Attention Consistency for Vision Models

Anonymous CVPR submission

Paper ID 3714

Abstract

Attention–or attribution–maps methods are methods designed to highlight regions of the model's input that were discriminative for its predictions. However, different attention maps methods can highlight different regions of the input, with sometimes contradictory explanations for a prediction. This effect is exacerbated when the training set is small. This indicates that either the model learned incorrect representations or that the attention maps methods did not accurately estimate the model's representations. We propose an unsupervised fine-tuning method that optimizes the consistency of attention maps and show that it improves both classification performance and the quality of attention maps. We propose an implementation for two state-of-the-art attention computation methods, Grad-CAM and Guided Backpropagation, which relies on an input masking technique. We evaluate this method on our own dataset of event detection in continuous video recordings of hospital patients aggregated and curated for this work. As a sanity check, we also evaluate the proposed method on PASCAL VOC. With the proposed method, with small training sets, we achieve a 6.6 points lift of F1 score over the baselines on our video dataset, a 2.9 point lift of F1 score on PASCAL, and a 1.8 points lift of mean Intersection over Union over Grad-CAM for weakly supervised detection on PASCAL. Those improved attention maps may help clinicians better understand vision model predictions and ease the deployment of machine learning systems into clinical care.

1. Introduction

In many real-world problems such as healthcare, labeled training data can be scarce [3, 17], which can drive models to learn partially-incorrect representations and overfit to their training set [2, 30]. Consequently, in small datasets, researchers need to ensure that correct representations were learned. Those representations should match human understanding, be generalizable to unseen data, and not focus on potential bias in the dataset. This is especially essential to healthcare machine learning systems, where interpretability can justify predictions and decisions.

Attention–or attribution–map methods can be used the evaluate the representations of a model by highlighting regions in the input signal that are discriminative for the model's predictions. They have established themselves as one of the main methods for analyzing the interpretability of neural networks [32], and verify that the model did not leverage bias present in the data. However, attention map results can greatly vary depending on the chosen attention computation method, to the point of being sometimes contradictory [1]. Moreover, some of these methods have been shown to be biased to irrelevant patterns in the data, such as regions of high-intensity gradients in images [1]. Attention maps also become more dissimilar when the task becomes more challenging and the chance of overfitting increases. For example, Dubost et al. [7] show that the difference of weakly supervised detection performance between the Grad-CAM attention map [19] and the Grad attention map [20] is larger in datasets for which the overall detection performance is worse.

Consequently, do models with harmonized attention, i.e. having similar attention maps computed from multiple methods, result in improved representations?

We propose to enforce consistency between attention maps computed using different methods to improve the representations learned by the model and increase its classification performance on unseen data. More specifically, we design an attention consistency loss function for two state-of-the-art attention maps methods: Grad-CAM [19] and Guided Backpropagation [23]. We propose to optimize this loss function as an unsupervised fine-tuning step, to improve the representations of pretrained models.

We show that the proposed method can improve classification performance in video clip event classification with our own dataset curated for this project. The video dataset consists of clips extracted from continuous video recordings of hospital patients in their rooms. As a sanity check, we also show improvement in image classification with PASCAL VOC [8] when the size of the training set is reduced. We show that attention consistency improves the quality of attention maps. Attention maps are analyzed qualitatively on the video dataset, and quantitatively on PASCAL by computing the overlap between thresholded attention maps and ground truth bounding boxes. The benefits of the method are demonstrated for multiple network architectures: ResNet 50, Inception-v3, and a 3D 18 layers ResNet. For the video dataset, we show that the proposed method can leverage the state-of-the-art self-supervised method SimCLR [4] to further boost performance. The improved attention maps may help clinicians better understand model predictions, ease the deployment of machine learning systems into clinical care, and eventually improve patients' outcomes.

2. Related works

3. Attention Maps

Below, we detail the major types of attention maps methods. Zhou et al. [33] 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* [19] 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. [20] to compute attention maps in a completely different manner. Simonyan et al. [20] 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. [23] notice that Simonyan et al. [20]'s method 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. [23] 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 [27] 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. [18] compute attention maps with masking perturbations. Fong et al. [9] proposed several other perturbation techniques including replacing a region with a constant value, injecting noise, and blurring the input.

3.1. Attention Map Consistency

To the best of our knowledge, consistency between multiple attention maps methods has not yet been used to evaluate and optimize a model's performance. The closest works entail inspecting the consistency of the same attention maps method under different augmented versions of the input [10], or across the layers of the same network [29]. Li et al. [16] extend the idea of Guo et al. [10] to attention consistency between image shared similar features. Xu et al. [31] combine the articles cited above to enforce consistency of attention maps both across augmented images and across network layers.

