# Deep Reinforced Attention Learning for Quality-Aware Visual Recognition

Duo Li and Qifeng Chen

The Hong Kong University of Science and Technology
duo.li@connect.ust.hk    cqf@ust.hk

**Abstract.** In this paper, we build upon the weakly-supervised generation mechanism of intermediate attention maps in any convolutional neural networks and disclose the effectiveness of attention modules more straightforwardly to fully exploit their potential. Given an existing neural network equipped with arbitrary attention modules, we introduce a meta critic network to evaluate the quality of attention maps in the main network. Due to the discreteness of our designed reward, the proposed learning method is arranged in a reinforcement learning setting, where the attention actors and recurrent critics are alternately optimized to provide instant critique and revision for the temporary attention representation, hence coined as Deep REinforced Attention Learning (DREAL). It could be applied universally to network architectures with different types of attention modules and promotes their expressive ability by maximizing the relative gain of the final recognition performance arising from each individual attention module, as demonstrated by extensive experiments on both category and instance recognition benchmarks.

**Keywords:** Convolutional Neural Networks, Attention Modules, Reinforcement Learning, Visual Recognition

## 1 Introduction

Attention is a perception process that aggregates global information and selectively attends to the meaningful parts while neglects other uninformative ones. Mimicking the attention mechanism has allowed deep Convolutional Neural Networks (CNNs) to efficiently extract useful features from redundant information contexts of images, videos, audios, and texts. Consequently, attention modules further push the performance boundary of prevailing CNNs in handling various visual recognition tasks. Recently, popularized attention operators usually follow the modular design which could be seamlessly integrated into feed-forward neural network blocks, such as channel attention [11] and spatial attention [40] modules. They learn to recalibrate feature maps via inferring corresponding importance factors separately along the spatial or channel dimension.

These attention modules are critical components to capture the most informative features and guide the allocation of network weights to them. Nevertheless, existing attention emerges automatically along with the weak supervision of the topmost classification objective, which is not dedicatedly devised for the intermediate attention generation. Thus, this weakly-supervised optimization scheme may lead to sub-optimal outcomes regarding the attention learning process. In other words, the attention maps learned in such a manner might be opaque in its discrimination ability. Linsley *et al.* propose to supervise the intermediate attention maps with human-derived dense annotations [18], but the annotation procedure could be both labor-intensive and easily affected by subjective biases. To dissolve the above deficiency, we propose Deep REinforced Attention Learning (DREAL) to provide direct supervision for attention modules and fully leverage the representational power of their parameters, thus promoting the final recognition performance. Our method does not require additional annotations and is generic to popular attention modules in CNNs. In addition to the conventional weakly-supervised paradigm, we introduce critic networks in parallel to the main network to evaluate the quality of intermediate attention modules. After investigating the source feature map and the inferred attention map to predict the expected critique[1], the critic network straightforwardly transmits a supervisory signal to the attention module based on the variation of the final recognition performance with or without the effect of this attention module. With this introspective supervision mechanism, the attention module could promptly identify to what degree its behavior benefits the whole model and adapt itself accordingly. If the allocation of attention weights is not favored at the moment, the attention module would correct it instantly according to the feedback from the critic network. In practice, to avoid the unacceptable cost of high-capacity modules, we adopt the recurrent LSTM cell as the critic network, which imposes a negligible parameter and computational burden on the whole network. Furthermore, it implicitly bridges the current layer and the previous layers, enhancing the interactions of features and attention maps at different depths in order to inject more contextual information into the critique.

Considering the supervision for optimizing the critic network, we develop an intuitive criterion that reflects the effect of attention on the amelioration of the final recognition results. This evaluation criterion is non-differentiable so the conventional back-propagation algorithm is hardly applicable. To solve this discrete optimization problem, we encompass the attention-equipped main network and the critic meta network into a reinforcement learning algorithm. Our proposed model can be served as the contextual bandit [15], a primitive instance of reinforcement learning model where all actions are taken in one single shot of the state. Specifically, in a convolutional block, the intermediate feature map is defined as the *state* while the relevant *action* is the attention map conditioned on its current feature map at a training step. The critic network takes the state and action as input and estimates the corresponding critic value. With the joint optimization of the attention actor and the recurrent critic, the quality of attention could be boosted progressively, driven by the signal of reward which measures the relative gain of attention modules in the final recognition accuracy. In a quality-aware style, attention modules would be guided with the direct supervi-

[1] "Critique" refers to the critic value outputted from the critic network in this paper.

sion from critic networks to strengthen the recognition performance by correctly emphasizing meaningful features and suppressing other nuisance factors.

On the ImageNet benchmark, DREAL leads to consistently improved performance for baseline attention neural networks, since attention maps are obtained in a more quality-oriented and reinforced manner. It can be applied to arbitrary attention types in a plug-and-play manner with minimal tunable hyperparameters. To explore its general applicability, the reinforced attention networks are further applied to the person re-identification task, achieving new state-of-the-art results on two popular benchmarks including Market-1501 and DukeMTMC-reID among recent methods which involve the attention mechanism. We also visualize the distribution of some attention maps for a clearer understanding of the improved attention-assisted features, illustrating how the critic network acts on these attention maps. Quantitative and qualitative results provide strong evidence that the learned critic not only improves the overall accuracy but also encodes a meaningful confidence level of the attention maps.

Summarily we make the following contributions to attention-equipped neural network architectures:

❏ We propose to assess the attention quality of existing modular designs using auxiliary critic networks. To the best of our knowledge, it has never been well studied in the research field to explicitly consider the attention quality of features inside backbone convolutional neural networks before us.

❏ We further bridge the critic networks and the backbone network with a reinforcement learning algorithm, providing an end-to-end jointly training framework. The formulation of reinforced optimization paves a creative way to solve the visual recognition problem with a quality-aware constraint.

