.03	20.54±2.15

5 RESULTS

In this section we analyse in detail the classification capabilities of RecViT models on original data and AEs (further referred to as the robustness of the networks).

5.1 Robustness Evaluation

We start with training RecViT networks for each combination of the two methods of loss computation, and four strategies of data manipulation, with varying number of maximum iterations: $k \in \{2,3,4\}$. This results in 24 models and for each of those we have 5 runs, to ensure statistically sound evaluation. These networks are then tested for robustness, as well as for the accuracy on clean data. In Table 1 we display a subset of the results, where for each RViT variant we choose the best k. For baseline we trained 10 runs of unmodified ViT (Baseline ViT).

As we can see in comparison with the Baseline ViT, there is a slight increase in robustness for RandT and Vanilla networks, without significant decrease in accuracy. On the other hand, the networks with blurred input yield higher robustness, though with a slight ($\approx 3\%$) drop in accuracy. This trade-off is further elaborated on in subsection 5.3.

5.2 Comparison with ViT

Since our best performing model (regarding the robustness and clean accuracy) used blurred data, we further analyse the contribution of recurrent connection. In order to do a fair comparison, the ViT should use the same data as the RecViT. For that means we simulated the input conditions for a Baseline ViT. When comparing with RecViT with a given k (Blur or InvBlur), we train the ViT using data with the same amount of blur, as are in the inputs to RecViT (in each iteration, the exact amount was chosen randomly amongst the possible values). In the testing phase, instead of computing a single forward pass (for the original image), we compute k passes, one for each of the possible inputs with varying amount of blur. The final classification is computed by averaging the logits

Table 2: Comparison of accuracy and robustness of RecViT models with ViT Blur models.

PET		
	Test-set	C-V AEs
RecViT variant	Mean±Std	Mean±Std
Baseline ViT	88.80±0.57	10.66±1.42
InvBlur M1, k=2	**86.99±0.98**	**18.93±6.63**
InvBlur M1, k=3	86.32±1.25	14.83±5.48
InvBlur M1, k=4	85.33±2.74	10.58±2.16
InvBlur M2, k=2	**85.27±1.28**	**20.54±2.15**
InvBlur M2, k=3	78.09±7.14	21.83±4.91
InvBlur M2, k=4	74.31±5.54	22.36±2.26
ViT Blur k=2	**86.21±0.83**	**18.20±1.25**
ViT Blur k=3	79.19±1.50	15.33±2.03
ViT Blur k=4	68.45±3.37	12.71±2.21

CIFAR-10		
	Test-set	C-V AEs
RecViT variant	Mean±Std	Mean±Std
Baseline ViT	97.64±0.11	45.78±1.92
InvBlur M1, k=2	97.44±0.09	43.89±2.98
InvBlur M1, k=3	97.43±0.12	44.85±2.12
InvBlur M1, k=4	**97.44±0.04**	**45.09±2.04**
InvBlur M2, k=2	**96.46±0.14**	**67.67±1.67**
InvBlur M2, k=3	95.12±0.25	69.88±1.95
InvBlur M2, k=4	94.97±0.15	68.32±1.01
ViT Blur k=2	**97.53±0.05**	**50.96±3.13**
ViT Blur k=3	96.04±1.25	54.61±2.04
ViT Blur k=4	93.51±1.47	53.11±2.91

Table 3: Average accuracy and robustness (in %) of the top 3 runs (according to the test-set performance) from InvBlur RecViT models trained on PET dataset.

	Test-set	C-V AEs
InvBlur M1, k=2	88.13	28.44
InvBlur M1, k=3	87.68	21.11
InvBlur M1, k=4	88.15	12.47
InvBlur M2, k=2	87.14	24.02
InvBlur M2, k=3	85.59	24.28
InvBlur M2, k=4	84.01	26.53

from individual forward passes. This way we generate networks similar to our RecViTs with blurred inputs, referred to as ViT Blur.

The resulting Table 2 including PET and CIFAR-10 results show consistent trend. Our RecViT outperforms the ViT Blur mainly regarding the robustness–accuracy trade-off, suggesting that the recurrent connection in RecViT indeed utilises the better computational capabilities of the networks.

5.3 Robustness–Accuracy Trade-Off

In previous two sections we determined the most robust models to be the RecViT InvBlur trained with the method 2. However, those have in some cases (mostly for the PET dataset) somewhat unstable performance, the accuracy and robustness seems to vary a lot. From this we deduce that since the images are blurred and the PET is more challenging task than CIFAR-10, the instability may result from sub-optimal hyperparametes. The solution would be to perform a more thorough hyperparameter search, or to train multiple runs and choose the best performing model on the validation set. We chose the second option and trained a larger sample of InvBlur RecViT models (15 runs for each combination of training method and $k \in \{2,3,4\}$). Given the results in Table 3 where we take the top 3 best performing models from each RecViT group, we end up with substantially better robustness-accuracy trade-off.

