

# Learning to ignore: rethinking attention in CNNs

Firas Laakom\*, Kateryna Chumachenko\*, Jenni Raitoharju, Alexandros Iosifidis, and Moncef Gabbouj

**Abstract**—Recently, there has been an increasing interest in applying attention mechanisms in Convolutional Neural Networks (CNNs) to solve computer vision tasks. Most of these methods learn to explicitly identify and highlight relevant parts of the scene and pass the attended image to further layers of the network. In this paper, we argue that such an approach might not be optimal. Arguably, explicitly learning which parts of the image are relevant is typically harder than learning which parts of the image are less relevant and, thus, should be ignored. In fact, in vision domain, there are many easy-to-identify patterns of irrelevant features. For example, image regions close to the borders are less likely to contain useful information for a classification task. Based on this idea, we propose to reformulate the attention mechanism in CNNs to learn to ignore instead of learning to attend. Specifically, we propose to explicitly learn irrelevant information in the scene and suppress it in the produced representation, keeping only important attributes. This implicit attention scheme can be incorporated into any existing attention mechanism. In this work, we validate this idea using two recent attention methods Squeeze and Excitation (SE) block and Convolutional Block Attention Module (CBAM). Experimental results on different datasets and model architectures show that learning to ignore, i.e., implicit attention, yields superior performance compared to the standard approaches.

**Index Terms**—Computer vision, CNNs, attention mechanisms, CBAM, SE



## 1 INTRODUCTION

INSPIRED by the properties of the human visual system, attention mechanisms have been recently applied in the field of deep learning, resulting in improved performance of the existing models across multiple applications. In the context of computer vision, learning to attend, i.e., learning to highlight and emphasize relevant attributes of images, have led to development of novel approaches [1], [2] in Convolutional Neural Networks (CNNs), improving their capabilities in many tasks [3], [4], [5].

Related to the concept of attention, recent studies in neuroscience suggest that the ability of humans to successfully perform visual tasks is related to the ability to ignore and suppress distractive information [6], [7], [8]. For example, the authors of [7] show that differences in visual working memory capacity, i.e., ability to remember visual features of multiple objects, are specifically related to distractor-suppression activity in visual cortex. This idea is reinforced in [8], where the authors provide evidence on an inhibitory mechanism of suppression of salient distractors aimed at preventing them from capturing attention and being further processed by humans. Additional studies [9] report that ignoring the irrelevant information is a powerful learning tool for human cognition with ubiquitous effectiveness. Inspired by these findings, we investigate the intuition of learning to explicitly ignore irrelevant information in the field of computer vision and reformulate attention mechanisms commonly utilized in CNNs under the framework of learning to ignore rather than learning to attend.

Existing attention mechanisms used in CNNs learn the attention masks by directly optimizing for the high re-

sponse of attributes of the image that are important for the prediction and, thus, should be focused on more. The learned attention masks are applied to feature representations, leading to higher emphasis put on the attributes of interest, and, therefore, resulting in implicit ignorance of the irrelevant features. In our work, we propose to rethink this logic and instead explicitly focus on ignoring irrelevant regions, hence achieving the attention to important regions implicitly. We argue that learning of features that should be ignored is an easier task than learning to attend and, therefore, optimization with such an objective leads to better training. Arguably, discriminative features of samples of different classes are harder to capture and often require more advanced feature learning. On the other hand, irrelevant attributes or attributes common between classes are often related to easy-to-identify patterns, such as borderline locations on the image or background features that can already be learned at early stages of training. Following this intuition, we design our method to explicitly optimize which attributes of the image should be ignored, and based on this, the important attributes that should be attended are derived implicitly. We validate this idea using two recent attention methods Squeeze and Excitation (SE) block and Convolutional Block Attention Module (CBAM) and show that indeed our intuition holds and explicitly learning features to ignore leads to better model performance.

Our contributions can be summarized as follows:

- We propose a new perspective on attention in computer vision where the main aim is to learn to ignore instead of learning to attend.
- We propose an implicit attention scheme which explicitly learns to identify the irrelevant parts of the scene and suppress them. The proposed approach can be incorporated into any existing attention mechanism.
- We validate this idea using two attention mech-

\* Equal contribution

F. Laakom, K. Chumachenko and M. Gabbouj are with Department of Computing Sciences, Tampere University, Tampere, Finland, Tampere University, Tampere, Finland.

J. Raitoharju is with the Programme for Environmental Information, Finnish Environment Institute, Jyväskylä, Finland.

A. Iosifidis is with the Department of Electrical and Computer Engineering, Aarhus University, Aarhus, Denmark.

anisms. Specifically, we reformulate Squeeze-and-Excitation (SE) block and Convolutional Block Attention Module (CBAM) using our paradigm, i.e., *learn to ignore*, and show the superiority of such an approach.

## 2 RELATED WORK

**Attention mechanisms in vision.** The idea of attention in vision tasks stems from the properties of selective focus in the human visual system, i.e., that humans do not perceive images as a whole, but rely on certain salient parts of them. This property gave rise to a variety of attention-based learning mechanisms aimed to enhance the performance in computer vision domain [3], [4], [10], finding its applications in a variety of tasks, including sequence learning [11], image captioning [5], and others [12], [13]. A subset of attention-driven methods is directed at CNNs and aims at selecting and highlighting relevant attributes in the feature space during training [1], [2]. Conventionally, this is achieved by learning attention masks over feature representations that encode the importance of different attributes in form of weights and applying these masks on intermediate feature representations. This results in higher influence of features relevant for decision making in subsequent layers.