❏ Our critic networks introduce negligible parameters and computational cost, which could also be completely removed during inference. The critic networks could slot into network models with arbitrary attention types, leading to accuracy improvement validated by comprehensive experiments.

## 2 Related Work

We revisit attention modules in the backbone network design and reinforcement learning applications associated with attention modeling in previous literature. We clarify the connections and differences of our proposed learning method with these existing works.

**Attention Neural Networks.** Recently, the attention mechanism is usually introduced to modern neural networks as a generic operation module, augmenting their performance with minimal additional computation. ResAttNet [36] stacks residual attention modules with trunk-and-mask branches. The auxiliary mask branch cascades top-down and bottom-up structure to unfold the feed-forward and feedback cognitive process, generating soft weights with mixed attention in an end-to-end fashion. The pioneering SENet [11] builds the foundation of a research area that inserts lightweight modular components to improve the functional form of attention. The proposed SE block adaptively recalibrates channel-wise feature responses by explicitly modeling interdependencies between channels, substantially improving the performance when adapted to any state-of-the-art neural network architectures. The follow-up GENet [10] gathers contextual information spreading over a large spatial extent and redistributes these aggregations to modulate the local features in the spatial domain. To take one step further, MS-SAR [39] collects all responses in the neighborhood regions of multiple scales to compute spatially-asymmetric importance values and reweights the original responses with these recalibration scores. CBAM [40] and BAM [23] come up with to decompose the inference of the three-dimensional attention map along spatial and channel dimensions and arrange them in a sequential or parallel layout for feature refinement. SRM [16] summarizes response statistics of each channel by style pooling and infers recalibration weights through the channel-independent style integration, leveraging the latent capability of style information in the decision making process. ECA-Net [37] applies a local cross-channel interaction strategy that is efficiently implemented by the fast 1D convolution with a kernel of adaptive size. As stated above, most existing methods are dedicated to developing sophisticated feature extraction and recalibration operations, but attention maps are sustained by weakly long-distance supervision. Probably [18] is most related to us regarding the *motivation*, which also attempts to augment the weakly-supervised attention derived from category-level labels. The referred approach first introduces an extra large-scale data set ClickMe with human-annotated salient regions. It then incorporates ClickMe supervision to the intermediate attention learning process of their proposed GALA module (an extension of the seminal SE module). In stark contrast to prior works, we do not propose any new attention modules or leverage external data and annotations for supervision. By employing a shared LSTM to evaluate these attention modules, our approach concentrates on promoting the quality-aware evolution of attention maps via a novel reinforcement learning design. Recently, the non-local modules [38] thrive as a self-attention mechanism. We also elaborate on this subarea of attention research in the supplementary materials. Generally speaking, our DREAL method could be readily applied to neural networks armed with all aforementioned attention modules regardless of their specific forms.

**Deep Reinforcement Learning.** Unlike conventional supervised machine learning methods, reinforcement learning has been originated from humans' decision making process [19]. It aims at enabling the agent to make decisions or select actions optimally based on rewards it receives from an environment. Recently, the field of reinforcement learning resurrects with the strong support of deep learning techniques. Deep Reinforcement Learning (DRL), as a principal paradigm, can be roughly divided into two categories: deep Q learning [8][21] and policy gradient [1][30]. In the former class, the goal of deep Q Networks is to fit a Q-value function to capture the expected return for taking a particular action at a given state. In the latter class, policy gradient methods approximate the policy which maximizes the expected future reward using gradient descent.

Deep reinforcement learning has been adopted in the selection procedure of attended parts for computer vision applications. For example, locating the most discriminative ones among a sequence of image patches can be naturally formulated as an MDP process and contributory to a wide array of tasks such as single-label [20] or multi-label [5] image classification, face hallucination [2] and person re-identification [13]. In these exemplars, a policy-guided agent usually traverses the spatial range of a single image to dynamically decide the attended regions via progressively aggregating regional information collected in the past. Distinct from spatially attentive regions in the image space, our research focuses on the attention modules in the backbone networks that are represented with feature-level attention maps instead of image-level saliency maps. In the same spirit, deep reinforcement learning is also utilized in the video space to find appropriate focuses across frames. This kind of attention indicates discarding the misleading and confounding frames within the video for face [25] or action [7] recognition. For comparison, the attention is defined in the spatial domain of an image or the temporal domain of a video segment in the aforementioned works while our formulation is shaped inside the convolutional blocks with attention operators. DRL has also been applied to the field of neural network architecture engineering but mainly focused on network acceleration and automated search, which is depicted in detail in the supplementary materials. Unlike this research line, we propose to measure and boost the quality of attention generation under the reinforcement learning framework. To the best of our knowledge, little progress with reinforcement learning has been made in the fundamental problem of handcrafted attention-equipped CNNs, which is of vital importance in the neural architecture design.

## 3 Approach

In this section, we first overview the proposed formulation of Deep REinforced Attention Learning (DREAL) and then elaborate on the critic and actor modules within this regime. Finally we describe the optimization procedure in detail.

### 3.1 Overview

Formally, let $\mathbf{X}$ denote the input image example, the intermediate feature map in a convolutional block is represented as the state $\mathbf{F}(\mathbf{X}; \boldsymbol{\mathcal{W}})$, where $\boldsymbol{\mathcal{W}}$ is the weight matrix of the backbone network. The corresponding attention action conditioned on the feature map emerges with an auxiliary operation module, represented as $\mathbf{A}(\mathbf{F}; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ defines the parameters of the attention module. Given the predefined state-action pair above, a critic network predicts the state-action value (Q-value) function as $Q(\mathbf{A}|\mathbf{F}; \boldsymbol{\phi})$, where $\boldsymbol{\phi}$ symbolizes the weights of this critic network (deep Q network).

