ATCON: Attention Consistency for Vision Models Supplementary Materials

Ali Mirzazadeh*[1,2], Florian Dubost*[1], Maxwell Pike[1]
Krish Maniar[1], Max Zuo[2], Christopher Lee-Messer[1], and Daniel Rubin[1]

[1]Stanford University
[2]Georgia Institute of Technology

alimirz@gatech.edu, floriandubost1@gmail.com, {cleemess, rubin}@stanford.edu
*equal contribution

1. Related Work on Attention Maps

Below, we detail the major types of attention maps methods. Zhou et al. [10] proposed the Class Activation Maps (CAM) method. Attention maps are computed as a linear combination of the feature maps of the last convolutional layer of a neural network. The network needs to have a global pooling layer after this last convolutional layer, subsequently followed by a fully connected layer to map to the outputs. For a given output neuron–e.g. a class in multiclass classification–the weights of the linear combination of feature maps are chosen as the weights of the fully connected layer mapping to that output.

This approach requires a specific architecture (global pooling and fully connected layers), which limits its applicability. *Grad-CAM* [6] also computes attention maps as linear combination of features maps but computes the weights differently, using the backpropagation algorithm. The global pooling layer is not needed anymore, and attention maps can be computed from any layer in any network architecture.

The backpropagation algorithm is also used by Simonyan et al. [7] to compute attention maps in a completely different manner. Simonyan et al. [7] propose to compute attention maps by estimating the gradient of the output with respect to the input signal, which consequently creates a bijective mapping between the input signal and corresponding attention map. Springenberg et al. [8] notice that Simonyan et al.'s method [7] creates interference patterns on the attention maps and that these patterns originate from negative gradients flowing back in the rectified linear unit (ReLU) activations. Springenberg et al. [8] propose to modify the behavior of ReLU during backpropagation for the creation of an attention map, and set these negative gradients to zero. This effectively removes the interference patterns. The authors call their method: *Guided Backpropagation*.

In practice, attention maps often have a higher resolution with Guided Backpropagation –that of the input signal– than with Grad-CAM, where the attention maps are often computed from pooled feature maps. On the same note, Grad-CAM tends to highlight larger regions of the input, while Guided Backpropagation focused more on details, and is sometimes biased toward saliency, e.g. image regions with high-intensity gradients [1].

Recently, Transformer Networks [9] have been also been used to compute attention. The attention mechanism is directly incorporated to the network architecture. While this makes the interpretation of attention more explicit, it also limits the type of architecture that can be used, which is in a way, similar to the model-specific CAM.

The last category of attention map computation methods is perturbation methods. These methods compute attention maps by applying random perturbations to the input and observe the changes in the network output. They are completely model-agnostic. For example, Petsiuk et al. [5] compute attention maps with masking perturbations. Fong et al. [3] proposed several other perturbation techniques including replacing a region with a constant value, injecting noise, and blurring the input.

2. Experiments on SVHN

In Table 1, we show results on SVHN dataset [4]. Those results are not as strong as those on the other datasets (PASCAL, video). We assumed that this is because the difficulty of SVHN is not related to locating the objects (digits) but rather differentiating between digits. The location is not a strong discriminative feature there, and consequently forcing attention to improve localization does not tremendously

Table 1. **Classification results (F1) on SVHN. Validation set.**

Imgs Per Class	2	4	8	12	16
ResNet	0.137	0.224	0.324	-	-
Proposed	0.160	0.247	0.331	-	-

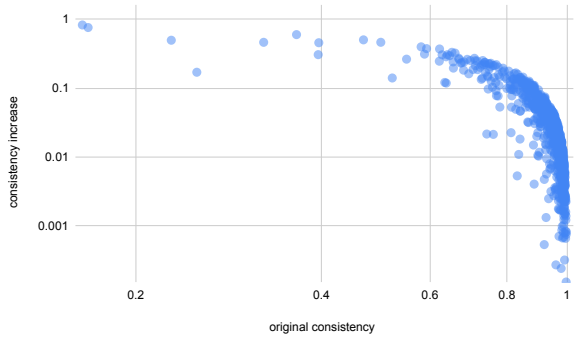

Figure 1. **Gain of attention consistency.** The x-axis shows the attention map consistency of the model before attention map consistency fine-tuning. The y-axis shows the gain of attention map consistency after fine-tuning with the proposed method.

improve classification performance.

3. Statistics of the hospital video dataset

In Table 2, 3, 4, and 5, we show statistics of the hospital video dataset such as the number of clips (windows) per class, statistics of the length of each clip, and number of patients per class.