Other tasks adjacent to this line of research include saliency estimation, image segmentation, and weakly-supervised object localization. In saliency estimation, the goal is to estimate salient, i.e., significant regions of the scene without any prior knowledge on the scene in unsupervised [14], [15] or supervised manner [16], [17], [18]. In image segmentation, the task is to partition a given image into a set of segments, based on either semantics (semantic segmentation) or individual objects (instance segmentation) [19]. In weakly-supervised object localization, the goal is to predict the location of the object given only image-level labels [20].

Within the attention mechanisms utilized in CNNs, two of the notable ones include Squeeze-and-Excitation block (SE) [1] and Convolutional Block Attention Module (CBAM) [2]. In SE, an attention mask is learned channel-wise based on global average-pooled features of intermediate representations and applied at multiple layers of the ResNet architecture [21]. A further extension is the CBAM method that enriches the SE mechanism by additional max-pooled input and learns spatial attention in addition to channel-wise one. The learned attention weight masks are then applied channel-wise or pixel-wise to corresponding feature maps. These methods were shown to lead to superior performance across various domains and can be incorporated in any CNN architecture.

**Learning by ignoring.** Learning by ignoring is a powerful learning paradigm, which has been used in various machine learning applications [22], [23], [24]. It has been leveraged in the context of saliency estimation [14], [23], [25], [26]. For example, the authors of [14] propose an unsupervised graph-based saliency estimation approach, where auxiliary variables are used to encode prior knowledge on regions to be ignored, such as dark regions, as it is assumed that they are less-likely to contain salient object. A similar approach was proposed for the color constancy problem

[27]. In the context of machine translation, it has been shown that learning to ignore spurious correlations in the data can improve the performance of neural networks in zero-shot translation [22]. In the context of domain adaptation, a learning framework assigning and learning an ‘ignoring’ score for each training sample and re-weighting the total loss based on these scores was proposed in [24].

## 3 LEARNING TO IGNORE IN CNNs

Attention in CNNs is generally formulated in a form of a learned attention mask that emphasizes relevant information in a feature map. Formally, given a feature map  $\mathbf{F}$ , attention can be defined as follows:

$$\mathbf{F}' = \mathbf{F} \otimes f_{\theta}(\mathbf{F}), \quad (1)$$

where  $\mathbf{F}'$  is the attended feature map output,  $\otimes$  is the element-wise multiplication and  $f_{\theta}(\cdot)$  is an attention function with learnable parameters  $\theta$ , which takes as input a feature map  $\mathbf{F}$  and returns an attention mask  $f_{\theta}(\mathbf{F}) \in [0, 1]$ . This mask is then element-wise multiplied with the original map  $\mathbf{F}$  in order to produce the output map  $\mathbf{F}'$ . The mask  $f_{\theta}(\mathbf{F})$  is expected to identify relevant spatial or channel information and output the ‘importance score’ for each attribute, producing high response for most relevant regions and smaller values for regions of lesser interest. This can be seen as an explicit attention mechanism, where the model  $f_{\theta}(\cdot)$  learns to directly identify and highlight relevant information.

In this work, we develop a new formulation of the concept of attention in CNNs, where the main target is learning to ignore instead of learning to attend. By training the model to predict irrelevance of features, rather than their importance, we expect to simplify the training objective and, hence, to improve the learning of the model. Our approach consists of a function which learns to identify irrelevant or confusing parts of the feature map in order to suppress them, followed by inversion of predicted irrelevance scores. Formally, this can be formulated as follows:

$$\mathbf{F}' = \mathbf{F} \otimes T(g_{\theta}(\mathbf{F})), \quad (2)$$

where  $g_{\theta}(\cdot)$  is a function with learned parameters  $\theta$  that is expected to learn to highlight information in the feature map that is irrelevant or confusing for the prediction. This can be seen as an *ignoring mask* that outputs high values for attributes and regions that should be suppressed in the feature map. The function  $T(\cdot)$  is a function flipping with an output  $T(x)$  inversely proportional to  $x$ , hence flipping the learned ignoring mask and transforming it into an attention mask. Similarly to Eq. (1), the final feature map  $\mathbf{F}'$  is obtained by element-wise multiplication of the input map  $\mathbf{F}$  and the flipped ignoring mask  $T(g_{\theta}(\mathbf{F}))$ .

Given an ignoring mask  $g_{\theta}(\mathbf{F})$ , the function  $T(\cdot)$  can be any function satisfying the condition of being inversely proportional to its input and bounded between  $[0, 1]$ . In this work, we propose three variants:

$$T_1(x) = 1 - \alpha x, \quad (3)$$

$$T_2(x) = \text{sigmoid}\left(\frac{1}{x}\right), \quad (4)$$

$$T_3(x) = \text{sigmoid}(-x). \quad (5)$$

The first variant  $T_1(\cdot)$  linearly converts the ignoring mask to an attention one, and  $\alpha$  is a hyper-parameter controlling this linear scaling. The extreme case  $\alpha = 0$  corresponds to the extreme case  $\mathbf{F}' = \mathbf{F}$ , i.e., none of the features are emphasized or suppressed. For the second and third variants  $T_2$  and  $T_3$ , a sigmoid function is applied to ensure that the output is bounded between  $[0, 1]$ .