To guide the critic network to predict the actual quality of our attention module, we design a reward $R$ as its direct supervision signal. The reward function reflects the relative gain for the entire network regarding one specific attention module. This reward concerning the $l^{\text{th}}$ attention module is defined as

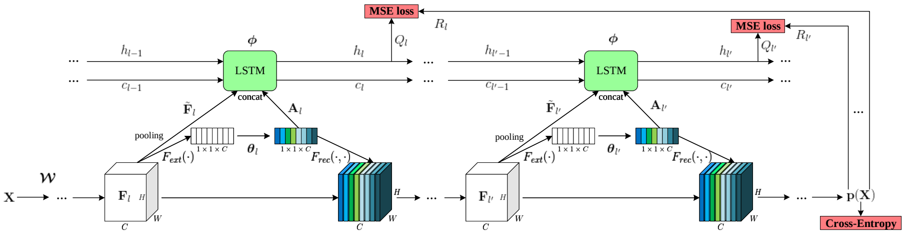

**Fig. 1.** Schematic illustration of our proposed Deep REinforced Attention Learning, built with the SENet [11] as an instance. Two selected building blocks in the same stage are presented for the purpose of conciseness. Best viewed in color and zoomed in.

$$R_l = \begin{cases} 1 - \frac{\mathbf{p}_c(\mathbf{X}|\mathbf{A}_1,\mathbf{A}_2,\cdots,\mathbf{A}_{l-1},\bar{\mathbf{A}}_l,\mathbf{A}_{l+1},\cdots,\mathbf{A}_L)}{\mathbf{p}_c(\mathbf{X}|\mathbf{A}_1,\mathbf{A}_2,\cdots,\mathbf{A}_L)}, & \\ \qquad\qquad\qquad \text{if } \mathbf{p}_c(\mathbf{X}|\mathbf{A}) \geq \mathbf{p}_i(\mathbf{X}|\mathbf{A}) \ \forall i = 1,2,\cdots,K, & \\ -\gamma, \qquad\qquad\quad \text{otherwise}, & \end{cases} \tag{1}$$

where $\mathbf{p}(\mathbf{X}|\mathbf{A})$ or $\mathbf{p}(\mathbf{X}|\mathbf{A}_1,\mathbf{A}_2,\cdots,\mathbf{A}_L)$ denotes the probabilistic prediction of the fully attention-based network with respect to an image sample $\mathbf{X}$, with the subscript $i$ being an arbitrary category label and $c$ being the corresponding ground truth category label drawn from a total of $K$ classes. For further clarification, $\mathbf{p}_c(\mathbf{X}|\mathbf{A}_1,\mathbf{A}_2,\cdots,\mathbf{A}_{l-1},\bar{\mathbf{A}}_l,\mathbf{A}_{l+1},\cdots,\mathbf{A}_L)$ defines the prediction output after substituting the attention map from the $l^{\text{th}}$ attention module with its mean vector $\bar{\mathbf{A}}_l$ during inference, which helps to bypass the emphasizing or suppressing effect of a specific attention module while retaining all others to isolate its influence on the final prediction. On the first condition of Eqn. 1, under the premise that the fully attention-equipped network should have satisfactory recognition ability, we tend to assign large reward value to a certain attention module if the output probability for the true class declines significantly (*i.e.*, the fraction in Eqn. 1 becomes small) when this attention module loses its recalibration effect, *i.e.*, substituted by its mean vector. On the second condition of Eqn. 1, incorrect prediction of the ground truth label would lead to penalization on all attention modules with a negative reward $-\gamma$, where $\gamma$ is established as a tunable positive factor. We set the parameter $\gamma$ as 1 in our main experiments by cross-validation to strike a balance between the positive and negative reward in the above two conditions. Intuitively, this criterion could effectively incentivize attention modules to bring more benefits to the final prediction results.

In the above statement, we have a glance at the general formulation of our proposed DREAL method where the actor generates the attention map and the critic analyzes the gain from the attention actor and guides the actor to maximize this gain. We leave the detailed architectural design of the critic and actor together with the optimization pipeline in the following subsections.

### 3.2 Recurrent Critic

We take the representative SENet [11] as an exemplar, with the network architecture and computation flow illustrated in Fig. 1. It could be readily extended to other types of attention-equipped networks. The raw feature map $\mathbf{F}_l \in \mathbb{R}^{H \times W \times C}$ in the $l^{\text{th}}$ building block is processed with the extraction function $F_{ext}(\cdot)$ to capture non-local context information, which often takes the form of global average pooling in the spatial domain. This processed tensor is fed into the subsequent attention module $\mathbf{A}(\cdot; \boldsymbol{\theta})$ to produce its corresponding attention map $\mathbf{A}_l$, which is then applied to the original feature map $\mathbf{F}_l$ through the recalibration function $F_{rec}(\cdot, \cdot)$. Typically, $F_{rec}(\mathbf{F}_l, \mathbf{A}_l)$ obtained the output tensor through an element-wise multiplication of the state $\mathbf{F}_l$ and action $\mathbf{A}_l$ (broadcast if necessary to match the dimension). With the dynamically selective mechanism, a spectrum of features could be emphasized or suppressed respectively in a channel-wise manner.

