Contrastively-reinforced Attention Convolutional Neural Network for Fine-grained Image Recognition

BMVC 2020 Submission # 656

Abstract

Fine-grained visual classification is inherently challenging because of its inter-class similarity and intra-class variance. However, by contrasting the images with same/different labels, a human can instinctively notice that the key clues lie in certain objects while other objects are ignorable. Inspired by this, we propose Contrastively-reinforced Attention Convolutional Neural Network (CRA-CNN), which reinforces the attention awareness of deep activations. CRA-CNN mainly contains two parts: the classification stream and attention regularization stream. The former classifies the input image and simultaneously proposes to divide the input visual information into attention and redundancy. The latter evaluates the attention/redundancy proposal by classifying the attention and contrasting the attention/redundancy of various inputs. The evaluation information is backpropagated and forces the classification stream to improve its awareness of visual attention, which helps classification. Experimental results on CUB-Birds and Stanford Cars show that CRA-CNN distinctly outperforms the baselines and is comparable with state-of-art studies despite its simplicity.

1 Introduction

Fine-grained visual classification (FGVC) aims to differentiate visually similar categories, such as different breeds of birds or models of cars. As an interesting yet challenging task, FGVC has recently attracted much attention. The difficulty of FGVC is principally caused by its inter-class similarity and intra-class variations.

To avoid the difficulty, existing approaches generally first locate the discriminative parts or objects and then specify the category with Convolutional neural networks (CNNs) [9, 12, 14, 16, 17, 22, 25, 26]. The benefit is that the located parts or objects preserve the information useful for classification and discard the useless information. By doing so, such approaches reduce inter-class similarity and intra-class variation.

However, attention localization is extremely complicated in itself. Inaccurate attention brings about low accuracy, while accurate attention always requires heavy effort, such as extra manual attention annotations [14, 22, 25, 34, 36] or sophisticated algorithms [9, 12, 16, 17, 26]. Thus, such attention-based approaches are always labor-intensive or/and computationally expensive, which is an obstacle for practical use. Especially in real-world applications, while the overhead in training procedure can be to some extent avoided by training beforehand, the overhead in testing (utilization) procedure is inevitably unfeasible.

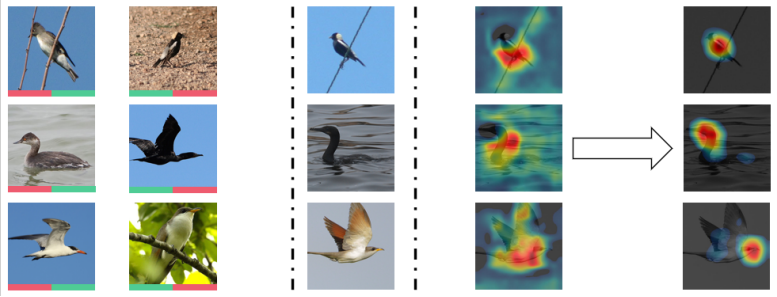

Figure 1: Motivation. (Best view in color). The middle column between two dash-dotted lines shows some images that are wrong classified by the a fine-tuned ResNet-50. The pair of colored bars below each image in the left two columns represents ground truth (left bar) or prediction result (right bar). The bar's color denotes whether the image in the same row, middle column is labeled/predicted as the same category of the image above the bar (green) or not (red). The right two columns show the middle-column images' CAM [40] generated by the ResNet-50 [11] (second column on the right) and wanted in this work (first column on the right). This work aims to reinforce the network's awareness of condition-invariant attention to improve generalization.

In this work, we do not follow the typical localize-and-recognize techniques but focus on reinforcing CNN activations' awareness of visual attention. Specifically, take the bird images as an example. As shown in Fig. 1, traditional CNNs struggle with two problems. First, bird images usually have certain habitat backgrounds, such as tree branches and water surface. If the CNNs rely on the habitats, they might make mistakes when a congeneric bird happens to be in another habitat. Second, the birds of the same species may look very different in different poses, illumination, etc., while the differences between different species are subtle. Affected by these problems, traditional CNNs sometimes fail to respond to visual attention, and that is why they need localize-and-recognize techniques. To address the problems, we attempt to regulate the networks to respond more to the core attention (e.g., bird head) that is invariant in different conditions (background, poses, illumination, etc.). In that way, networks respond to correct clues in whatever conditions and less be affected by visual redundancy.

For this purpose, we propose Contrastively-reinforced Attention Convolutional Neural Network (CRA-CNN), which consists of two network streams: classification stream N_{cls} and attention regularization stream N_{ar}. N_{cls} has two tasks: (a) predict the correct category of a given image; (b) generate a set of attention parameters conditioned on the information learned from the given image. Then, the attention parameters, along with the input image, are fed into the proposed Attention-redundancy Transformer module (ART module). The ART module divides the visual information of the input into attention and redundancy. Thereafter, N_{ar} evaluates the attention/redundancy proposal of N_{cls}, and regulates the activations of N_{cls} through standard backpropagation. We train N_{cls} and N_{ar} together during training. For the testing process, we remove N_{ar} and only use N_{cls}, and thus our approach requires no heavier overhead than basic models such as the ResNets.