However, none of these articles inspect consistency across multiple attention map methods. The core mechanics of different attention map methods can be substantially different, and their ranking in terms of weakly supervised detection performance can greatly vary between datasets [7]. In supplementary materials, we detail the major types of attention maps methods, including those utilized in this article, namely Grad-CAM [19] and Guided Backpropagation [23].

3.2. Semi-supervised and Self-supervised Learning

When training labels are scarce, the go-to approach becomes unsupervised and semi-supervised learning, and especially self-supervised learning. In image classification, most state-of-the-art semi-supervised methods are based on self-supervision and use contrastive learning approaches [4, 11]. Momentum Contrast (MoCo) [11] encodes and matches query images to keys of a dynamic dictionary. SimCLR [4] improves upon MoCo by removing the need for

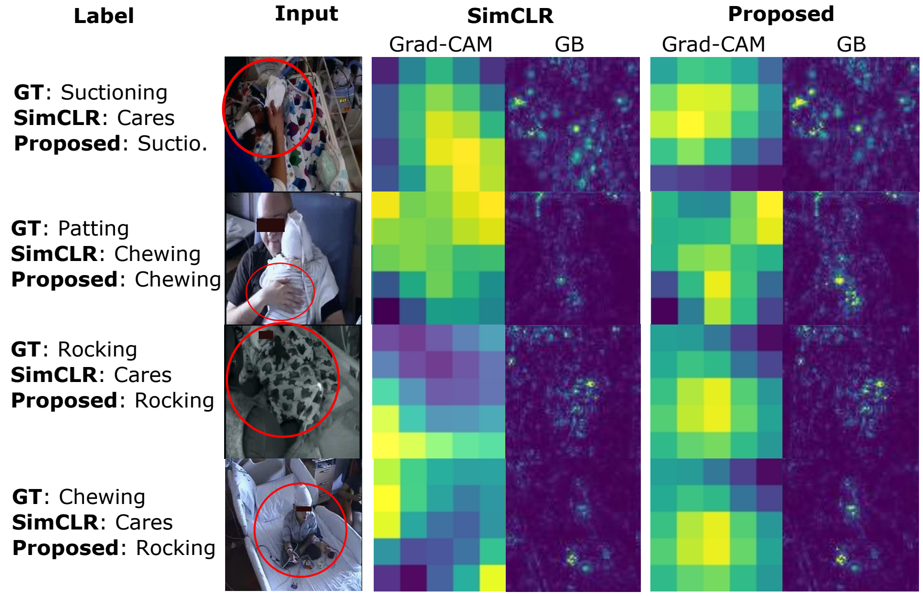

Figure 1. **Examples of attention maps for hospital video data.** Comparison between SimCLR and SimCLR plus the proposed unsupervised attention consistency fine-tuning (Proposed). The first column indicates the ground truth (GT) label, the label predicted by SimCLR, and by the proposed method. The second column shows one of 16 input frames. We circled in red the region where the event is occurring. The two middle columns show Grad-CAM and Guided Backpropagation (GB) for SimCLR, and the last two columns show the same attention maps for the proposed method. Models were training with 16 samples per class. Examples showing substantial differences between methods were selected.

specialized architectures. The authors of SimCLR stress that the composition of data augmentation is crucial in achieving high performance. Another key parameter is the batch size. SimCLR works best with large batch sizes, which can become a strong limitation when working with high dimensional data such as video data. Jing et al. [13] proposed a semi-supervised learning method designed for video classification, using pseudo-labels and normalization probabilities of unlabeled videos to improve the classification performance.

Few-shot learning methods can also leverage small datasets by requiring only a few labeled samples. For example, FixMatch [22] creates pseudo labels for unlabeled data using the current model's logits and include a cross-entropy term on those pseudo labels in the loss function. Prototypical network [21] is another few-shot learning method which allows networks to generalize to new classes with limited samples for those classes.

Yet, most unsupervised, semi-supervised or self-supervised learning methods assume that a large quantity of (unlabeled) data are available, together with large computational resources to store and process them. In many real-world applications, such as medical applications, both data and computational resources are often lacking, especially for smaller institutions or emerging modalities. Acquisition costs can be high because of the price of the acquisition device, or because experimental settings deviate from clinical practice and require the setup of a dedicated research acquisition environment. For example, MRI scanners are expensive and only allow a limited number of patients to be scanned at once in an hospital. Experiments with MRI settings that target resolutions beyond 1mm isotropic are rarely used in clinical practice and require a specific environment. For example, Cicek et al. [5] published a prominent article in medical image analysis that only use 3 sparsely labeled 3D microscopy images in their study, and many medical image analysis competitions only provide a few scans for participants [15,25] also partially due to administrative cost relating to patient consent and personal health information.