Taking consideration of the critic model, even injecting a miniaturized auxiliary network separately into each layer will increase the total amount of parameters as the network depth grows. Furthermore, following this way, critique results of previous layers will be overlooked by subsequent ones. Therefore, we introduce a recurrent critic network design that benefits from parameter sharing and computation re-use to avoid heavily additional overheads. Specifically, an LSTM model is shared by all residual blocks in the same stage, where successive layers have the identical spatial size and similar channel configurations [9]. The dimension of the raw feature map is first reduced to match that of the attention map (usually using global average pooling along the channel or spatial dimension depending on the specific attention types to be evaluated), then they are concatenated and fed into the LSTM cell as the temporary input, together with the hidden and cell state from the previous layer. The LSTM network generates the current hidden state $h_l \in \mathbb{R}$ and cell state $c_l \in \mathbb{R}$ as

$$h_l, c_l = \mathrm{LSTM}(\mathrm{concat}(\tilde{\mathbf{F}}_l, \mathbf{A}_l), h_{l-1}, c_{l-1}; \boldsymbol{\phi}), \tag{2}$$

where $\tilde{\mathbf{F}}_l$ denotes the reduced version of $\mathbf{F}_l$ as stated above. The cell state stores the information from all precedent layers in the same stage, while the new hidden state is a scalar that would be directly extracted to be the output critic value for current attention assessment, written as

$$Q_l(\mathbf{A}_l | \mathbf{F}_l; \boldsymbol{\phi}) = h_l. \tag{3}$$

It is noted that if spatial and channel attention coexist, *e.g.* in the CBAM [40], two individual LSTM models will be employed to process attention maps with different shapes.

The LSTM models not only incorporate the features and attention maps in the current residual block but also recurrently integrate the decisions from previous layers in the same stage, exploring complicated non-linear relationships between them. Thus, the attention-aware features could adjust in a self-adaptive fashion as layers going deeper. The recurrent critic network implicitly captures the inter-layer dependencies to provide a more precise evaluation regarding the influence of the current attention action on the whole network.

**Complexity Analysis.** The recurrent characteristic permits the critic network to maintain reasonable parameter and computational cost. Both additional parameters and FLOPs approximately amount to $4 \times (2C \times 1 + 1 \times 1)$ for each stage, which is economic and negligible compared to the main network. Specifically, there exist 4 linear transformations that take the concatenated vector with the size of $2C$ and a one-dimensional hidden state as the input to compute two output scalars, *i.e.*, hidden and cell state. Furthermore, since an LSTM is shared throughout the same stage, the number of parameter increments may remain constant with the growing depths, referring to the comparisons between ResNet-50 and ResNet-101 with various attention types in Table 1.

### 3.3 Attention Actor

We explore various attention types as the actors, including channel, spatial and style modules, which are developed in SENet [11], CBAM [40] and SRM [16] respectively. The detailed forms of these operators are reviewed in the following.

**Channel Attention.** Different channels in the feature map could contain diverse representations for specific object categories or visual patterns. The channel attention action exploits to emphasize more informative channels and suppress less useful ones. The attention map is represented as

$$\mathbf{A}_c = \sigma(\mathbf{W}_1 \delta(\mathbf{W}_0 \mathrm{AvgPool}(\mathbf{F}))), \tag{4}$$

where $\mathbf{W}_0 \in \mathbb{R}^{\frac{C}{r} \times C}$ and $\mathbf{W}_1 \in \mathbb{R}^{C \times \frac{C}{r}}$ are weight matrices of two consecutive Fully Connected (FC) layers composing the bottleneck structure, with $r$ being the reduction ratio. $\sigma$ denotes the sigmoid function and $\delta$ refers to the ReLU [22] activation function. $\mathrm{AvgPool}(\cdot)$ indicates the global average pooling operation.

**Spatial-Channel Attention.** Non-local context information is of critical importance on object recognition, which reflects long-range dependence in the spatial domain. The spatial attention action further aggregates such kinds of information and redistribute them to local regions, selecting the most discriminative parts to allocate higher weights. The spatial attention is represented as

$$\mathbf{A}_s = \sigma(\mathrm{conv}_{7\times 7}(\mathrm{concat}(\mathit{AvgPool}(\mathbf{F}), \mathit{MaxPool}(\mathbf{F})))), \tag{5}$$

where $\mathrm{conv}_{7\times 7}(\cdot)$ defines the convolution operation with the kernel size of $7 \times 7$. The concatenation and pooling operations (denoted as *AvgPool* and *MaxPool*) here are along the channel axis, in contrast to the ordinary AvgPool above in the spatial axis. Here, the channel attention map is generated leveraging the clue of highlighted features from global maximum pooling, reformulated as

$$\mathbf{A}_c = \sigma(\mathbf{W}_1 \delta(\mathbf{W}_0 \mathrm{AvgPool}(\mathbf{F})) + \mathbf{W}_1 \delta(\mathbf{W}_0 \mathrm{MaxPool}(\mathbf{F}))). \tag{6}$$

The above two attention modules are placed in a sequential manner with the channel-first order.

**Style Recalibration.** Recently it is revealed that the style information also plays an important role in the decision process of neural networks. The style-based attention action converts channel-wise statistics into style descriptors through a Channel-wise Fully Connected (CFC) layer and re-weight each

**Algorithm 1:** **D**eep **RE**inforced **A**ttention **L**earning

```
Input: Training dataset $\mathcal{D}$, maximal iterations $M$, network depth $L$
Output: Parameters of the backbone network $\boldsymbol{\mathcal{W}}$, attention actors $\boldsymbol{\theta}$ and
        recurrent critics $\boldsymbol{\phi}$
Initialize the model parameters $\boldsymbol{\mathcal{W}}$, $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$
for $t \leftarrow 1$ to $M$ do
    Randomly draw a batch of samples $\mathcal{B}$ from $\mathcal{D}$
    foreach $\mathbf{X}$ in $\mathcal{B}$ do
        Compute feature state $\mathbf{F}(\mathbf{X}; \boldsymbol{\mathcal{W}})$
        Derive attention action $\mathbf{A}(\mathbf{F}; \boldsymbol{\theta})$
        Estimate critic value $Q(\mathbf{A}|\mathbf{F}; \boldsymbol{\phi})$
        Bypass the recalibration effect of the attention module and forward
          to infer the corresponding reward $R$
        Calculate loss functions $\mathcal{L}_c$, $\mathcal{L}_q$, $\mathcal{L}_r$
        Update $\boldsymbol{\mathcal{W}}$ with $\Delta\boldsymbol{\mathcal{W}} \propto \frac{\partial}{\partial \boldsymbol{\mathcal{W}}}\mathcal{L}_c$
        Update $\boldsymbol{\theta}$ with $\Delta\boldsymbol{\theta} \propto \frac{\partial}{\partial \boldsymbol{\theta}}(\mathcal{L}_c + \mathcal{L}_q)$
        Update $\boldsymbol{\phi}$ with $\Delta\boldsymbol{\phi} \propto \frac{\partial}{\partial \boldsymbol{\phi}}\mathcal{L}_r$
    end
end
return $\boldsymbol{\mathcal{W}}$, $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$
```