We design N_{ar} inspired by the fact that humans can effectively notice informative attention by contrasting different fine-grained images. For example, given an FGVC task of bird species, after contrasting images corresponding to different labels, humans instinctively notice that the key clues for classification lie in the birds rather than the background. Then the background objects, such as the trees, become worthless in human eyes. We use N_{ar} to recognize the proposed attention. If N_{cls} correctly proposes the attention/redundancy of an input image,

for N_{ar}'s perspective, the redundancy should be similar to the redundancy of other images and contrasting to the attention of itself. Thus, in addition to softmax loss, we train N_{ar} with triplet loss to separate the attention-redundancy pair of the same image and pull closer the redundancy of different images.

Our contributions are: (1) we propose a novel neural network model that contrastively reinforces networks' awareness of condition-invariant attention; (2) Our approach is simple to implement and computationally cheap, especially in the test procedure. (3) Our approach is quite close to the state-of-the-art approaches on CUB Birds, and reaches state-of-the-art performance on Stanford Cars, while we only use a basic ResNet-101 model for testing.

2 Related Work

Region-based Attention Learning. Mainstream studies overcome inter-class similarity and intra-class variance by localizing attentional regions for classification. For learning region localization, many prior studies introduce extra annotations, such as bounding boxes and part annotations [14, 22, 25, 25, 34, 36]. However, such approaches require heavy effort in manual annotations before learning features, and thus not practical for practical use.

Recently, there have been emerging studies on automatically searching attention regions by unsupervised or weakly-supervised approaches. For example, [7, 16, 21] embeds certain learnable mechanisms within deep neural networks, and train the attention-learning mechanism together with other components of the networks. However, the problem is two-fold. First, although the localization becomes learnable, researchers have to manually fix the sizes, shapes, etc., of the attention region [7, 12, 16]. Otherwise, the networks will become too difficult to train. Second, cropping local regions brings an inevitable loss of some visual information. Because of these problems, region-based attention learning suffers from the dilemma of either keeping redundancy or losing useful information. To reduce the dilemma, researchers have to construct multi-stream architectures to obtain complementary information. Such effort includes multi-part [9, 16] or multi-level [7, 12, 17, 26] attention regions, which, however, require huge computational expenses in both training and testing procedures.

In the training procedure, our work uses a somehow similar architecture to the ones used in some region-based attention studies, such as [16]. However, our work does not recognize local regions instead of entire images but utilizes the designed loss occurred by local regions to regulate the networks. Thus, CRN-CNN explores entire visual information for utilization.

Attention-aware Deep Features. Some researchers focus on learning attention by applying selective or weighting schemes on deep features. [27] takes CNNs' activation maps as the inputs of their LSTM-based attention model. [33] uses spatially-correlated information to cluster the activation maps from different CNN channels, and then applies weighting on the grouped features to learn attention. [29] learns multiple weighted feature maps, and then uses an OSME module to enforce the attentional correlations among the feature maps.

Although our work aims to reinforce the deep activations' awareness of attention, it is quite different from all the above-mentioned studies on attention-aware deep feature learning. Our work does not focus on designing extra sophisticated schemes to apply some kinds of weighting to CNN features. Instead, our work reinforce the attention awareness by the attentional regulation from a sub-network with the designed losses. Therefore, our work is much simpler to implement than above-mentioned studies.

3 Proposed Approaches

The overview structure of proposed CRA-CNN is illustrated in Fig 2. The classification stream (N_{cls}) is the target stream that we want to improve the attention awareness. The