In Fig. 4 we visualise the robustness–accuracy trade-off for one of the best performing RecVit models with the ViT Blur counterpart. Due to having different number of runs, we include ViT Blur models with varying k. Moreover, we computed the correlation coefficient between accuracy and robustness for all InvBlur RecViT models. To our surprise, we detected average (across k) correlation coefficients of 0.57 and 0.81 for InvBlur RecViT networks trained with M1 and M2 respectively. This also means that further enhancing the accuracy via deeper optimisation might heighten the robustness levels as well.

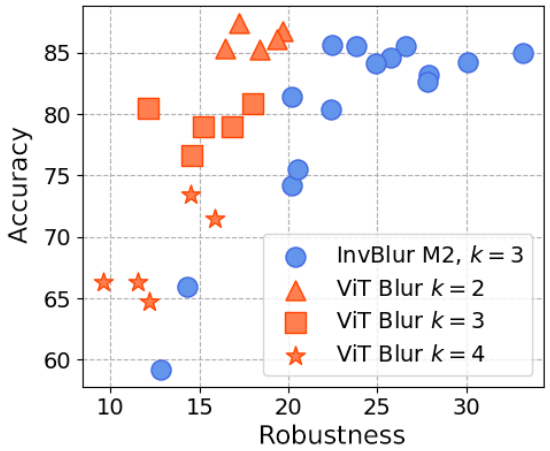

Figure 4: Visualisation of the robustness vs accuracy of individual InvBlur RecViT networks trained with M2 and k=3, compared to ViT Blur models with various k.

6 HEATMAP COMPARISON

Unlike traditional convolutional networks, ViTs come with an inherent way to depict their inner behaviour. The same holds for the RecViT. By having self-attention module partly defined by the equation

$$\text{attention}(Q,K,V) = \text{softmax}(\frac{QK^{\text{T}}}{\sqrt{\text{d}_\text{k}}})V, \quad (4)$$

we are able to directly extract the information flow for the class token in the form of importances per each image patch (Vaswani et al., 2017). To our advantage, we can compare the heatmaps extracted from RecViT with the trimap segmentation masks for PET dataset. Our goal is to compare the heatmaps in RecViT over input from various sources.

The key comparison is to visualise the activations of AEs vs original images. There are studies, which compare heatmaps of AEs vs originals (Dong et al., 2017; Xu et al., 2019; Kotyan and Vargas, 2021). However, those were based either on other visualisation methods (not inherent) or different aspects of AEs.

Inspired by recent work (Rieger and Hansen, 2020), we compute the cosine similarity between produced attention maps and the trimap segmentation masks. To normalise both to the same range, we first scale the attention maps linearly to the range $[-1,1]$. Then we also modify the segmentation maps, so that the value of -1 corresponds to the background (and the black padded area), 1 to the object of interest, and 0 to the border area. The similarity between the segmentation mask S and the attention scores A is computed according to equation:

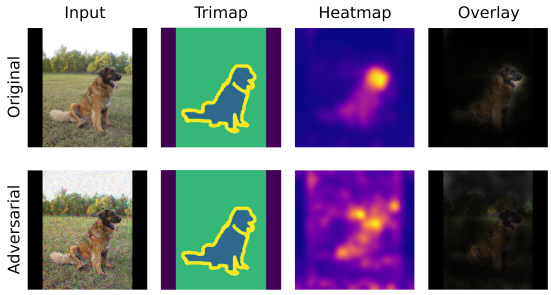

Figure 5: Comparison of attention activation for an AE and the original image in a vanilla RecViT network on layer 8, trained with method 1, k=3.

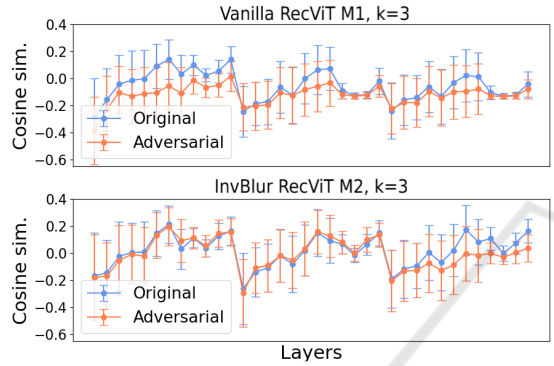

Figure 6: Development of similarity scores of AEs vs original images on two selected networks, computed using 150 AE/original pairs. The x-axis (layers) represents individual attention modules.

$$\text{cosine sim}(S,A) = \frac{\sqrt{\sum_{i=1}^{N} A_i S_i}}{\sqrt{\sum_{i=1}^{N} A_i^2}\sqrt{\sum_{i=1}^{N} S_i^2}}. \quad (5)$$

After visually examining the attention maps, we clearly see that the AEs have worse overlay with the true object — we can see an example of that in Fig. 5. To support our claim that this phenomenon occurs consistently we plot the average cosine similarity between segmentation masks and attention scores of AEs and the original images on a selected RecViT network (Fig. 6). High deviation in the similarity scores, present in the graph, is mostly caused by the varying area of the objects of interest in the images. As the matter of fact, the AEs only rarely produce a better match with the segmentation masks than the original examples, and this holds for all of our tested networks. Interestingly, in some RecViT networks (particularly more robust ones) we notice that the overlap of AEs is on par with that of the original examples. This observation could be investigated in a future, more detailed analysis.