channel with the corresponding importance factor. The style recalibration map is represented as

$$\mathbf{A}_t = \sigma(\mathrm{BN}(\mathbf{W} \cdot \mathrm{concat}(\mathrm{AvgPool}(\mathbf{F}), \mathrm{StdPool}(\mathbf{F})))), \tag{7}$$

where StdPool defines the global standard deviation pooling akin to global average pooling and each row in the weight matrix $\mathbf{W} \in \mathbb{R}^{C \times 2}$ of the CFC layer is multiplied individually to each channel representation.

### 3.4 Reinforced Optimization

Unlike standard reinforcement learning, there does not exist an explicit sequential relationship along the axis of the training step or network depth. The attention action is conditioned on the feature state in a one-shot fashion, which is essentially a one-step Markov Decision Process (MDP). It could be also viewed as a contextual bandit [15] model. Furthermore, the action is a continuous value thus its optimum could be searched through gradient ascent following the solution of continuous Q-value prediction. In order to provide positive guidance for the attention module, the loss function for Q-value prediction is defined as the negative of Eqn. 3

$$\mathcal{L}_q = -Q(\mathbf{A}(\mathbf{F}; \boldsymbol{\theta})|\mathbf{F}; \boldsymbol{\phi}). \tag{8}$$

With the critic network $\boldsymbol{\phi}$ frozen, the attention actor $\boldsymbol{\theta}$ is updated to obtain higher value of critique via this loss function, which implies higher quality of attention.

In the meanwhile, the critic network is optimized via regression to make an accurate quality estimation of the attention action conditioned on the feature state. The Mean Squared Error (MSE) loss is constructed through penalizing the squared Euclidean distance between the predicted Q-value and the actual reward $R$, represented as

$$\mathcal{L}_r = \|Q(\mathbf{A}(\mathbf{F};\boldsymbol{\theta})|\mathbf{F};\boldsymbol{\phi}) - R\|^2. \tag{9}$$

With the attention actor $\boldsymbol{\theta}$ frozen this time, the critic network $\boldsymbol{\phi}$ is updated to acquire more precise quality-aware evaluation.

The supervised training has been largely in place, which employs the conventional cross-entropy for classification correctness, represented as

$$\mathcal{L}_c = -\frac{1}{|\mathcal{B}|}\sum_{\mathbf{X}\in\mathcal{B}} \log \mathbf{p}_c(\mathbf{X};\boldsymbol{\mathcal{W}},\boldsymbol{\theta}), \tag{10}$$

where $\mathcal{B}$ is a randomly sampled mini-batch within the entire dataset $\mathcal{D}$ and $\mathbf{X}$ denotes an image example with $c$ indicating its corresponding ground truth label.

In this regime, we combine the strength of supervised and reinforcement learning, alternately training the backbone architecture, attention actor models and LSTM-based critic networks. The learning scheme is summarized in Algorithm 1. During inference, recurrent critic networks are all discarded so that the computational cost is exactly identical to that of the original attention-based backbone network.

## 4 Experiments

In this section, we evaluate the proposed DREAL method on close- and open-set visual recognition tasks: image classification and person re-identification. We make comparisons with extensive baseline attention networks to demonstrate the effectiveness and generality of our method.

### 4.1 Category Recognition

We employ several attention-based networks as the backbone models, including SENet [11], CBAM [40] and SRM [16], which feature channel, spatial-channel and style attention respectively. We evaluate the reinforced attention networks on the ImageNet [6] dataset, which is one of the most large-scale and challenging object classification benchmarks up to date. It includes over 1.2 million natural images for training as well as 50K images reserved for validation, containing objects spreading across 1,000 predefined categories. Following the common practice of optimization, we adopt the Stochastic Gradient Descent (SGD) optimizer with the momentum of 0.9, the weight decay of 1e-4 and the batch size of 256. We keep in accordance with SENet [11] and train all networks for 100 epochs. The learning rate is initiated from 0.1 and divided by 10 every 30 epochs. For data augmentation, we randomly resize and crop training images to patches of 224 × 224 size with random horizontal flipping. For evaluation, we resize the shorter sides of validation images to 256 pixels without changing their aspect ratios and

**Table 1.** Recognition error comparisons on the ImageNet validation set. The standard metrics of top-1/top-5 errors are measured using the single-crop evaluation. It is noted that the additional parameters and FLOPs of our proposed reinforced attention networks exist only during the training process, originating from critic networks.