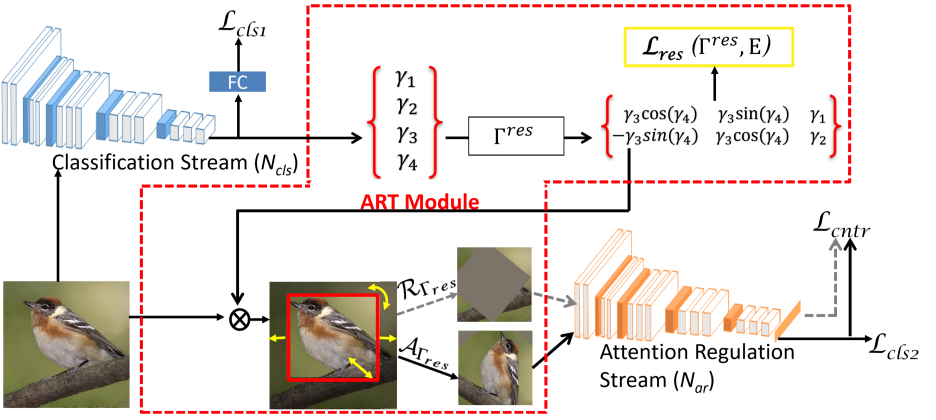

Figure 2: Pipeline of CRA-CNN. Given an input, beside classification (trained with $\mathcal{L}_{cls1}$), N_{cls} is required to output a set of transformation parameters Γ. Γ is limited to be in a reasonable range and becomes Γ_{res}, which is further restricted by $\mathcal{L}_{res}$. Parametrized transformations $\mathcal{A}_{\Gamma_{res}}$ and $\mathcal{R}_{\Gamma_{res}}$ are then applied to the input and form attention and redundancy. Γ allows localization, zooming, and rotation. Then the N_{ar} regulates N_{cls} by recognizing the attention (L_{cls2}) and contrasting the attention/redundancy (L_{cntr}).

attention regulation stream (N_{ar}) is used to force the N_{cls} to do so. The N_{cls} and N_{ar} are connected by the proposed ART module, which contains attention transformation module (AT module) and redundancy transformation module (RT module). The AT module uses the attentional information proposed by N_{cls} to define an attentional transformation on the input image. The RT module gathers the visual information uncovered in the attention and defines redundancy. The restriction loss $\mathcal{L}_{res}$ is applied to encourage the proposed transformation to be in a reasonable range. Then N_{ar} recognizes the attention's category, which should be same as the category of the input, and pull away/closer the transformed images.

Clearly, N_{cls} is forced to simultaneously: (a) predict the category of a given image; (b) predict a transformation conditioned on the attention information inside N_{cls}; (c) ensure the visual information included/excluded in the transformed image to be discriminative/redundant. By doing so, N_{cls} gradually learns the attention that helps classification. After the training, we only keep N_{cls} for testing. More details are given below.

3.1 Attention-redundancy Transformer module

ART module plays an important role as a bridge between N_{cls} and N_{ar}. ART module must be able to effectively propagate the current attention awareness of N_{cls}'s activations to N_{ar}, and then back-propagate N_{ar}'s regulation to update N_{cls}. Therefore, the ART module has to meet three requirements: (a) ART module should be differentiable and thus can be embedded within neural networks. (b) ART module should be simple and easy to optimize. Otherwise, the whole architecture might be too difficult to train, and the introduction of the ART module brings performance decrease. (c) ART module should automatically adjust the attention's locations, sizes, and angles of the for efficiently reflecting the attention awareness of the N_{cls}. The angle adjustment applies attention alignment. [□] points out that the activation maps of strong visual semantics help to align objects. We assume the converse to be also true: solving alignment tasks encourages N_{cls} to increase visual semantics.

To meet the above requirements, we turn our eyes to the Spatial Transformer (ST) module

of the Spatial Transformer Networks (STNs)[16]. The ART module is adapted from the ST module by overcoming ST module's drawbacks to meet the requirements of this work.

Spatial transformer module. ST module generates a transformed image I_t conditioned on the input I_{in}. Neglect the numbers of channels in I_t and assume the 2-D size of I_{in} to be $H \times W$ (height, width), and $H \in [1,h]$, $w \in [1,w]$. $G = \{(y_1,x_1),(y_1,x_2),...(y_2,x_1),(y_2,x_2),...(y_w,x_{h-1}),(y_h,x_w)\}$ is a regular spatial grid that defines the I_t. Similarly, assume the 2-D size of I_{in} to be $H' \times W'$ (height, width), and $H' \in [1,h']$, $w' \in [1,w']$. Let $G' = \{(y'_1,x_1)',(y'_1,x'_2),...(y_{2'},x'_1),(y'_2,x'_2),...(y'_w,x'_{h-1}),(y'_h,x'_w)\}$ to be the grid that defines the I_{in}. Then the transformation applied by the ST module can be mathematically written as:

$$G' = \mathcal{T}_\Theta(G), \qquad where\ \Theta = f_{loc}(I_{in}). \tag{1}$$

Here, f_{loc} denotes a learnable spatial transformation predictor that is named as the localization network. f_{loc} takes as input the I_{in} and predicts a set of transformation parameters $\Theta = \begin{bmatrix}\theta_{11} & \theta_{12} & \theta_{13}\\ \theta_{21} & \theta_{22} & \theta_{23}\end{bmatrix}$, which is actually a 6-dimensional affine transformation matrix. Let $g_{ij} = (x_i, y_j)$ to be a coordinate in G. Then the parametrized transformation $\mathcal{T}$ is defined as

$$\begin{pmatrix} x'_i \\ y'_j \end{pmatrix} = \mathcal{T}_\Theta(g_{ij}) = \begin{bmatrix}\theta_{11} & \theta_{12} & \theta_{13}\\ \theta_{21} & \theta_{22} & \theta_{23}\end{bmatrix}\begin{pmatrix} x_i \\ y_j \\ 1 \end{pmatrix}, \tag{2}$$

where (x'_i, y'_j) is a coordinate in G', and defines the sampling location in I_{in} for the g_{ij} location in I_t. Coordinates of G and G' are normalized by width/height. Thus, $x, y \in [-1,1]$, and $x', y' \in [-1,1]$. We follows [16] to use bilinear sampling kernel to sample points.