In this article, we propose a regularization method that improves the classification and interpretability of neural networks on very small datasets, without requiring additional data.

4. Method

We have seen in the literature that attention maps become more dissimilar when the task becomes more challenging and the chance of overfitting increases [7]. Consequently, we propose to improve the representations learned by classifiers by enforcing consistent representations across different types of attention maps. Firstly, we detail the concept of attention map consistency, and secondly, we propose an implementation for two state-of-the-art attention functions:

Grad-CAM [19] and Guided Backpropagation [23].

4.1. Attention Map Consistency

Let us consider a set X of N samples x_n with their corresponding labels y_n, and a classifier f with parameters θ, such that $f(x) = \hat{y}$. By definition, an attention function g allows us to compute an attention map A for a given input x and classifier f such that $g(f, x) = A_x$. The attention map A has values in $\mathbb{R}$ and highlights the subset of x that was most informative to the predictor, or that relates the most to its prediction. For example, if x is an image, the attention map A highlights areas of the images that are discriminative for the target prediction. While the dimensions of A and x do not have to be the same, we assume that there is at least a surjective mapping from x to A, such that any element of the input x can be linked to an element of A (its discriminative power for the task).

Let us now consider a set of M attention functions g_m such that $g_m(f, x) = A_{x,m}$ for an input x. We want to maximize the correlation h between the attention maps $A_{x,m}$ for the full set of input samples x_n such that we seek to solve:

$$\max_{\theta}\{\mathbb{E}_{x\in X}[h(A_{x,1}, ..., A_{x,M})]\} \tag{1}$$

The correlation h should be chosen to be high when the attention maps highlight similar regions of the input. The choice of the correlation function h depends on the type of attention map to compare. We detail our choice of h for Grad-CAM and Guided Backpropagation in section 4.3 and propose ablation studies.

The corresponding attention map consistency loss function is defined as

$$L_A = -\sum_{x\in X} h(A_{x,1}, ..., A_{x,M}). \tag{2}$$

4.2. Training Strategy: Unsupervised Fine-Tuning

The attention consistency loss does not require any training labels, which means that unsupervised training is possible. However, for the loss to converge to a desirable minimum, the classifier needs to have already learned meaningful representations prior to optimizing the consistency between attention maps. Hence, we propose the optimize the attention map consistency a posteriori, as an unsupervised fine-tuning step.

In the ablation study of the experiment section, we compare unsupervised fine-tuning to adding the consistency loss to the standard classification loss during training (categorical cross-entropy) by computing a linear combination of the two losses. We call this other strategy *combined* optimization. We also compare to alternating between the two losses batch-wise as proposed by Chen et al. [3]. We call this last strategy *alternated* optimization.

Further details about training hyperparameters, transformations, optimizers, and libraries are given in experiment section, Section 5.2.

4.3. Consistency between Grad-CAM and Guided Backpropagation

We chose to implement the attention consistency loss using two state-of-the-art attention functions: Grad-CAM [19] and Guided Backpropagation [23]. We selected these methods as they can be computed for any type of convolutional network and because their computation is different enough to witness significant changes in the attention consistency loss during training.

Grad-CAM attention map $A_{Grad-CAM}$ is computed as a linear combination of the k feature maps f_k of a target convolutional layer:

$$A_{Grad-CAM} = \sum_{k}^{N} \alpha_k f_k \tag{3}$$

$$\alpha_k = \frac{1}{Z}\sum \frac{\partial \hat{y}}{\partial f_k}, \tag{4}$$

where each weight α_k is computed as the average of the derivative of the output $\hat{y}$ with respect to the feature maps f_k, and where Z is the size of the feature map f_k. We choose the last convolutional layer of our network's architecture to compute Grad-CAM, as it is the most closely related to the output. More generally, the earlier the layer, the less related to output the features can be expected to be.

Guided Backpropagation is computed by estimating the gradient of the network's output $\hat{y}$ with respect to the network's input x:

$$A_{GB} = \frac{\partial \hat{y}}{\partial x}. \tag{5}$$

For multiclass classification problems, both Grad-CAM and Guided Backpropagation compute class-wise attention maps. We compute the attention map consistency loss only using the attention map of the top predicted class.

As mentioned in the related work section, Guided Backpropagation attention maps often have a higher resolution than that of Grad-CAM attention maps because of pooling layers in the architecture. Consequently, often there exists no bijective mapping between both attention maps which complicates the computations of the attention map correlation function h. Moreover, Grad-CAM and Guided Backpropagation tend to focus on semantically different regions of the input, such that a simple resizing operation would still incorrectly represent the correlation between the attention maps.