7 CONCLUSION

We proposed a novel transformer-like architecture with recurrence in class token (RecViT), to simulate a top-down connection, providing the network an option to reconsider its initial prediction. Since it can also process a sequence of inputs instead of a single image, the computational capacity of the RecViT seems to be better utilised. We hypothesise that the capacity increase could be the reason why some variants of RecViT demonstrate higher robustness against the tested adversarial examples. This behaviour could be further investigated more in depth using more diverse attack methods.

One of the most prominent results is the fact that BlurInv RecViT networks, which showed the highest accuracy on AEs, exhibit high positive correlation between clean and adversarial accuracy. This seems to contradict the robustness–accuracy trade-off. This positive correlation also suggests that clean accuracy (which is often known) could be used as a guide to pick the best performing networks, without risking a drop in robustness. Choosing the best performing models we achieved $\approx$ 18% increase in robustness with only 1 % drop in clean accuracy on PET dataset. On CIFAR-10 the drop in clean accuracy was similar, while robustness increase reached up to 22 %.

Since the idea of having a top-down connection in ViTs has proven useful, it would be beneficial to further investigate this model. Some ideas for a follow-up, which we hope will bring improvements, are to use adversarial training and to have more robust data augmentation while categorising a single input.

Yet another usage of RecViT is to exploit the differences in heatmaps of the self-attention modules, when passing through a normal example vs an adversarial example. This discrepancy could be used as a partial defence against adversarial attacks.

ACKNOWLEDGEMENTS

This research was carried out in the framework of the Horizon Europe project TERAIS, GA No. 101079338. It was funded in part by Horizon 2020 project TAILOR, GA No. 952215 and by national project KEGA 022UK-4/2023.

REFERENCES

Bai, Y., Mei, J., Yuille, A. L., and Xie, C. (2021). Are transformers more robust than CNNs? In *Advances in*

Neural Information Processing Systems, volume 34, pages 26831–26843. Curran Associates, Inc.

Carlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., Goodfellow, I., Madry, A., and Kurakin, A. (2019). On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*.

Dong, Y., Su, H., Zhu, J., and Bao, F. (2017). Towards interpretable deep neural networks by leveraging adversarial examples. *arXiv preprint arXiv:1708.05493*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Gani, H., Naseer, M., and Yaqub, M. (2022). How to train vision transformer on small-scale datasets? *arXiv preprint arXiv:2210.07240*.

Gehrig, M. and Scaramuzza, D. (2023). Recurrent vision transformers for object detection with event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13884–13893.

Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., and Carreira, J. (2021). Perceiver: General perception with iterative attention. In *International Conference on Machine Learning*, pages 4651–4664. PMLR.

Kietzmann, T., Spoerer, C., Sörensen, K., Cichy, R., Hauk, O., and Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, 116:201905544.

Kotyan, S. and Vargas, D. V. (2021). Deep neural network loses attention to adversarial images. *arXiv preprint arXiv:2106.05657*.

Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. Technical report, University of Toronto.

Liu, Y., Chen, X., Liu, C., and Song, D. (2016). Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*.

Messina, N., Amato, G., Carrara, F., Gennaro, C., and Falchi, F. (2022). Recurrent vision transformer for solving visual reasoning problems. In *International Conference on Image Analysis and Processing*, pages 50–61. Springer.

Nicolae, M.-I., Sinn, M., Tran, M. N., Buesser, B., Rawat, A., Wistuba, M., Zantedeschi, V., Baracaldo, N., Chen, B., Ludwig, H., Molloy, I., and Edwards, B. (2018). Adversarial robustness toolbox v1.2.0. https://arxiv.org/pdf/1807.01069.

Parkhi, O. M., Vedaldi, A., Zisserman, A., and Jawahar, C. V. (2012). Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Rieger, L. and Hansen, L. K. (2020). A simple defense against adversarial attacks on heatmap explanations. *arXiv preprint arXiv:2007.06381*.

Stollenga, M. F., Masci, J., Gomez, F., and Schmidhuber, J. (2014). Deep networks with internal selective attention through feedback connections. *Advances in Neural Information Processing Systems*, 27.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2014). Intriguing properties of neural networks. In *International Conference on Learning Representations*.

Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., and McDaniel, P. (2018). Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. (2019). Robustness may be at odds with accuracy. In *International Conference on Learning Representations*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Vilone, G. and Longo, L. (2020). Explainable artificial intelligence: a systematic review. *arXiv preprint arXiv:2006.00093*.

Wightman, R. (2019). Pytorch image models. *GitHub repository*.

Xu, K., Liu, S., Zhang, G., Sun, M., Zhao, P., Fan, Q., Gan, C., and Lin, X. (2019). Interpreting adversarial examples by activation promotion and suppression. *arXiv preprint arXiv:1904.02057*.