Details of attention-redundancy transformer module. Although the ST module has some effects of resolving spatial variations, it also faces some problems in practice. The most common problem is that the localization network struggles with noise at the early stages and introduces large errors because of the irreversible loss of visual information (also pointed out in [23]). Besides, as pointed out in [28], the ST module fails to eliminate some spatial variations. Those problems make the training of ST module a rather difficult task. To Avoid this difficulty, [16] fixes θ_{11}, θ_{12}, θ_{21} and θ_{22}, and only optimizes θ_{13} and θ_{23} when learning attentional information in the FGVC task. However, this greatly limits the variety of attention.

ART module consists of AT and RT modules. The AT module aims to efficiently reflect the attention awareness of the N_{cls} without heavy training difficulty. As mentioned, the AT module should be able to automatically adjust the attention's locations, sizes, and angles. Thus, in addition to the classifiers, N_{cls} outputs $\Gamma = \begin{bmatrix}\gamma_1 & \gamma_2 & \gamma_3 & \gamma_4\end{bmatrix}$, which is a 4-dimensional vector and we refer this vector as attention parameters. γ_1 to γ_4 respectively defines the horizontal location, vertical location, scale and alignment angle. Here, we use I_{in}, I_t and G following the same formulation as given above. G^a and G^r are respectively the grids define attention and redundancy of I_{in}. The transformation implemented by the AT module is therefore similarly defined as

$$G^a = \mathcal{A}_\Gamma(G), \qquad where\ \Gamma = f^{att}_{cls}(I_{in}). \tag{3}$$

Here, f^{att}_{cls} denotes the N_{cls}'s function of predicting attention parameters. The transformation $\mathcal{A}$ for g_{ij} is defined as

$$\begin{pmatrix} x^a_i \\ y^a_j \end{pmatrix} = \mathcal{A}_\Gamma(g_{ij}) = \begin{bmatrix}\gamma_3 cos(\gamma_4) & -\gamma_3 sin(\gamma_4) & \gamma_1\\ \gamma_3 sin(\gamma_4) & \gamma_3 cos(\gamma_4) & \gamma_2\end{bmatrix}\begin{pmatrix} x_i \\ y_j \\ 1 \end{pmatrix}. \tag{4}$$

With reduced parameters to learn, the AT module is easier to optimize than the ST module. Despite its simplicity, the AT module is capable of abundant attention transformations conditioned on the information from N_{cls}. G^r is defined as:

$$G^r = G - (G \cap G^a). \tag{5}$$

Restrictions on the attention parameters. We apply restrictions on the ART module to avoid meaningless transformation such as sampling largely outside the boundary of I_{in}. Since the coordinates of G, G^a and G^r are normalized to $[-1,1]$, the location factors γ_1,γ_2 and the scale factor γ_3 should be limited to a reasonable range. Otherwise, the transformation may be irreversibly meaningless and misdirect the optimization of the networks. Therefore, we apply a restriction on Γ as:

$$\Gamma_{res} = \begin{bmatrix} \gamma_1^{res} & \gamma_2^{res} & \gamma_3^{res} & \gamma_4^{res} \end{bmatrix} = \begin{bmatrix} \alpha_p tanh(\gamma_1) & \alpha_p tanh(\gamma_2) & \alpha_s tanh(\gamma_3) & \gamma_4 \end{bmatrix}, \tag{6}$$

where $\alpha_p \in [0,1]$, $\alpha_s \in [0,1]$ restrict $\gamma_1^{res},\gamma_2^{res} \in [-\alpha_p,\alpha_p]$ and $\gamma_3^{res} \in [-\alpha_s,\alpha_s]$. We keep $\gamma_4^{res} = \gamma_4$ because γ_4 is actually restricted by trigonometric functions. In practice, we use Γ_{res} instead of Γ to propagate attentional information between the streams, and Γ_{res} helps to prevent overfitting.