To alleviate those issues, we propose to use a masking strategy. First, we compute Grad-CAM $A_{Grad-CAM,1}$ and Guided Backpropagation A_{GB} attention maps. Then, a mask P is derived from Guided Backpropagation to mask the input. The mask is computed according to the spatial attention masking defined by Wang et al. [28]:

$$P(i) = \frac{1}{1 + \exp(-(A_{GB}(i) - \mu)/\sigma)}, \tag{6}$$

where μ is the mean over A_{GB}, σ the variance, and i spans the A_{GB}. Subsequently, the forward propagation is run a second time using the masked input $P \odot x$. We compute Grad-CAM $A_{Grad-CAM,2}$ a second time using the feature maps of this second forward propagation. Eventually, we compute the correlation between the two instances of Grad-CAM attention maps by vectorizing the attention maps and computing the Pearson correlation $L_A(\theta, x) = Pearson(A_{Grad-CAM,1}, A_{Grad-CAM,2})$. Correlating the two Grad-CAM maps $A_{Grad-CAM,1}$ and $A_{Grad-CAM,2}$ indirectly correlates the original Grad-CAM map $A_{Grad-CAM,1}$ to the Guided Backpropagation map A_{GB} because the masking with A_{GB} forces $A_{Grad-CAM,2}$ to highlight regions already highlighted by A_{GB} itself.

This masking method was inspired by perturbation methods for attention map computation [18], and can be generalized to compute attention map consistency with any type of attention map. Algorithm 1 summarizes the masking and computation of the attention consistency loss function.

Algorithm 1 Proposed attention consistency for Grad-CAM and Guided Backpropagation

```
Input: sample $x$, convolutional neural network $f$
Output: attention consistency loss $L_A$
1: Forward propagate $x$ through $f$
2: $A_{Grad-CAM,1} \leftarrow$ Grad-CAM (Equation 3)
3: $A_{GB} \leftarrow$ Guided Backpropagation (Equation 5)
4: $P \leftarrow$ mask computed from $A_{GB}$ (Equation 6)
5: $x_{masked} \leftarrow P \odot x$
6: Forward propagate $x_{masked}$ through $f$
7: $A_{Grad-CAM,2} \leftarrow$ Grad-CAM computed from $x_{masked}$ (Equation 3)
8: $L_A \leftarrow Pearson(A_{Grad-CAM,1}, A_{Grad-CAM,2})$
```

5. Experiments

We show that the proposed method can improve classification performance, while improving the representations of the classifier. We show that the performance gain varies with the training set size, and we show the benefits of the method for three architecture: ResNet 50 [12], Inception-v3 [24], and 3D 18 layers ResNet [26]. The proposed method is evaluated on a real-world dataset, which we curated for this project: event recognition in continuous recordings of hospital patients. Sanity checks are performed on the PASCAL-VOC dataset [8]. We also present an ablation study where we compare unsupervised fine-tuning, to combined and alternated optimization, and where attention consistency with Grad-CAM and Guided Backpropagation is computed using four different resolution matching strategies and three different correlation measures.

5.1. Datasets

We aggregated and curated a dataset of continuous video recordings of hospital patients in an epilepsy center unit for children and neonates. We identified video clips displaying five types of events: patting of the neonates by nurses, suctioning of neonates' mouth liquid by nurses, rocking of neonates, patient chewing food, and finally cares being done on the patient by nurses. Those events are selected as they can mislead automated seizure detection systems. Clips are post-processed to be 4 sec long, sampled at 4 frames per second. Consequently, the task is defined as a five-classes video clip classification problem. The curated dataset includes 59 patients and 2 hours 18 minutes of video recordings, with a median event clip length of 25 seconds. The frame resolution is reduced from 320x240 to 80x80 using subsampling. The video clips were separated into three balanced sets of similar size, using data of different patients for each sets. The first set has 32 4-sec clips per class, the second set 49 and the last set 35. More statistics about the dataset can be found in supplementary materials.

PASCAL VOC [8] is used for sanity check and is reframed as a 20 classes multi-label classification dataset. The bounding box ground truth annotations are not used for training. They are only used during inference to evaluate the weakly supervised detection capability of the generated attention maps. 500 random images sampled from the 5717 images of the training set are used a validation set, and a varying number of images, from 2 to 16 images per class, are sampled from the remaining set and used for training. The publicly released validation set of 5823 images is left out during training, and used a separate test set to evaluate the methods.

5.2. Models and Training

For the experiements on hospital videos, we use a 3D 18 layers ResNet [26] pretrained on Kinetics-400 [14]. For the experiments on PASCAL, we use state-of-the-art 2D networks: ResNet 50 [12] and Inception-v3 [24] (only for PASCAL), both pretrained on ImageNet [6].