We argue that formulating the objective as learning of irrelevant features that should be ignored, as opposed to focusing on important features, is beneficial, as optimization of a model with such an objective is easier. This is due to potential presence of many easy-to-identify patterns of irrelevant attributes, such as borderline pixel locations, color and lighting perturbations, or background properties that are not correlated with the groundtruth labels. At the same time, information responsible for predictions is generally label-specific and harder to capture. Moreover, learning of discriminative attributes that can be regarded as important often requires learning of complex feature representations that can be achieved only at latter stages of training, while patterns irrelevant for decision making can often be identified already at the early stages.

It can be argued that standard attention, i.e., Eq. (1), is also learning to ignore as it is expected to indirectly assign smaller values for less important regions. However, function  $f_\theta(\cdot)$  is optimized directly for highlighting relevant information and, hence, this can be seen as an implicit and indirect strategy of learning to ignore. In our approach, Eq. (2), the model  $g_\theta(\cdot)$  is explicitly optimized for identifying the irrelevant or confusing parts and the function  $T(\cdot)$  suppresses them. This can be seen as an implicit learning to attend approach and explicit learning to ignore approach, as opposed to the standard attention which has an explicit learning to attend formulation.

As can be seen, the main difference between implicit and explicit attention formulations is the presence of a flipping function  $T(\cdot)$ . It can be seen from Eq. (1) and Eq. (2) that  $f_\theta(\cdot)$  can be directly replaced by  $T(g_\theta(\cdot))$ . This makes it straightforward to reformulate any existing explicit attention method to learn to ignore instead of learning to attend by applying an inversion function  $T(\cdot)$  on top of the learned mask. This way, the model  $g_\theta(\cdot)$  can be learned as the model  $f_\theta(\cdot)$  in conventional attention methods, while its parameters will be optimized to detect irrelevant or confusing regions instead of relevant ones. In this paper, for the choice of the function  $f_\theta(\cdot)$ , we consider two state-of-the-art attention mechanisms, namely SE [1] and CBAM [2], and we show how to reformulate them using our paradigm in the following subsections.

### 3.1 Ignoring with Squeeze-and-Excitation blocks

Squeeze-and-Excitation (SE) block [1] presents a mechanism to learn channel-wise attention, focusing on which features of the representation are important for prediction. This is achieved by squeezing the spatial information into a channel representation, followed by an excitation operation that highlights important channels via a bottleneck block. Formally, given a feature map  $\mathbf{F}$ , this is defined as follows:

$$f_\theta(\mathbf{F}) = \sigma(\mathbf{W}_2\delta(\mathbf{W}_1GAP(\mathbf{F}))), \quad (6)$$

where  $GAP(\cdot)$  denotes Global Average Pooling,  $\delta$  is a ReLU activation,  $\sigma$  is the sigmoid function,  $\mathbf{W}_1 \in \mathbb{R}^{c \times \frac{c}{r}}$  and  $\mathbf{W}_2 \in \mathbb{R}^{\frac{c}{r} \times c}$  are linear layers,  $c$  is the number of channels in  $\mathbf{F}$ , and  $r$  is the reduction rate in the bottleneck block. Given the output  $f_\theta(\mathbf{F})$ , the attended feature map is obtained by applying the learned mask element-wise between corresponding channels.

To incorporate our ignoring paradigm into SE, we apply  $T(\cdot)$  to the output  $f_\theta(\mathbf{F})$ , hence transforming its objective into learning features that should be ignored. Specifically, we define the three variants as:  $f_\theta^1(\mathbf{F}) = 1 - \alpha\sigma(\mathbf{W}_2\delta(\mathbf{W}_1GAP(\mathbf{F})))$ ;  $f_\theta^2(\mathbf{F}) = \sigma(\frac{1}{\sigma(\mathbf{W}_2\delta(\mathbf{W}_1GAP(\mathbf{F})))})$ ;  $f_\theta^3(\mathbf{F}) = \sigma(-\mathbf{W}_2\delta(\mathbf{W}_1GAP(\mathbf{F})))$  using the definitions of  $T_1$ ,  $T_2$ , and  $T_3$ , respectively. As can be noticed, in the first two variants  $T(\cdot)$  is applied directly on  $f_\theta(\mathbf{F})$ , while in the third case it is applied on pre-sigmoid output to ensure sufficiently wide range for attention scores.