Moreover, we propose the attention-restriction loss to further restrict the transformation. Let $E = \begin{bmatrix} e_1,e_2,e_3,e_4 \end{bmatrix}$ is the expectation of Γ. That is, the regions obtained with the Γ equal to E can likely well represent the attentional information in the most average situation. Then the attention-restriction loss is defined as:

$$\begin{aligned} \mathcal{L}_{res} = (&\frac{max(0,|\gamma_1^{res}-e_1|-t_1)}{|\gamma_1^{res}-e_1|-t_1+eps}(\gamma_1^{res}-e_1)^2 + \frac{max(0,|\gamma_2^{res}-e_2|-t_2)}{|\gamma_2^{res}-e_2|-t_2+eps}(\gamma_2^{res}-e_2)^2 \\ &+\frac{max(0,|\gamma_3^{res}-e_3|-t_3)}{|\gamma_3^{res}-e_3|-t_3+eps}(\gamma_3^{res}-e_3)^2 + \frac{max(0,|\gamma_4-e_4|-t_4)}{|\gamma_4-e_4|-t_4+eps}(\gamma_4-e_4)^2)^{\frac{1}{2}}, \\ &where\ eps\ denotes\ the\ epsilon. \end{aligned} \tag{7}$$

Here, $T = \begin{bmatrix} t_1 & t_2 & t_3 & t_4 \end{bmatrix}$ is a set of thresholds. It is clear that $\mathcal{L}$ penalizes the distances between Γ and E if the distance is larger than the thresholds. In practice, we manually set E to be the half-length and half-width center crop. Unlike the previous studies that only learns location [7, 12, 16], our work learns various attentional information restricted by the thresholds.

3.2 Contrastive Reinforcement

Let (I_{in}^m, I_{in}^n) be a pair of input images and (a^m,a^n),(r^m,r^n) are respectively the attention and redundancy pairs of the image pairs. $\mathcal{L}_{cntr}$ is defined as:

$$\begin{aligned} \mathcal{L}_{cntr} = &max(d(f_{ar}(r^m),f_{ar}(r^n)) - d(f_{ar}(r^m),f_{ar}(a^m)) + margin,0) + \\ &max(d(f_{ar}(r^n),f_{ar}(r^m)) - d(f_{ar}(r^n),f_{ar}(a^n)) + margin,0). \end{aligned} \tag{8}$$

where d denotes the Euclidean distance and f_{ar} denotes the deep feature of N_{ar} (e.g., the output of the final fully connected layer in N_{ar}). In the training procedure, we minimize a multi-task objective function. The loss function $\mathcal{L}$ for an input image is defined as:

$$\mathcal{L} = \mathcal{L}_{cls1} + \mathcal{L}_{cls2} + \mathcal{L}_{res} + \mathcal{L}_{cntr}. \tag{9}$$

To minimize $\mathcal{L}_{cls2}$ and $\mathcal{L}_{cntr}$, the N_{cls} has to provide effective attentional information to the N_{ar} by the ART module, which forces N_{cls} to reinforce its awareness of attention. $\mathcal{L}_{res}$ ensures the transformation to be in a reasonable range. $\mathcal{L}_{res}$ penalizes the AT module if the transformation exceeds the range while giving full freedom otherwise. Moreover, $\mathcal{L}_{res}$ works as a noise to N_{cls}, and forces N_{cls} to continuously explore more attentional information, which helps prevent overfitting. In the testing procedure, we remove N_{ar} and the ART module, and only use the classifiers of N_{cls}. Therefore, our approach is as light as the basic backbone networks in the testing procedure.

4 Experiments

4.1 Experiments Setup

Datasets. We carried out comparison experiments on benchmark datasets CUB-200-2011 [33] and Stanford Cars [18]. CUB-200-2011 is a bird image dataset across 200 species with totally 11,788 images. Stanford Cars is a car image dataset contains totally 16,185 images across 196 car classes.

Baselines. We use as baseline the ResNet-50 and ResNet-101 [10] pre-trained on ImageNet [5] and fine-tuned on CUB-200-2011 and Stanford Cars. We fine-tunes ResNet-50 and ResNet-101 following the very careful fine-tuning techniques in [20]. For CRA-CNN, we adopt the ResNet-50 and ResNet-101 as the backbone and fine-tunes respectively on the two datasets. N_{cls} and N_{ar} always have the same network as the backbone. We add a fully-connected layer on the top of the last pooling layer of the ResNets to obtain Γ. As mentioned before, we remove the N_{ar} and ART module during testing. Therefore, in the testing stage, CRA-CNN has the same structure as the baselines.