On top of ImageNet or Kinetics pretraining, we train on PASCAL or the video dataset using N samples per class and a categorical cross-entropy loss function and the Adam optimizer with a learning rate of 0.001. Then, we perform unsupervised fine-tuning with the proposed method using the same training samples. The batch size is 4 for PASCAL, and 12 for the video dataset. We select the model that maximizes the mean average precision on PASCAL (because it

Table 1. **Classification results on the video dataset.** We compare the baseline method, the baseline fine-tuned with the proposed attention consistency (B + Attention), SimCLR pretraining [4] (SimCLR), and SimCLR pretraining fine-tuned with the proposed attention consistency (S + Attention). The first four rows indicate models trained with 16 samples per class, and the last four rows with 32 samples per class. We show F1 scores rescaled to [0,100]. Mean F1 is the F1 averaged over the five classes. F1 suctioning, chewing, rocking, cares and patting show class-wise F1s. Bootstrapped confidence intervals are indicated in brackets. The highest performance is indicated in bold.

Method	Mean F1	F1 suctioning	F1 chewing	F1 rocking	F1 cares	F1 patting
Baseline – 16 samples	15.5 [11.6-19.4]	37.9 [25.0-49.3]	0.0 [0.0-0.0]	0.0 [0.0-0.0]	**30.7 [21.7-39.7]**	**9.3 [0.0-20.0]**
B + Attention	17.4 [13.3-21.9]	42.1 [31.2-52.2]	8.2 [0.0-16.2]	16.4 [4.8-29.0]	20.2 [11.6-28.6]	0.0 [0.0-0.0]
SimCLR	23.1 [18.8-27.7]	47.1 [35.1-58.3]	27.6 [16.3-38.9]	17.1 [6.6-29.2]	23.8 [14.7-33.3]	0.0 [0.0-0.0]
S + Attention	**29.7 [25.2-34.3]**	**50.4 [37.5-62.2]**	**40.7 [29.4-51.4]**	**43.2 [32.5-53.9]**	14.5 [5.5-24.3]	0.0 [0.0-0.0]
Baseline – 32 samples	26.5 [21.6-31.7]	30.1 [19.7-40.4]	36.4 [24.7-46.9]	9.3 [0.0-20.5]	44.0 [33.0-54.5]	13.0 [0.0-25.0]
B + Attention	27.5 [22.7-32.8]	**33.6 [23.5-43.0]**	**46.7 [35.9-57.1]**	5.4 [0.0-15.8]	29.2 [16.7-41.0]	**23.2 [10.9-35.7]**
SimCLR	29.3 [23.8-34.8]	21.3 [10.5-32.0]	43.9 [30.2-56.5]	**30.9 [18.2-43.5]**	42.0 [31.2-51.9]	7.7 [0.0-16.7]
S + Attention	**31.4 [25.9-37.0]**	31.2 [17.2-43.3]	35.4 [22.9-47.5]	30.5 [17.5-42.9]	**48.0 [38.3-57.6]**	12.2 [3.6-23.3]

Table 2. **Classification Results on PASCAL.** We compare the classification results of ResNet without data augmentation (ResNet no Aug), ResNet with data augmentation (ResNet), ResNet with data augmentation (ResNet) and layer attention consistency (ResNet + Layer Att), and ResNet with data augmentation and the proposed attention consistency fine-tuning (ResNet + Proposed Att (ours)). Results are shown for models trained with varying numbers of training samples per class. Stars indicate significant differences.

Methods	F1, Training Sample per Class						mAP, Training Sample per Class					
	2	4	8	12	16	135	2	4	8	12	16	135
ResNet no Aug [12]	23.9	43.1	57.4	59.9	64.7	76.9	51.4	65	73.2	76.9	78.6	85.2
ResNet [12]	38.3	46.9	60.2	64.1	68.3	**77.6**	57.4	67.2	74.4	77.5	79.4	**85.7**
ResNet + Layer Att [29]	37.8	48.8	60.4	64.1	**68.6**	**77.6**	57.2	67.0	74.4	**77.6**	**79.5**	**85.7**
ResNet + Proposed Att (ours)	**41.2***	**51.6***	**62.5***	**64.8**	68.2	77.3	**58.1**	**67.9**	**74.9**	77.2	**79.4**	**85.7**

is multi-label) and the mean F1 score on the video dataset, computed over the validation set. The test is completely left out of the training loop for final evaluation.

The models are regularized with data augmentation. For the video dataset, data augmentation includes random color jitters with brightness, contrast, and saturation up to 0.8, and hue up to 0.4, random crops of 90 percent of the image size in x and y, and random horizontal flips. For PASCAL, data augmentation includes random rotations up to 10 degrees, 50 percent chance of horizontal or vertical flip, and resizing to 256x256 for ResNet or to 299x299 for Inception-v3 due to architecture constraints.