### 3.2 Ignoring with Convolutional Block Attention Modules

Following the approach of SE, Convolutional Block Attention Module (CBAM) [2] extends it to incorporate spatial attention as well as to enrich channel attention with an additional input representation. Under the definition of attention in Eq. (1), this is formulated as follows:

$$\begin{aligned} f^{ch}(\mathbf{F}) &= \sigma(\mathbf{W}_2\delta(\mathbf{W}_1(GAP(\mathbf{F}))) + \mathbf{W}_2\delta(\mathbf{W}_1(GMP(\mathbf{F})))), \\ \mathbf{F}^{ch} &= \mathbf{F} \otimes f^{ch}(\mathbf{F}), \\ f^{sp}(\mathbf{F}^{ch}) &= \sigma(Conv^{7 \times 7}(GAP(\mathbf{F}^{ch}) \frown GMP(\mathbf{F}^{ch}))), \end{aligned} \quad (7)$$

where  $f^{ch}$  and  $f^{sp}$  denote channel and spatial attention, respectively,  $GAP(\cdot)$  and  $GMP(\cdot)$  correspond to Global Average Pooling and Global Max Pooling, respectively,  $\delta$  is a ReLU activation,  $\sigma$  is the sigmoid activation,  $\mathbf{W}_1 \in \mathbb{R}^{c \times \frac{c}{r}}$  and  $\mathbf{W}_2 \in \mathbb{R}^{\frac{c}{r} \times c}$  are linear layers,  $c$  is the number of channels in  $\mathbf{F}$ , and  $r$  is the reduction rate in the bottleneck block, similarly to SE.  $\mathbf{F}^{ch}$  is the channel-wise attended feature map,  $Conv^{7 \times 7}$  denotes a convolutional layer with  $7 \times 7$  kernel, and  $\frown$  denotes concatenation.

As can be seen, channel and spatial attention masks are applied sequentially and channel-attended feature representations are used as input to compute spatial attention. Following this, we transform CBAM for ignoring by addition of inversion function  $T(\cdot)$  on top of both channel function  $f^{ch}(\cdot)$  and spatial function  $f^{sp}(\cdot)$  to reformulate their objectives as learning of features and regions to ignore. In both cases, variants of  $T_1(\cdot)$  and  $T_2(\cdot)$  are applied directly on the output of corresponding functions, and  $T_3(\cdot)$  is applied on pre-sigmoid output.

## 4 EXPERIMENTAL RESULTS

### 4.1 CIFAR10 & CIFAR100

We start by validating our approach on image classification task using CIFAR10 and CIFAR100 [28] datasets. To show invariance of the proposed approach to specific model architectures, we evaluate two state-of-the-art CNNs, namely,

ResNet50 [21] and DenseNet [29] architectures. We report the results of standard models with no attention, models with applied CBAM and SE attention blocks, and models with our proposed ignoring approach with both CBAM and SE variants with the three inversion function variants presented in Section 3.

All the models are optimized using Stochastic Gradient Descent (SGD) [30] with a momentum of 0.9 [31], weight decay of 0.0001 [32], and a batch size of 128. The initial learning rate is set to 0.1 and is then decreased by a factor of 5 after 60, 120, and 160 epochs, respectively. The models are trained for 200 epochs with the best performance on the validation set used for testing. Each experiment is repeated three times and the average performance is reported. 40k images are used for training and 10k for validation. Standard data augmentation is used [33], [34].

In Table 1, we report the experimental results of the standard model, i.e., no attention, SE, and our different SE-based variants, namely, SE-Ign<sub>i</sub> where *i* indicates the flipping function used ( $T_1$  or  $T_2$  or  $T_3$ ). For the first variant, i.e., SE-Ign<sub>1</sub>, we experiment with three different values of hyper-parameter  $\alpha$ : 1, 0.8, and 0.5. We note that for both architectures applying an explicit or implicit attention mechanism consistently outperforms the standard model. On CIFAR10, the best performance is achieved using our third variant, i.e., SE-Ign<sub>3</sub>, which improves the results by 1% compared to standard and +0.3% compared SE using ResNet50 architecture. On CIFAR100, the lowest top1-% error rates are achieved by SE-Ign<sub>3</sub> and SE-Ign<sub>1</sub>( $\alpha=0.5$ ) for ResNet50 and DenseNet architectures, respectively. In fact, on this dataset our third variant boosts the accuracy by 4% compared to the standard and 1.85% compared to SE. This can be explained by the fact that for this dataset only 500 training samples per class are available, thus making it hard to directly learn the relevant visual features for each class. At the same time, the irrelevant features are more universal and typically independent of the class, thus making them easier to learn in a scarce data context.

In Table 2, we report the empirical results for the different CBAM-based variants. As can be seen, the results with this attention variant are consistent with our findings using SE. For both datasets and for both architectures, learning to ignore yields better performance compared to both the standard model and the SE attention. The top performance is achieved by either by CBAM-Ign<sub>1</sub>( $\alpha=0.5$ ) or CBAM-Ign<sub>1</sub>( $\alpha=0.8$ ) variant. More results can be found Supplementary material Table 1.

## 4.2 ImageNet

To further validate the effectiveness of our learning to ignore framework, we perform additional experiments on ImageNet dataset [35] using ResNet50. For training on ImageNet, optimization is done with SGD with the same weight decay and momentum as used for CIFAR datasets. The initial learning rate is set to 0.1 and reduced by a factor of 10 after 30, 60, and 80 epochs, respectively. The models are trained for 90 epochs with batch size of 256 with the results reported on the validation set.