Training details. As mentioned above, we fine-tune the baselines like [20], which has the best fine-tuning results to our knowledge. Therefore, here we mainly introduce the details of CRA-CNN. We manually set the margin in Equation (8) as 0.7 and $\alpha_p = 1$, $\alpha_s = 0.5$, $E = \begin{bmatrix}0 & 0 & 0.5 & 0\end{bmatrix}$, $T = \begin{bmatrix}0.4 & 0.4 & 0.4 & \pi\end{bmatrix}$. When training networks, we resize the input image so the shorter side is 512 but the aspect ratio does not change, then random crop a 448×448 region as the input. The 448×448 images are fed into N_{cls} for classification and obtaining the transformation parameters. We set the output size of the ART module as 224×224. The batch size for training is 64, and we set weight decay factor as 5×10^{-4}, momentum as 0.9. The initial learning rate is set as 10^{-3}, which decays to 10^{-4} after 50 epochs, and then decays by 10^{-1} for every 45 epochs. Moreover, after the first 50 epochs, we repeatedly turn off $\mathcal{L}_{res}$ for 45 epochs and then turn on $\mathcal{L}_{res}$ for 45 epochs. For validation, we first resize the images in the same way as training. Then we center crop the image (Subsection 4.2) or average the final outputs of classifiers without cropping (Subsection 4.3).

4.2 Comparison with the Baselines

Table 1: Comparison results with baselines.

	ResNet-50		ResNet-101	
	Baseline	CRA-CNN	Baseline	CRA-CNN
CUB-200-2011	84.2%	86.2%	86.1%	87.6%
Stanford Cars	90.0%	92.6%	91.8%	93.4%

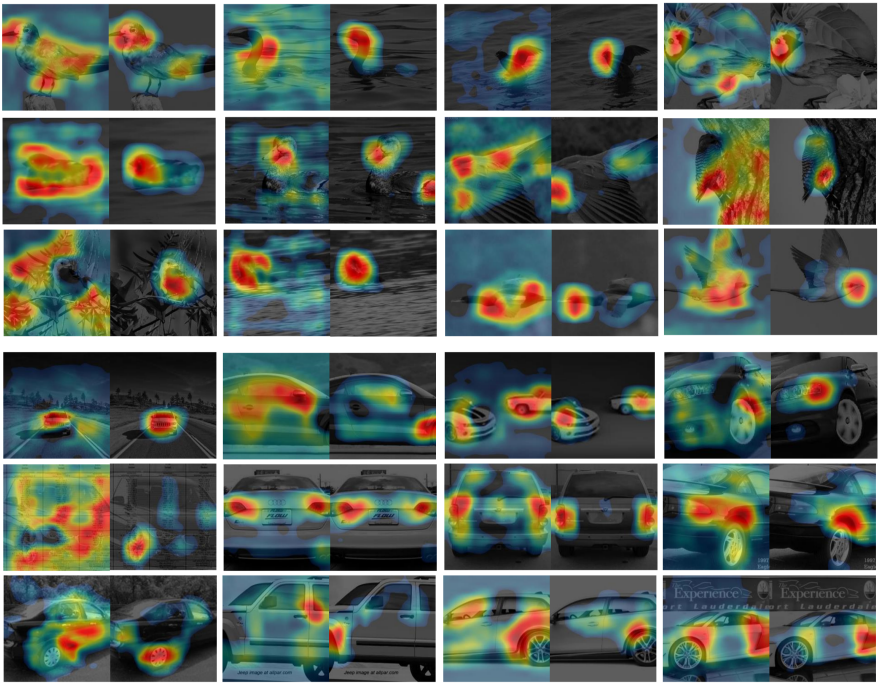

Figure 3: CAM maps respectively generated by the baseline ResNet-50 (the left image of each pair) and the CRA-CNN whose backbone is ResNet-50 (the right image of each pair). It is clear that our approach is much more focused than the baseline.

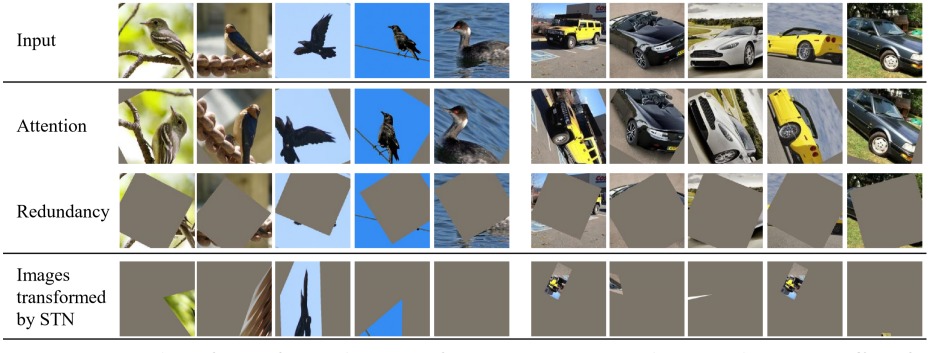

Figure 4: Examples of transformed images from CRA-CNN and STN. The STN suffers from training difficulty and cannot capture meaningful attention. The CRA-CNN tends to both capture and aligns the objects conditioned on certain attention focus, such as head, tire, etc.