Our code is available online on GitHub at *removed for blind review*, and builds on PyTorch 1.9.0 and Torchvision 0.10.0. We used two NVIDIA Titan RTX for training.

5.3. Classification Results

On the video dataset, we compare the proposed method to training a 3D 18 layers ResNet without attention consistency. We also include a comparison to SimCLR pretraining [4] using the same training samples. All training procedures use the same training samples. We vary the number of training samples per class from 16 to 32. Table 1 shows the F1 scores. The proposed attention consistency method provides a lift of mean F1. This lift increases when SimCLR is used for pretraining.

To verify that these results generalize to other datasets, we repeat these experiments on PASCAL. We compare the proposed method to training ResNet 50 without attention consistency. We also compare to not using data augmentation during this training stage. Lastly, we compare the proposed attention consistency method to the layer attention consistency method proposed by Wang et al. [29]. As recommended by the authors, we use the last convolution layers of the last two blocks in ResNet to compute the attention consistency. The layer attention consistency is implemented as a fine-tuning step, which shows better results than simultaneous training with categorical cross-entropy. We repeat the experiments with varying numbers of training samples from 2 to 135 (the maximum possible) per class (2, 4, 8, 12, 16 and 135). The F1 scores and mean average precision on the test are shown in Table 2. Experiments on PASCAL confirm the findings of the curated video dataset, and show that fine-tuning can improve classification performance when training data is scarce. The proposed attention consistency method outperforms the layer attention consistency baseline, and combining both methods did show higher results than the proposed attention consistency method alone.

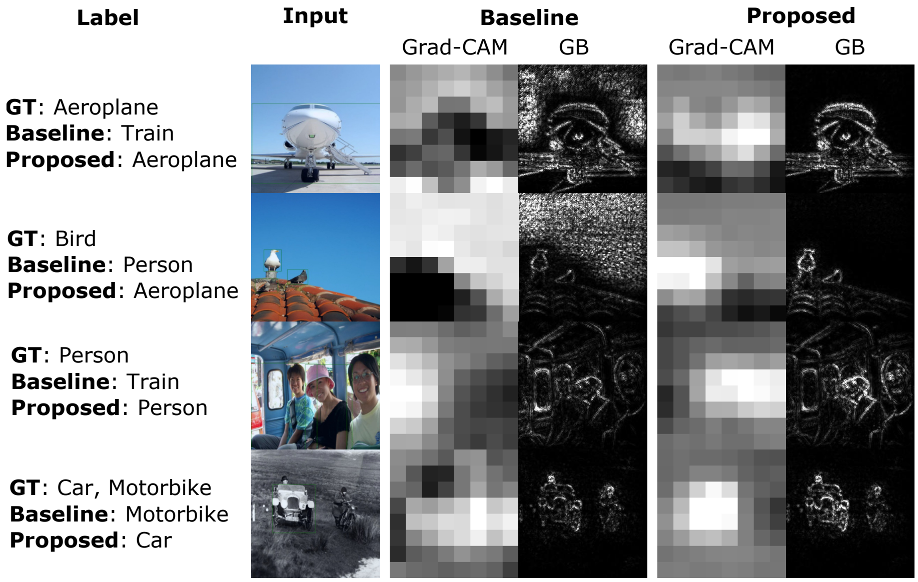

Figure 2. **Examples of attention maps for PASCAL.** Comparison between the baseline ResNet with data augmentation (Baseline) and the same method plus the proposed unsupervised attention consistency fine tuning (Proposed). The first column indicates the ground truth (GT) label, the label predicted by the baseline, and by the proposed method. The second column shows the input frames. The two middle columns show Grad-CAM and Guided Backpropagation (GB) for the baseline, and the last two columns show the same attention maps for the proposed method. Models were training with 4 samples per class. Examples with clear differences between methods were selected for display.

Table 3. **Overlap on PASCAL.** Intersection over Union between Grad-CAM attention maps and ground truth bounding boxes. The overlap is computed only for true positives classifications. Results are shown for Inception with data augmentation (Inception), Inception with data augmentation and the proposed attention consistency fine-tuning (Inception + Attention), ResNet without data augmentation (ResNet no Aug), ResNet with data augmentation (ResNet), ResNet without data augmentation (ResNet no Aug), ResNet with data augmentation and layer attention consistency (ResNet + Layer Att), and ResNet with data augmentation and the proposed attention consistency fine-tuning (ResNet + Attention). Each column corresponds to a model trained with a varying number of samples per class. The highest performance is indicated in bold. Stars indicate significant statistical improvement.