Table 3 shows the results on ImageNet dataset, where Top-1 and Top-5 errors are reported. As can be seen, our

results are consistent with findings on CIFAR10 and CIFAR100 datasets. Specifically, we find that applying attention, whether explicit or implicit, outperforms standard model. At the same time, the proposed framework based on ignoring outperforms the conventional attention in a vast majority of cases. In SE variant, SE-Ign<sub>1</sub>( $\alpha=1$ ) and SE-Ign<sub>3</sub> outperform the conventional approach, while other variants report competitive results with minimal gap. Best result of SE-Ign<sub>3</sub> outperforms the standard model by 1.1%. In CBAM, all variants of CBAM-Ign<sub>1</sub> outperform conventional approach on both Top-1 and Top-5 metric, and CBAM-Ign<sub>2</sub> and CBAM-Ign<sub>3</sub> outperform conventional CBAM on Top-5 metric, while being competitive on Top-1 metric. More results can be found Supplementary material Table 2.

## 4.3 NTU-RGBD

To further demonstrate the effectiveness of our approach, we additionally evaluate the proposed method in the multimodal fusion setting. Here, we rely on the Multimodal Transfer Module (MMTM) [36] architecture for our evaluation. MMTM is a method for fusing information from multiple modalities in multiple-stream architectures, which has recently shown good performance in a variety of tasks, including activity recognition, gesture recognition, and audiovisual speech enhancement.

The method relies on an architecture inspired from Squeeze-and-Excitation blocks placed between network branches. Specifically, considering a two-stream scenario, intermediate feature representations from two network branches corresponding to two modalities are first spatially squeezed into channel descriptors by applying global average pooling in each branch. The squeezed representations are subsequently concatenated and projected into a joint lower-dimensional space. The resulting features are transformed with two projection matrices corresponding to each of the two modalities to the spaces of original dimensionalities, and sigmoid activation is then applied to obtain attention masks. The masks are further multiplied element-wise with original feature representations in each branch.

As can be seen, the fusion module is essentially a multimodal SE-block with joint squeeze and modality-specific excitation operations, to which we apply our ignoring framework as described in Section 3.1. We perform experiments on NTU-RGBD dataset [37] for human action recognition, where we fuse the skeleton and RGB modalities, similarly to MMTM [36]. We follow our ignoring paradigm and replace the SE attention mask in each branch with our proposed approach. The rest of the architecture and training protocol follows that of MMTM. We initialize the model from ImageNet+Kinetics pretrained weights, finetune for 10 epochs with batch size 8, and report the test set performance of the model that performed best on validation set. The results are reported in Table 4. As can be seen, the proposed ignoring approaches outperform the baseline in the vast majority of cases.

## 4.4 Discussion

As can be seen from the experimental results in previous sections, learning to ignore consistently yields superior performance compared to the baselines. We argue that this

		CIFAR 10	CIFAR 100	
		Top-1 Error%	Top-1 Error%	Top-5 Error%
ResNet50	Standard	08.27 ± 0.54	34.06 ± 1.02	10.97 ± 0.54
	SE	07.63 ± 0.37	32.80 ± 0.11	09.97 ± 0.50
	SE-Ign <sub>1</sub> ( $\alpha=1$ )	07.42 ± 0.29	32.50 ± 0.26	09.92 ± 0.37
	SE-Ign <sub>1</sub> ( $\alpha=0.5$ )	07.61 ± 0.46	31.40 ± 0.68	<b>09.39 ± 0.19</b>
	SE-Ign <sub>1</sub> ( $\alpha=0.8$ )	07.76 ± 0.73	32.71 ± 1.15	10.07 ± 0.64
	SE-Ign <sub>2</sub>	07.66 ± 0.13	32.78 ± 0.77	10.11 ± 0.56
	SE-Ign <sub>3</sub>	<b>07.28 ± 0.17</b>	<b>30.95 ± 0.08</b>	09.49 ± 0.36
DenseNet	Standard	07.07 ± 0.33	29.25 ± 0.10	08.26 ± 0.12
	SE	06.96 ± 0.05	29.43 ± 0.44	08.36 ± 0.33
	SE-Ign <sub>1</sub> ( $\alpha=1$ )	06.94 ± 0.07	29.17 ± 0.07	08.22 ± 0.13
	SE-Ign <sub>1</sub> ( $\alpha=0.5$ )	06.69 ± 0.04	<b>27.64 ± 0.30</b>	<b>07.30 ± 0.10</b>
	SE-Ign <sub>1</sub> ( $\alpha=0.8$ )	06.95 ± 0.14	27.73 ± 0.41	07.39 ± 0.07
	SE-Ign <sub>2</sub>	06.80 ± 0.09	28.08 ± 0.35	07.39 ± 0.23
	SE-Ign <sub>3</sub>	<b>06.41 ± 0.08</b>	27.77 ± 0.54	07.65 ± 0.20

TABLE 1

Results of SE variants on CIFAR10 and CIFAR100 datasets.