Table 1 shows the comparison results between the CRA-CNN and baselines (our implementation). Clearly, our approach surpass the baselines in both datasets. It is noticeable that our work actually uses the same structure as baselines for testing without additional overhead, which indicates the significance of our work. Fig. 3 visualizes some example CAM maps respectively generated by the baseline ResNet-50 and CRA-CNN whose backbone is ResNet-50. It is obvious that CRA-CNN focus more on the core attention, while the baseline ResNet-50 tends to be distracted. This phenomenon indicates that: (a) the proposed approach efficiently forces the network to focus on core attention; (b) The core attention is very helpful.

Table 2: Comparison results on CUB.

STNs (4×ST-CNN 448px) [16]	84.1%
RA-CNN (scale 1+2+3) [7]	85.3%
Kernel Pooling [3]	86.2%
MA-CNN ($L_{cls} + L_{cng}$) [38]	86.5%
ResNet-101+OSME+MAMC [29]	86.5%
PC-DenseNet-161 [6]	86.9%
TASN [39]	87.9%
HSE [1]	88.1%
PAIRS[9]	89.2%
WS-DAN [13]	89.4%
Stacked LSTM [8]	90.4%
Ours(ResNet-50)	86.7%
Ours(ResNet-101)	88.3%

Table 3: Comparison results on Cars.

BoostCNN [24]	88.5%
Kernel Pooling [3]	92.0%
RA-CNN (scale 1+2+3) [7]	92.5%
MA-CNN ($L_{cls} + L_{cng}$) [38]	92.8%
ResNet-101+OSME+MAMC [29]	90.3%
MPN-COV [19]	93.3%
TASN [39]	93.8%
MGE-CNN [35]	93.9%
WS-DAN [13]	94.5%
EfficientNet [31]	94.7%
AutoAugment [2]	94.8%
Ours(ResNet-50)	93.3%
Ours(ResNet-101)	94.8%

Fig. 4 shows the examples of the original input, attention, and redundancy. Clearly, the CRA-CNN captures strong visual semantics. As the ART module is adapted from the ST module, we also train an STN as a reference. For a fair comparison, we do not fix the first four parameters of the ST module like [16] but optimize all the six parameters. Clearly, STN fails to propose any useful visual information in this circumstance, which suggests the extreme training difficulty of the STN. Actually, in practice, we observe that the STN refuses to converge at all if we do not fix the firs four parameters.

4.3 Comparison with State-of-art

In this subsection, we compare work work with the state-of-art studies on CUB-200-2011 and Stanford Cars. For the experiment in this subsection, we do not crop the resized images but average the final prediction scores. We find that the results improve in this way.

Table 2 shows comparison results with prior work on CUB-200-2011. The results suggest that work is quite close to the state-of-art studies. Although our best result is a little behind the state-of-the-art results, such as [8], our work is competitive because our work is quite easy to implement. [8] involves LSTMs and a Mask-RCNN that needs to be pretrained on additional data. The network optimization in [8] requires multiple complex stages and is computationally expensive (also pointed out in [10, 30]). [9] uses Earth Mover's Distance (EMD) to measure the distance between datasets, and requires much extra data for transferring the knowledge. In comparison, our work is simple and light, especially in the test procedure.

Table 3 shows comparison results with prior work on Stanford Cars, and our best result reaches the state-of-art result reported in [2]. [2] aims to find the best augmentation policy with a search scheme, which, however, is very computationally-expensive. Besides, as pointed out in [37], the policies acquired on proxy tasks may not be suitable for target tasks. In comparison, our work provides a computationally-affordable and effective solution.

5 Conclusion

In this paper, we propose Contrastively-reinforced Attention Convolutional Neural Network (CRA-CNN) to improve the attention awareness of deep activations. CRA-CNN consists of two streams that are connected by the proposed attention-redundancy transformer (ART) module. The attention regulation stream forces the classification stream to continuously explore core attention by evaluating the attention/redundancy information from the classification stream. Our work is simple to implement and computationally cheap. Despite its simplicity, our approach is very competitive with state-of-art studies in terms of accuracy.

References

[1] Tianshui Chen, Wenxi Wu, Yuefang Gao, Le Dong, Xiaonan Luo, and Liang Lin. Fine-grained representation learning and recognition by exploiting hierarchical semantic embedding. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 2023–2031, 2018.

[2] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 113–123, 2019.

[3] Yin Cui, Feng Zhou, Jiang Wang, Xiao Liu, Yuanqing Lin, and Serge Belongie. Kernel pooling for convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2930, 2017.

[4] Yin Cui, Yang Song, Chen Sun, Andrew Howard, and Serge Belongie. Large scale fine-grained categorization and domain-specific transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4109–4118, 2018.

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

[6] Abhimanyu Dubey, Otkrist Gupta, Pei Guo, Ramesh Raskar, Ryan Farrell, and Nikhil Naik. Pairwise confusion for fine-grained visual classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 70–86, 2018.

[7] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4438–4446, 2017.