4. Gain of attention consistency

In Figure 1, we show plot the gain of attention consistency provided by the proposed method on the PASCAL test set.

5. Details of few-shot learning experiments (LaSO [2])

The few shot learning baseline is LaSO [2]. This is a multi-label few shot learning method. LaSO compute operations on label sets (such as union or intersection) for image pairs in feature space and consequently creates combination of labels that are not present in the original training set. We apply LaSO to our PASCAL dataset, where only the classes of object presents in the image (not the bounding boxes) are used as image-level multi-label for training. Few shots learning experiments presented in the main body use ResNet architecture with data augmentation. Experiments settings are the same as for the other PASCAL experiments and are described in section 4.2 in the main body. The data splits are also the same as those utilized in the rest of the article.

6. Discussion on threshold for ATCON

In our experiments, we observed that ATCON is only beneficial when the training dataset was small. When the training set size increases, there is no significant difference between the proposed method and baseline. We believe that this indicates that when the training dataset is large enough, more accurate representations are learnt, and forcing the conception of attention consistency alone is not sufficient to further improve the accuracy of learnt representations. More specifically, although the F1 score in Table 2 (main body) is slightly lower for 16 and 135 images per class, after statistical testing, these differences were found not significant. Note that, for these dataset sizes, the mAPs are the same for the proposed method and baseline (Table 2, main body), and that the object detection metrics (Table 3, main body) are slightly better for the proposed method, although, again this is not statistically significant. Finding a general threshold that determine when ATCON is beneficial is challenging. For PASCAL, that threshold lies between 8 and 12 images per class, while for the video dataset a substantial improvement was still observed for 16 clips per class. In addition to the number of samples per class, we suppose that this threshold depends on the difficulty of the dataset, and the number of classes. We assume that the threshold relates to the number of sample necessary to correctly detect the object in the image. For new datasets, we suggest using the method on a series of small training sets and extrapolating the performance gain for larger training set sizes.

Table 2. **Statistics of the full hospital video dataset.**

Label	Nbr of windows	Average length (sec)	Median length (sec)	Standard Deviation (sec)	Total length (sec)	Number of patients
Suctioning	45	14.36	10	14.46	646	20
Chewing	15	89.07	45	91.23	1336	12
Rocking	21	54.67	24	71.08	1148	10
Cares	44	88.66	46	170.59	3901	23
Patting	33	38.64	21	40.49	1275	9
All	158	52.57	25	104.47	8306	59

Table 3. **Statistics of the training split of the hospital video dataset.**

Label	Nbr of windows	Average length (sec)	Median length (sec)	Standard Deviation (sec)	Total length (sec)	Number of patients
suctioning	12	20.92	16	14.92	251	5
chewing	5	147.4	130	98.04	737	4
rocking	6	92.33	56	97.53	554	2
cares	18	122.56	41	255.59	2206	8
patting	10	42.2	18	56	422	3

Table 4. **Statistics of the validation split of the hospital video dataset.**

Label	Nbr of windows	Average length (sec)	Median length (sec)	Standard Deviation (sec)	Total length (sec)	Number of patients
suctioning	18	13.5	11	10.96	243	7
chewing	6	44.33	20	42.13	266	5
rocking	8	45.75	24	53.7	366	4
cares	15	75.33	65	50.24	1130	10
patting	17	31.41	17	28.8	534	4

Table 5. **Statistics of the testing split of the hospital video dataset.**

Label	Nbr of windows	Average length (sec)	Median length (sec)	Standard Deviation (sec)	Total length (sec)	Number of patients
suctioning	15	10.13	7	15.88	152	8
chewing	4	83.25	44	96.14	333	3
rocking	7	32.57	17	43.74	228	4
cares	11	51.36	34	50.54	565	5
patting	6	53.17	52	32.53	319	2

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, pages 9524–9535, 2018.

[2] Amit Alfassy, Leonid Karlinsky, Amit Aides, Joseph Shtok, Sivan Harary, Rogerio Feris, Raja Giryes, and Alex M Bronstein. Laso: Label-set operations networks for multi-label few-shot learning. In *CVPR*, pages 6548–6557, 2019.

[3] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[4] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[5] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *British Machine Vision Conference*, 2018.

[6] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017.

[7] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *International Conference for Learning Representations Workshop*, 2014.

[8] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. In *International Conference for Learning Representations*, 2015.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[10] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016.