| Architecture | Params | GFLOPs | Method | Top-1 / Top-5 Err.(%) |
|---|---|---|---|---|
| SE-ResNet-50 | 28.088M | 4.091 | *official* | 23.29 / 6.62 |
| | | | *self impl.* | 22.616 / 6.338 |
| | 28.119M | 4.092 | reinforced | **22.152 / 5.948** |
| SE-ResNet-101 | 49.326M | 7.806 | *official* | 22.38 / 6.07 |
| | | | *self impl.* | 21.488 / 5.778 |
| | 49.358M | 7.811 | reinforced | **20.732 / 5.406** |
| CBAM-ResNet-50 | 28.089M | 4.095 | *official* | 22.66 / 6.31 |
| | | | *self impl.* | 22.386 / 6.172 |
| | 28.154M | 4.097 | reinforced | **21.802 / 6.084** |
| CBAM-ResNet-101 | 49.330M | 7.812 | *official* | 21.51 / 5.69 |
| | | | *self impl.* | 21.518 / 5.812 |
| | 49.394M | 7.819 | reinforced | **20.682 / 5.362** |
| SRM-ResNet-50 | 25.587M | 4.089 | *official* | 22.87 / 6.49 |
| | | | *self impl.* | 22.700 / 6.392 |
| | 25.618M | 4.090 | reinforced | **22.348 / 6.084** |
| SRM-ResNet-101 | 44.614M | 7.801 | *official* | 21.53 / 5.80 |
| | | | *self impl.* | 21.404 / 5.740 |
| | 44.644M | 7.806 | reinforced | **20.474 / 5.362** |

crop center regions of the same size as that of training images. As a special note for meta networks of critic, hidden and cell states in the LSTM cells from each stage are initialized as zero scalars. During each training epoch, one building block in each stage is bypassed to measure the corresponding reward, avoiding much additional inference cost. This optimization strategy could guarantee that each LSTM belonging to one stage is optimized all the way along with the main network. All experiments are performed with the PyTorch [24] framework.

For the baseline attention networks, we re-implement each network and achieve comparable or even stronger performance compared to those from the original papers. The officially released performance and outcomes of our re-implementation are shown in Table 1, denoted as *official* and *self impl.* respectively. We also report the parameters, computational complexities and validation errors of our reinforced attention networks correspondingly. It is noteworthy that the increment of parameters and computation is completely negligible compared to the baseline counterparts. For SE-ResNet-50, the additionally introduced parameters only occupy 0.11% of the original amount. Thanks to the parameter sharing mechanism of recurrent critics, roughly the same number of network parameters is attached to SE-ResNet-101, which consists of the same number of stages as the 50-layer version. Consequently, the relative increase of parameters is further reduced to 0.06% for this deeper backbone network. Regarding computational cost, the most significant growth among all networks does not exceed 0.1%, which comes from the reinforced CBAM-ResNet-101 with double LSTMs for both spatial and channel attention modeling.

Our reinforced attention networks bring about clear-cut reduction of error rates compared to the strong re-implementation results. The ResNet-101 networks with three types of attention obtain more improvement than their 50-layer versions, which could be attributed to the capability of our method to exploit the potential of more attention representations in these deeper models. While we explore three types of attention networks to demonstrate its wide applicability

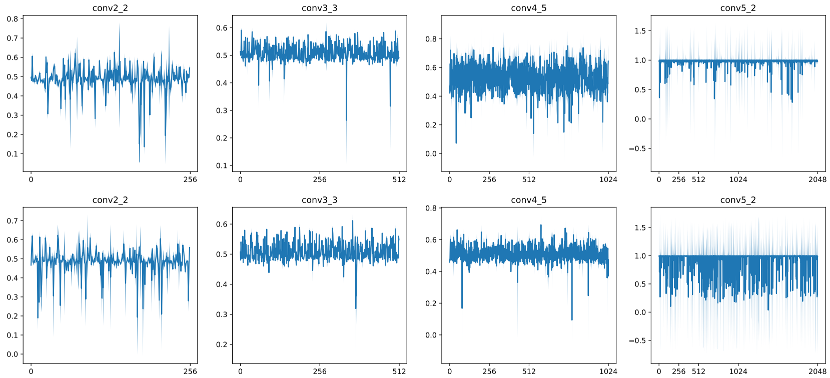

**Fig. 2.** Distributions of channel-attention vectors on the ImageNet validation set before (*top*) and after (*bottom*) applying DREAL. The x-axis represents channel index.

here, our DREAL method could be integrated into any other types of attention networks conveniently. We also explore more complicated recurrent neural network architectures for critics, but it brings marginally additional benefit with more parameters and computational costs.

**Visualization.** To provide better intuitive insight of our method, we take SE-ResNet-50 as an example and visualize the distribution of channel-attention vectors before and after applying our method. The attention maps are evaluated on the ImageNet validation set and the distributions of the last residual block in each stage are showcased in Fig. 2, where the solid line indicates the mean values among all validation image examples and the shadow area indicates 3× variance. By comparison, we observe that in certain layers (like `conv3_3` and `conv4_5`), DREAL encourages attention weights to become similar to each other across different channels. It echos the rationale that shallower layers capture fundamental visual patterns, which tend to be **category-agnostic**. In deeper layers (like `conv5_2`), with the guidance of the critic network, attention weights develop a tendence to fluctuate more but within a moderate range, flexibly extracting **category-oriented** semantic meaning for the final recognition objective. Visualization results of other layers are provided in the supplementary materials.

### 4.2 Instance Recognition

We further conduct experiments on the more challenging open-set recognition task to demonstrate the generalization ability of our learning approach. We evaluate the performance of reinforced attention networks on two widely used person re-identification benchmarks, *i.e.* Market-1501 [43] and DukeMTMC-reID [27].

**Datasets.** Person ReID is an instance recognition task with the target of retrieving gallery images of the same identity as the probe pedestrian image. The Market-1501 dataset is comprised of 32,668 bounding boxes of 1,501 identities generated by a DPM-detector, with original images captured by 6 cameras in front of the supermarket inside the campus of Tsinghua University. The conventional split contains 12,936 training images of 751 identities and 15,913 gallery images of 750 identities as well as 3,368 queries. The DukeMTMC-reID dataset consists of 36,411 images covering 1,812 identities collected by 8 cameras, where only 1,404 identities appear across camera views and the other 408 identities are regarded as distractors. The training split includes 16,522 image examples from 702 persons while the non-overlapping 17,661 gallery samples and 2,228 queries are drawn from the remaining 702 person identities.