		CIFAR 10	CIFAR 100	
		Top-1 Error%	Top-1 Error%	Top-5 Error%
ResNet50	Standard	08.27 ± 0.54	34.06 ± 1.02	10.97 ± 0.54
	CBAM	08.04 ± 0.03	31.46 ± 0.20	09.32 ± 0.15
	CBAM-Ign <sub>1</sub> ( $\alpha=1$ )	07.78 ± 0.28	31.03 ± 0.25	09.28 ± 0.27
	CBAM-Ign <sub>1</sub> ( $\alpha=0.5$ )	<b>07.17 ± 0.05</b>	30.58 ± 0.20	09.25 ± 0.23
	CBAM-Ign <sub>1</sub> ( $\alpha=0.8$ )	07.40 ± 0.23	<b>30.28 ± 0.39</b>	<b>09.08 ± 0.33</b>
	CBAM-Ign <sub>2</sub>	07.53 ± 0.29	31.42 ± 0.58	09.27 ± 0.21
	CBAM-Ign <sub>3</sub>	07.60 ± 0.10	30.88 ± 0.22	09.38 ± 0.32
DenseNet	Standard	07.07 ± 0.33	29.25 ± 0.10	08.26 ± 0.12
	CBAM	07.21 ± 0.23	30.63 ± 0.23	08.90 ± 0.14
	CBAM-Ign <sub>1</sub> ( $\alpha=1$ )	07.19 ± 0.26	29.63 ± 0.46	08.37 ± 0.39
	CBAM-Ign <sub>1</sub> ( $\alpha=0.5$ )	06.53 ± 0.14	27.92 ± 0.19	07.58 ± 0.27
	CBAM-Ign <sub>1</sub> ( $\alpha=0.8$ )	<b>06.40 ± 0.14</b>	<b>27.11 ± 0.08</b>	<b>07.33 ± 0.19</b>
	CBAM-Ign <sub>2</sub>	06.80 ± 0.02	27.88 ± 0.59	07.62 ± 0.05
	CBAM-Ign <sub>3</sub>	06.68 ± 0.05	27.94 ± 0.10	07.78 ± 0.21

TABLE 2

Results of CBAM variants on CIFAR10 and CIFAR100 datasets.

	Top-1 Error%	Top-5 Error%
Standard	23.73	06.85
SE	22.70	06.35
SE-Ign <sub>1</sub> ( $\alpha=1$ )	22.60	<b>06.29</b>
SE-Ign <sub>1</sub> ( $\alpha=0.5$ )	23.03	06.58
SE-Ign <sub>1</sub> ( $\alpha=0.8$ )	22.88	06.30
SE-Ign <sub>2</sub>	23.16	06.55
SE-Ign <sub>3</sub>	<b>22.59</b>	06.32
CBAM	22.91	06.58
CBAM-Ign <sub>1</sub> ( $\alpha=1$ )	<b>22.84</b>	06.50
CBAM-Ign <sub>1</sub> ( $\alpha=0.5$ )	<b>22.84</b>	06.52
CBAM-Ign <sub>1</sub> ( $\alpha=0.8$ )	<b>22.84</b>	06.40
CBAM-Ign <sub>2</sub>	23.02	<b>06.39</b>
CBAM-Ign <sub>3</sub>	23.10	06.44

TABLE 3

Results of CBAM and SE with variants of ignoring on ImageNet dataset

stems from the fact that learning irrelevant information is easier than identifying what should be attended. For example, in order to learn features that should be attended to, the model needs to first learn to extract patterns such as lines and edges and make associations with the class labels in order to produce a meaningful attention mask. On the other hand, irrelevant patterns, such as background textures and borderline pixels, are often shared across the dataset, are persistent and independent of the class labels, which makes them easier to learn. Therefore, it should be possible to learn them already in the early stages of training. Figure 1 shows

the validation loss curves of the baseline attention methods and the best-performing ignoring methods with ResNet50 on CIFAR100 dataset (more training curves can be found in supplementary material). As can be seen, especially at the earlier stages of training, our approach results in lower loss with less fluctuations and more stable training, hence supporting our claim. From an optimization point of view, in the case of  $\alpha=1$ , only the gradient of the attention blocks are flipped, and thus in the back-propagation, when they are summed with the gradient of the main block (which are not flipped), the total feedback carried to the earlier layers is different and does not correspond to a flipped version of the total sum of the standard attention. Thus, this yields different feedback and leads to a different optimal solution in the end of the training (Figure 7 in supplementary material).

Moreover, in Figure 2, we provide visual results of the class activation maps [38] produced by the different models on three different samples from validation set of ImageNet. As can be seen, the learning to ignore formulation leads to different attention maps compared to the explicit attention, i.e., learning to attend. Noticeably, standard CBAM attention tries to capture the relevant parts of the image directly, leading to the prediction being made based on the small part of the input that is considered by the model as the most important. This leads to the possibility that the model can miss some important parts of the class of interest on the image. As an example, only one of the plants on the lower

	MMTM	$\text{Ign}_1(\alpha=1)$	$\text{Ign}_1(\alpha=0.5)$	$\text{Ign}_1(\alpha=0.8)$	$\text{Ign}_2$	$\text{Ign}_3$
NTU-RGBD	89.98	89.99	<b>90.52</b>	88.70	90.21	90.36