[8] Weifeng Ge, Xiangru Lin, and Yizhou Yu. Weakly supervised complementary parts models for fine-grained image classification from the bottom up. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3034–3043, 2019.

[9] Pei Guo and Ryan Farrell. Aligned to the object, not to the image: A unified pose-aligned representation for fine-grained recognition. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1876–1885. IEEE, 2019.

[10] Harald Hanselmann and Hermann Ney. Elope: Fine-grained visual classification with efficient localization, pooling and embedding. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1247–1256, 2020.

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

[12] Xiangteng He, Yuxin Peng, and Junjie Zhao. Fast fine-grained image classification via weakly supervised discriminative localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(5):1394–1407, 2018.

[13] Tao Hu, Honggang Qi, Qingming Huang, and Yan Lu. See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification. *arXiv preprint arXiv:1901.09891*, 2019.

[14] Shaoli Huang, Zhe Xu, Dacheng Tao, and Ya Zhang. Part-stacked cnn for fine-grained visual categorization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1173–1182, 2016.

[16] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015.

[17] Zhong Ji, Yanwei Fu, Jichang Guo, Yanwei Pang, Zhongfei Mark Zhang, et al. Stacked semantics-guided attention model for fine-grained zero-shot learning. In *Advances in Neural Information Processing Systems*, pages 5995–6004, 2018.

[18] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013.

[19] Peihua Li, Jiangtao Xie, Qilong Wang, and Zilin Gao. Towards faster training of global covariance pooling networks by iterative matrix square root normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 947–955, 2018.

[20] Z. Li, Y. Yang, X. Liu, F. Zhou, S. Wen, and W. Xu. Dynamic computational time for visual attention. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 1199–1209, Oct 2017. doi: 10.1109/ICCVW.2017.145.

[21] Zhenyang Li, Kirill Gavrilyuk, Efstratios Gavves, Mihir Jain, and Cees GM Snoek. Videolstm convolves, attends and flows for action recognition. *Computer Vision and Image Understanding*, 166:41–50, 2018.

[22] Di Lin, Xiaoyong Shen, Cewu Lu, and Jiaya Jia. Deep lac: Deep localization, alignment and classification for fine-grained recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1666–1674, 2015.

[23] Pau Rodriguez Lopez, Diego Velazquez Dorta, Guillem Cucurull Preixens, Josep M Gonfaus Sitjes, Francesc Xavier Roca Marva, and Jordi Gonzalez. Pay attention to the activations: a modular attention mechanism for fine-grained image recognition. *IEEE Transactions on Multimedia*, 2019.

[24] Mohammad Moghimi, Serge J Belongie, Mohammad J Saberian, Jian Yang, Nuno Vasconcelos, and Li-Jia Li. Boosted convolutional neural networks. In *BMVC*, pages 24–1, 2016.

[25] Omkar M Parkhi, Andrea Vedaldi, CV Jawahar, and Andrew Zisserman. The truth about cats and dogs. In *2011 International Conference on Computer Vision*, pages 1427–1434. IEEE, 2011.

[26] Tianrong Rao, Xiaoxu Li, Haimin Zhang, and Min Xu. Multi-level region-based convolutional neural network for image emotion classification. *Neurocomputing*, 333: 429–439, 2019.

[27] Shikhar Sharma, Ryan Kiros, and Ruslan Salakhutdinov. Action recognition using visual attention. *arXiv preprint arXiv:1511.04119*, 2015.

[28] Chang Shu, Xi Chen, Chong Yu, and Hua Han. A refined spatial transformer network. In *International Conference on Neural Information Processing*, pages 151–161. Springer, 2018.

[29] Ming Sun, Yuchen Yuan, Feng Zhou, and Errui Ding. Multi-attention multi-class constraint for fine-grained image recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 805–821, 2018.

[30] Min Tan, Guijun Wang, Jian Zhou, Zhiyou Peng, and Meilian Zheng. Fine-grained classification via hierarchical bilinear pooling with aggregated slack mask. *IEEE Access*, 7:117944–117953, 2019.

[31] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[33] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, 2011.

[34] Han Zhang, Tao Xu, Mohamed Elhoseiny, Xiaolei Huang, Shaoting Zhang, Ahmed Elgammal, and Dimitris Metaxas. Spda-cnn: Unifying semantic part detection and abstraction for fine-grained recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1143–1152, 2016.

[35] Lianbo Zhang, Shaoli Huang, Wei Liu, and Dacheng Tao. Learning a mixture of granularity-specific experts for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8331–8340, 2019.

[36] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. Part-based r-cnns for fine-grained category detection. In *European conference on computer vision*, pages 834–849. Springer, 2014.

[37] Xinyu Zhang, Qiang Wang, Jian Zhang, and Zhao Zhong. Adversarial autoaugment. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ByxdUySKvS.

[38] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 5209–5217, 2017.

[39] Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5012–5021, 2019.

[40] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.