Method	Training Samples per Class					
	2	4	8	12	16	135
Inception [24]		44.5	44.1	44.5	43.9	**43.3**
Inc + Att (ours)		**44.6**	**44.2**	**44.9**	**44.6**	**43.3**
ResNet no Aug [12]	42.8	46.6	47.7	47.6	48.1	**45.2**
ResNet [12]	44.6	50.0	48.5	49.0	47.9	44.9
ResNet + Layer Att [29]	43.2	47.8	47.1	47.3	46.9	43.6
ResNet + Prop Att (ours)	**46.4***	**50.2***	**49.5**	**49.3**	**48.4**	45.1

5.4. Learned Representations

We consider the generated attention maps as a measure of the correctness of the representations. If the attention maps highlight the target objects in the image, we expect the model to have learned correct representations. We visualize attention maps of both the video dataset and PASCAL in Figures 1 and 2. We also plot the gain of attention consistency with respect to the original attention consistency on PASCAL (Supplementary Material).

There were no bounding box annotations to quantify the improvement of the attention maps localization on the video dataset. Instead, we perform the analysis on PASCAL. Using the intersection of union, we compute the overlap of the ground truth bounding boxes with Grad-CAM attention maps rescaled in $[0, 1]$ and thresholded at 0.5. We compute the overlap only for true positive image classifications (the number of true positive images increases with the classification performance and the number of training samples per class). Table 3 shows the overlap for varying number of training samples for ResNet, Inception-v3, and for the layer attention consistency baseline [29]. The overlap is always higher with the proposed method, for all training sample sizes and both network architectures. We also notice that forcing layer attention consistency [29] reduces the localization capacity of the model. Contrary to the proposed method, with layer attention consistency [29], the performance gain observed in Table 2 for smaller datasets is probably due to the induced smaller variance of the network weights than more accurate attention maps.

Qualitative inspection of the attention maps on PASCAL (Figure 2) and the video dataset (Figure 1) also reveal that attention consistency can improve the attention maps to focus on the target subject, while still predicting the incorrect label. Such an improvement has not been accounted for in our quantitative measure of the overlap (Table 3) as only true positives were utilized.

Note that attention consistency displayed smaller improvements on Inception-v3 than on ResNet (Table 3). This may indicate that Inception-v3 directly learns better representations. Inception was designed to be computationally efficient, with a reduced number of parameters. This may explain why the network seems to learn more generalizable representations [24].

6. Ablation Study

To justify our choice of attention consistency loss function, we present an ablation study varying its two main components: the resolution matching technique and the correlation metric used to estimate attention map overlap. We consider that an attention consistency loss function is good if a lower consistency loss is equivalent to a lower supervised classification loss–the cross-entropy in our experiments. To quantify this relationship, we measure the correlation between both the unsupervised consistency loss and the supervised classification loss during a fully supervised training procedure. In this experiment, only the fully supervised classification loss is used to update the network weights. The unsupervised consistency loss is only monitored.

We explore three correlation measures: Structural Similarity Index Measure (SSIM), cross-correlation and Pearson coefficient; and four resolution matching techniques: the proposed masking technique using Guided Backpropagation to create the mask, the proposed masking technique using Grad-CAM to create the mask, smoothing and upsampling of Grad-CAM with linear interpolation, and downsampling of Guided Backpropagation with max pooling.

Table 4 shows the results of the ablation study on PASCAL. The best combination of resolution matching strategy and correlation measure was Pearson and Guided Backpropagation as a mask. The second best was SSIM and Guided Backpropagation as a mask. Using Grad-CAM upsampling or Guided Backpropagation pooling for resolution matching was worse than Guided Backpropagation masking but still gave satisfying results, independently of the correlation metrics. Using Grad-CAM as a mask for resolution matching should be avoided.

We also compare the three different training strategies: optimizing the consistency loss as an unsupervised fine-tuning step (fine-tuning), together as a linear combination with the supervised loss (combined), and alternating between both losses batch-wise during training (alternated). Experiments are done on PASCAL with ResNet 50 and 4 training samples per class. The combined approach gets an attention map overlap of 49.3% IoU and a F1 of 48.9%. The alternated one 49.5% IoU and a F1 of 50.6%. As reported earlier, the fine-tuning approach got an 50.2% IoU and a F1 of 51.6%.

Table 4. **Ablation study showing the Pearson correlation between different attention consistency losses and the target supervised loss.** High correlations show that the attention consistency loss estimates the target supervised loss well. The training was realized with a supervised ResNet for 100 epochs on PASCAL, and the supervised cross-entropy loss was computed on the validation set. 135 samples per class, the maximum in our classification dataset, was used for this experiment. The rows indicate the resolution matching techniques, and the columns the correlation measures.