TABLE 4  
Accuracy on NTU-RGBD dataset

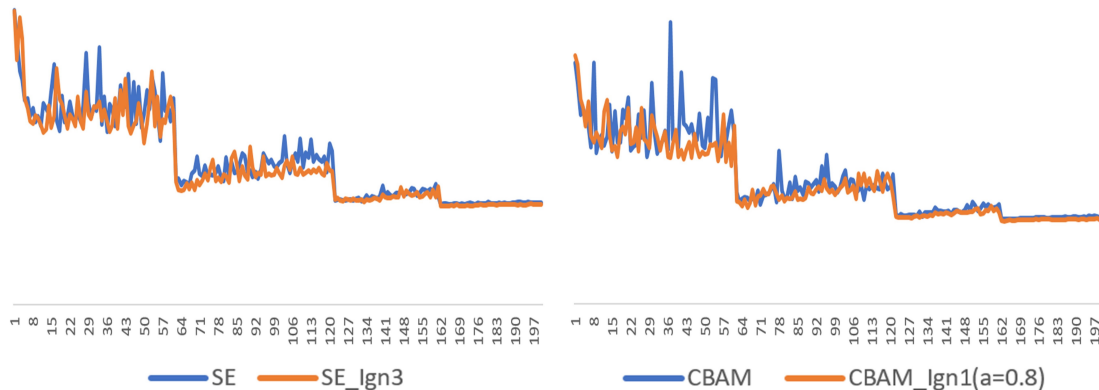


Fig. 1. Validation loss curves of ResNet50 on CIFAR100 using the different attention approaches.

figure is considered in CBAM model, as well as only a side of the bus in the middle image. On the other hand, our approach by learning to identify the non-relevant background regions first and subsequently suppressing them, simplifies the problem and typically results in an attention mask that is broader and captures the object of interest better, hence reducing the risk of suppressing relevant attributes of it.

## 5 CONCLUSION

In this paper, we provide a new perspective on attention in CNNs where the main target is learning to ignore instead of learning to attend. To this end, we propose an implicit attention scheme with three variants which can be incorporated into any existing attention mechanism. The proposed approach explicitly learns to identify the irrelevant and confusing parts of the scene and suppresses them. In addition, we reformulate two state-of-the-art attention approaches, namely SE and CBAM, using our learning paradigm. Experimental results on three image classification datasets show that learning to ignore, i.e., implicit attention consistently outperforms standard attention across multiple models.

## ACKNOWLEDGMENTS

This work has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 871449 (OpenDR), and the NSF-Business Finland Center for Visual and Decision Informatics (CVDI) project AMALIA. The authors wish to acknowledge CSC – IT Center for Science, Finland, for computational resources.

## REFERENCES

- [1] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.
- [2] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “Cbam: Convolutional block attention module,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.
- [3] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.
- [4] M. Jiang, Y. Yuan, and Q. Wang, “Self-attention learning for person re-identification,” in *British Machine Vision Conference*, 2018, p. 204.
- [5] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in *International conference on machine learning*. PMLR, 2015, pp. 2048–2057.
- [6] J. D. Cosman, K. A. Lowe, W. Zinke, G. F. Woodman, and J. D. Schall, “Prefrontal control of visual distraction,” *Current biology*, vol. 28, no. 3, pp. 414–420, 2018.
- [7] J. M. Gaspar, G. J. Christie, D. J. Prime, P. Joliceur, and J. J. McDonald, “Inability to suppress salient distractors predicts low visual working memory capacity,” *Proceedings of the National Academy of Sciences*, vol. 113, no. 13, pp. 3693–3698, 2016.
- [8] N. Gaspelin and S. J. Luck, “The role of inhibition in avoiding distraction by salient stimuli,” *Trends in cognitive sciences*, vol. 22, no. 1, pp. 79–92, 2018.
- [9] C. A. Cunningham and H. E. Egeth, “Taming the white bear: Initial costs and eventual benefits of distractor inhibition,” *Psychological science*, vol. 27, no. 4, pp. 476–485, 2016.
- [10] H. Larochelle and G. E. Hinton, “Learning to combine foveal glimpses with a third-order boltzmann machine,” *Advances in neural information processing systems*, vol. 23, pp. 1243–1251, 2010.
- [11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” *arXiv preprint arXiv:1706.03762*, 2017.
- [12] G. Wu, X. Zhu, and S. Gong, “Spatio-temporal associative representation for video person re-identification,” in *British Machine Vision Conference*, 2019, p. 278.
- [13] F. Zhang, B. Ma, H. Chang, S. Shan, and X. Chen, “Relation-aware multiple attention siamese networks for robust visual tracking,” in *British Machine Vision Conference*, 2019.
- [14] C. Aytekin, A. Iosifidis, and M. Gabbouj, “Probabilistic saliency estimation,” *Pattern Recognition*, vol. 74, pp. 359–372, 2018.
- [15] J. Zhang, T. Zhang, Y. Dai, M. Harandi, and R. Hartley, “Deep unsupervised saliency detection: A multiple noisy labeling perspective,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9029–9038.
- [16] N. Liu, J. Han, and M.-H. Yang, “Picanet: Learning pixel-wise contextual attention for saliency detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3089–3098.
- [17] N. Liu, N. Zhang, K. Wan, J. Han, and L. Shao, “Visual saliency transformer,” *arXiv preprint arXiv:2104.12099*, 2021.
- [18] N. Liu and J. Han, “Dhsnet: Deep hierarchical saliency network for salient object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 678–686.