**Implementation Details.** Following the common practice of experimental setup, we adopt the ResNet-50 model as the backbone network due to its strong track record of feature extraction. To achieve fast convergence, the backbone of ReID model is pre-trained on ImageNet for parameter initialization. The last down-sampling operation in the `conv5_x` stage is removed to preserve high resolution for a better output representation. We deploy sequential channel and spatial attention modules on the ResNet model, which resembles the arrangement of CBAM [40]. For data augmentation, input pedestrian images are first randomly cropped to the size of 384 × 128 for fine-grained representation. Then they are horizontally flipped with the probability of 0.5 and normalized with mean and standard deviation per channel. Finally, the Random Erasing [45] technique is applied to make the model robust to occlusion. In this ranking-based task, we further introduce a triplet loss to encourage inter-class separation and intra-class aggregation with a large margin, which is set as 0.5 in the experiments. To satisfy the demand for triplet loss, we employ the PK sampling strategy [28], randomly selecting P identities and K samples from each identity to form each mini-batch. In our main experiments, we set P=16 and K=8 to generate mini-batches with the size of 128. Furthermore, we apply the label-smoothing regularization [33] to the cross-entropy loss function to alleviate overfitting, where the perturbation probability for original labels is set as 0.1. We also add a Batch Normalization neck after the global average pooling layer to normalize the feature scales. The four losses in total are minimized with the AMSGRAD [26] optimizer ($\beta_1 = 0.9, \beta_2 = 0.999$, weight decay=5e-4). The learning rate initiates from 3e-4 and is divided by a factor of 10 every 40 epochs within the entire optimization period of 160 epochs. During evaluation, we feed both the original image and its horizontally flipped version into the model and calculate their mean feature representation. The extracted visual features are matched based on the similarities of their cosine distance.

**Evaluation Protocols.** We conduct evaluation under the single-query mode and adopt Cumulative Matching Characteristics (CMC) and mean Average Precision (mAP) as the evaluation metrics. CMC curve records the hit rate among the top-k ranks and mAP considers both precision and recall to reflect the performance in a more comprehensive manner. Here we choose to report the Rank-1 result in the CMC curve. For the purpose of fairness, we evaluate our method without any post-processing methods, such as re-ranking [44], which is applicable to our method and would significantly boost the performance of mAP especially.

**Performance Comparison.** As illustrated in the bottom groups of Table 2, we compare our proposed method with the baseline model as well as the attention-based one. We also compare it to other state-of-the-art methods that

**Table 2.** Comparison to state-of-the-art methods on the Market-1501 (*left*) and DukeMTMC-reID (*right*) benchmarks. Results extracted from the original publications are presented with different decimal points. Red indicates the best results while green the runner-up. ResNet-50 is employed as the backbone if no special statement.

| Method | Reference | Rank-1(%) | mAP(%) |
|---|---|---|---|
| IDEAL$^{\diamond}$ | BMVC 2017 [14] | 86.7 | 67.5 |
| MGCAM | CVPR 2018 [31] | 83.55 | 74.25 |
| AACN$^{\diamond}$ | CVPR 2018 [42] | 85.90 | 66.87 |
| DuATM$^{\dagger}$ | CVPR 2018 [29] | 91.42 | 76.62 |
| HA-CNN$^{\ddagger}$ | CVPR 2018 [17] | 91.2 | 75.7 |
| Mancs | ECCV 2018 [35] | 93.1 | 82.3 |
| AANet | CVPR 2019 [34] | 93.89 | 82.45 |
| ABD-Net | ICCV 2019 [4] | 95.60 | 88.28 |
| MHN-6 (PCB) | ICCV 2019 [3] | 95.1 | 85.0 |
| SONA$^{2+3}$ | ICCV 2019 [41] | 95.58 | 88.83 |
| baseline | This Paper | 93.5 | 82.8 |
| + attention | | 94.7 | 85.9 |
| + reinforce | | 96.1 | 89.6 |

| Method | Reference | Rank-1(%) | mAP(%) |
|---|---|---|---|
| AACN$^{\diamond}$ | CVPR 2018 [42] | 76.84 | 59.25 |
| DuATM$^{\dagger}$ | CVPR 2018 [29] | 81.82 | 64.58 |
| HA-CNN$^{\ddagger}$ | CVPR 2018 [17] | 80.5 | 63.8 |
| Mancs | ECCV 2018 [35] | 84.9 | 71.8 |
| AANet | CVPR 2019 [34] | 86.42 | 72.56 |
| ABD-Net | ICCV 2019 [4] | 89.00 | 78.59 |
| MHN-6 (PCB) | ICCV 2019 [3] | 89.1 | 77.2 |
| SONA$^{2+3}$ | ICCV 2019 [41] | 89.38 | 78.28 |
| baseline | This Paper | 84.8 | 72.5 |
| + attention | | 86.4 | 76.2 |
| + reinforce | | 89.6 | 79.8 |

$\diamond$ with the GoogleNet/Inception [32, 33] backbone.
$\dagger$ with the DenseNet-121 [12] backbone.
$\ddagger$ with the dedicatedly designed HA-CNN [17] backbone.

exploit various types of attention designs, as listed in the top groups of these two sub-tables. It is observed that harnessing the spatial and channel attention mechanism considerably enhances the performance of baseline models, while our proposed reinforced attention networks achieve further improvement over the vanilla attention networks. Specifically, with the proposed method, our model outperforms the vanilla attention network with a margin of 1.4%/0.7% regarding the Rank-1/mAP metric on the Market-1501 dataset. DukeMTMC-reID is a much more challenging dataset due to the wider camera views and more complex scene variations. In this context, our method could better demonstrate its superiority by leveraging the potential of attention representation. As a result, a more prominent performance gain of 3.2%/2.6% on the Rank-1/mAP metric is achieved. Even horizontally compared with other state-of-the-art methods that utilize dedicatedly designed backbone networks [17] or exploit higher-order attention forms [3, 41], our proposed method beats them with consistent margins on both Rank-1 accuracy and mAP results across different datasets. For example, on the Market-1501 benchmark, we surpass the nearest rival method of SONA by 0.5% and 0.8% on the Rank-1 and mAP measurement respectively.