	Pearson	Cross-correlation	SSIM
Grad-CAM Upsampling	65.4	67.5	60.3
GB Pooling	65.5	64.1	68.8
GB as mask	82.6	26.7	80.6
Grad-CAM as mask	-51.9	-67.2	-75.8

6.1. Limitations

Improving attention consistency was challenging for samples whose original attention maps did not overlap. Introducing a stochastic process in the computation of the attention consistency may allow attention maps to overlap and help the network converge to the target solution.

We refer to our training strategy as unsupervised fine-tuning, but we still monitor the generalization performance using the labels of the validation set. Knowing when to stop training without utilizing any labels could have substantial practical implications, where networks can be fine-tuned for attention consistency on deployment dataset without requiring any labels.

We first evaluated the layer attention consistency baseline [29] in PASCAL and did not observe quantitative improvements in the classification performance nor in attention maps quality, even combined with the proposed attention consistency method. While we may likely observe similar results in the video dataset, we have not performed the analysis.

7. Conclusions

We proposed a method to optimize the consistency of attention maps, and propose an implementation for Grad-CAM and Guided Backpropagation. We showed on our own video dataset and on PASCAL that the method can improve both classification performance and the quality of attention maps. The proposed method contributes to research in network interpretability and application in small datasets.

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. 1, 2

[2] Naomi Altman and Martin Krzywinski. The curse (s) of dimensionality. *Nat Methods*, 15(6):399–400, 2018. 1

[3] Po-Hsuan Cameron Chen, Yun Liu, and Lily Peng. How to develop machine learning models for healthcare. *Nature materials*, 18(5):410–414, 2019. 1, 4

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2, 6

[5] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016. 3

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[7] Florian Dubost, Hieab Adams, Pinar Yilmaz, Gerda Bortsova, Gijs van Tulder, M Arfan Ikram, Wiro Niessen, Meike W Vernooij, and Marleen de Bruijne. Weakly supervised object detection with 2d and 3d regression neural networks. *Medical Image Analysis*, 65:101767, 2020. 1, 2, 3

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010. 1, 5

[9] 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. 2

[10] Hao Guo, Kang Zheng, Xiaochuan Fan, Hongkai Yu, and Song Wang. Visual attention consistency under image transforms for multi-label image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 729–739, 2019. 2

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 2

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5, 6, 7

[13] Longlong Jing, Toufiq Parag, Zhe Wu, Yingli Tian, and Hongcheng Wang. Videossl: Semi-supervised learning for video classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1110–1119, 2021. 3

[14] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 5

[15] Hugo J Kuijf, J Matthijs Biesbroek, Jeroen De Bresser, Rutger Heinen, Simon Andermatt, Mariana Bento, Matt Berseth, Mikhail Belyaev, M Jorge Cardoso, Adria Casamitjana, et al. Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge. *IEEE transactions on medical imaging*, 38(11):2556–2568, 2019. 3

[16] Yang Li, Shichao Kan, and Zhihai He. Unsupervised deep metric learning with transformed attention consistency and contrastive clustering loss. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 141–157. Springer, 2020. 2

[17] Yubin Park and Joyce Ho. Tackling overfitting in boosting for noisy healthcare data. *IEEE Transactions on Knowledge and Data Engineering*, 2019. 1

[18] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *British Machine Vision Conference*, 2018. 2, 5

[19] 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. 1, 2, 4

[20] 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. 1, 2

[21] Jake Snell, Kevin Swersky, and Richard S Zemel. Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175*, 2017. 3

[22] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020. 3

[23] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. In *International Conference for Learning Representations*, 2015. 1, 2, 4

[24] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 5, 7, 8

[25] Kimberley M Timmins, Irene C van der Schaaf, Edwin Bennink, Ynte M Ruigrok, Xingle An, Michael Baumgartner, Pascal Bourdon, Riccardo De Feo, Tommaso Di Noto, Florian Dubost, et al. Comparing methods of detecting and segmenting unruptured intracranial aneurysms on tof-mras: The adam challenge. *NeuroImage*, page 118216, 2021. 3

[26] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal

convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 5

[27] 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. 2

[28] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2017. 5

[29] Lezi Wang, Ziyan Wu, Srikrishna Karanam, Kuan-Chuan Peng, Rajat Vikram Singh, Bo Liu, and Dimitris N Metaxas. Sharpen focus: Learning with attention separability and consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 512–521, 2019. 2, 6, 7, 8

[30] Chunming Xu and Scott A Jackson. Machine learning and complex biological data, 2019. 1

[31] Haotian Xu, Xiaobo Jin, Qiufeng Wang, and Kaizhu Huang. Multi-scale attention consistency for multi-label image classification. In *International Conference on Neural Information Processing*, pages 815–823. Springer, 2020. 2

[32] Quanshi Zhang and Song-Chun Zhu. Visual interpretability for deep learning: a survey. *arXiv preprint arXiv:1802.00614*, 2018. 1

[33] 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. 2