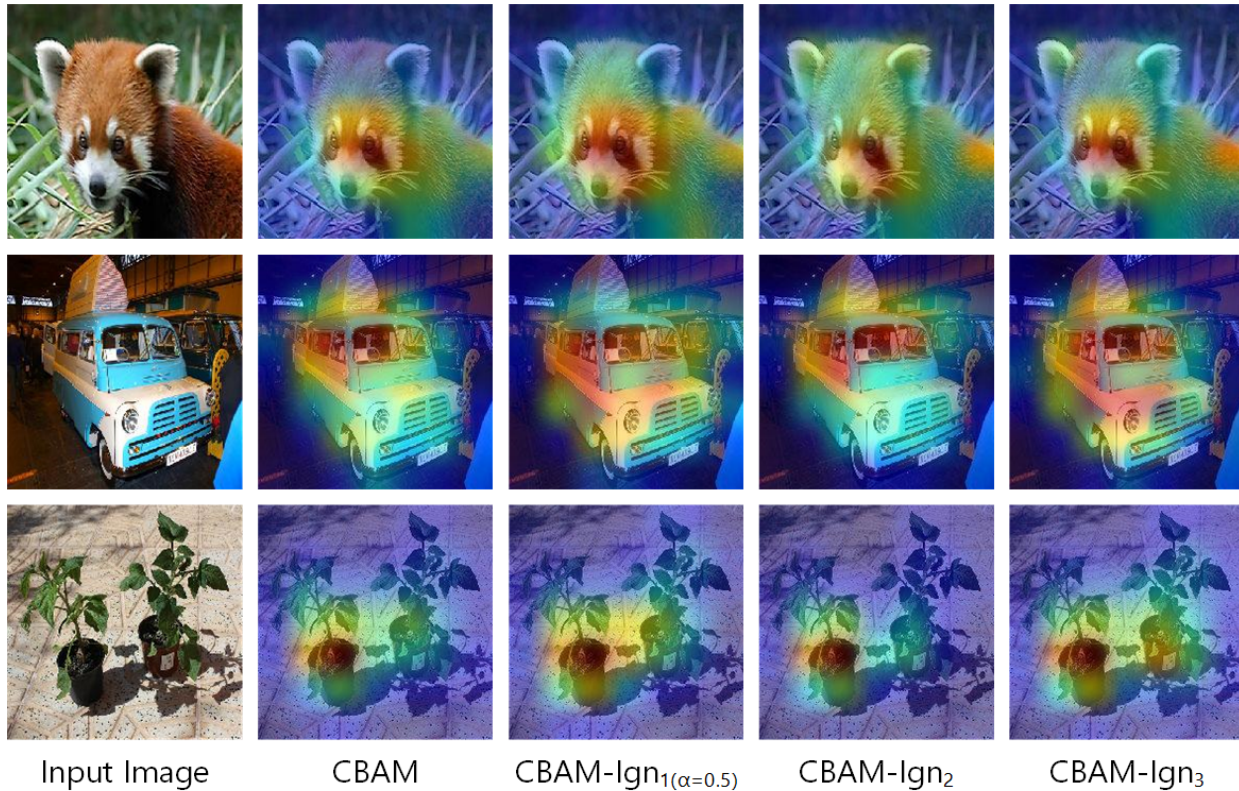


Fig. 2. Visual results of different CBAM-based attention mechanisms on three different samples from validation set of ImageNet. The attention masks are obtained as in [38].

- [19] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.
- [20] D. Zhang, J. Han, G. Cheng, and M.-H. Yang, "Weakly supervised object localization and detection: A survey," *IEEE transactions on pattern analysis and machine intelligence*, 2021.
- [21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [22] J. Gu, Y. Wang, K. Cho, and V. O. Li, "Improved zero-shot neural machine translation via ignoring spurious correlations," *arXiv preprint arXiv:1906.01181*, 2019.
- [23] B. Jiang, L. Zhang, H. Lu, C. Yang, and M.-H. Yang, "Saliency detection via absorbing markov chain," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 1665–1672.
- [24] X. Zhao, X. He, and P. Xie, "Learning by ignoring, with application to domain adaptation," *arXiv preprint arXiv:2012.14288*, 2020.
- [25] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang, "Saliency detection via dense and sparse reconstruction," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 2976–2983.
- [26] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 2814–2821.
- [27] F. Laakom, J. Raitoharju, A. Iosifidis, U. Tuna, J. Nikkanen, and M. Gabbouj, "Probabilistic color constancy," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 978–982.
- [28] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.
- [29] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.
- [30] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.
- [31] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.
- [32] A. Krogh and J. A. Hertz, "A simple weight decay can improve generalization," in *Advances in neural information processing systems*, 1992, pp. 950–957.
- [33] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.
- [34] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *International Conference on Learning Representations 2018*, 2018.
- [35] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
- [36] H. R. V. Joze, A. Shaban, M. L. Iuzzolino, and K. Koishida, "Mmtm: Multimodal transfer module for cnn fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 289–13 299.
- [37] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1010–1019.
- [38] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.