## 5 Conclusion

In this paper, we have proposed Deep REinforcement Attention Learning (DREAL) to facilitate visual recognition in a quality-aware manner. We employ recurrent critics that assess the attention action according to the performance gain it brings to the whole model. Wrapped up in a reinforcement learning paradigm for joint optimization, critic networks would promote the relevant attention actor to focus on the significant features. Furthermore, the recurrent critic could be used as a plug-and-play module for any pre-existing attention networks with negligible overheads. Extensive experiments on various recognition tasks and benchmarks empirically verify the efficacy and efficiency of our method.

## References

1. Ammar, H.B., Eaton, E., Ruvolo, P., Taylor, M.: Online multi-task learning for policy gradient methods. In: ICML (2014)
2. Cao, Q., Lin, L., Shi, Y., Liang, X., Li, G.: Attention-aware face hallucination via deep reinforcement learning. In: CVPR (2017)
3. Chen, B., Deng, W., Hu, J.: Mixed high-order attention network for person re-identification. In: ICCV (2019)
4. Chen, T., Ding, S., Xie, J., Yuan, Y., Chen, W., Yang, Y., Ren, Z., Wang, Z.: Abd-net: Attentive but diverse person re-identification. In: ICCV (2019)
5. Chen, T., Wang, Z., Li, G., Lin, L.: Recurrent attentional reinforcement learning for multi-label image recognition. In: AAAI (2018)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR (2009)
7. Dong, W., Zhang, Z., Tan, T.: Attention-aware sampling via deep reinforcement learning for action recognition. In: AAAI (2019)
8. Gu, S., Lillicrap, T., Sutskever, I., Levine, S.: Continuous deep q-learning with model-based acceleration. In: ICML (2016)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
10. Hu, J., Shen, L., Albanie, S., Sun, G., Vedaldi, A.: Gather-excite: Exploiting feature context in convolutional neural networks. In: NeurIPS (2018)
11. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR (2018)
12. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR (2017)
13. Lan, X., Wang, H., Gong, S., Zhu, X.: Deep Reinforcement Learning Attention Selection for Person Re-Identification. arXiv e-prints arXiv:1707.02785 (Jul 2017)
14. Lan, X., Wang, H., Gong, S., Zhu, X.: Deep reinforcement learning attention selection for person re-identification. In: BMVC (2017)
15. Langford, J., Zhang, T.: The epoch-greedy algorithm for multi-armed bandits with side information. In: NIPS (2008)
16. Lee, H., Kim, H.E., Nam, H.: SRM : A style-based recalibration module for convolutional neural networks. In: ICCV (2019)
17. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: CVPR (2018)
18. Linsley, D., Shiebler, D., Eberhardt, S., Serre, T.: Learning what and where to attend with humans in the loop. In: ICLR (2019)
19. Littman, M.L.: Reinforcement learning improves behaviour from evaluative feedback. Nature **521**, 445–451 (2015)
20. Mnih, V., Heess, N., Graves, A., kavukcuoglu, k.: Recurrent models of visual attention. In: NIPS (2014)
21. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., Hassabis, D.: Human-level control through deep reinforcement learning. Nature **518**, 529–533 (2015)
22. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: ICML (2010)
23. Park, J., Woo, S., Lee, J.Y., Kweon, I.S.: BAM: Bottleneck attention module. In: BMVC (2018)

24. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: NeurIPS (2019)
25. Rao, Y., Lu, J., Zhou, J.: Attention-aware deep reinforcement learning for video face recognition. In: ICCV (2017)
26. Reddi, S.J., Kale, S., Kumar, S.: On the convergence of adam and beyond. In: ICLR (2018)
27. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set formulti-target, multi-camera tracking. In: ECCV Workshops (2016)
28. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: CVPR (2015)
29. Si, J., Zhang, H., Li, C.G., Kuen, J., Kong, X., Kot, A.C., Wang, G.: Dual attention matching network for context-aware feature sequence based person re-identification. In: CVPR (2018)
30. Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., Riedmiller, M.: Deterministic policy gradient algorithms. In: ICML (2014)
31. Song, C., Huang, Y., Ouyang, W., Wang, L.: Mask-guided contrastive attention model for person re-identification. In: CVPR (2018)
32. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: CVPR (2015)
33. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: CVPR (2016)
34. Tay, C.P., Roy, S., Yap, K.H.: Aanet: Attribute attention network for person re-identifications. In: CVPR (2019)
35. Wang, C., Zhang, Q., Huang, C., Liu, W., Wang, X.: Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In: ECCV (2018)
36. Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X.: Residual attention network for image classification. In: CVPR (2017)
37. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: ECA-Net: Efficient channel attention for deep convolutional neural networks. In: CVPR (2020)
38. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR (2018)
39. Wang, Y., Xie, L., Qiao, S., Zhang, Y., Zhang, W., Yuille, A.L.: Multi-scale spatially-asymmetric recalibration for image classification. In: ECCV (2018)
40. Woo, S., Park, J., Lee, J.Y., So Kweon, I.: CBAM: Convolutional block attention module. In: ECCV (2018)
41. Xia, B.N., Gong, Y., Zhang, Y., Poellabauer, C.: Second-order non-local attention networks for person re-identification. In: ICCV (2019)
42. Xu, J., Zhao, R., Zhu, F., Wang, H., Ouyang, W.: Attention-aware compositional network for person re-identification. In: CVPR (2018)
43. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: ICCV (2015)
44. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: CVPR (2017)
45. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: AAAI (2020)