Destruction and Construction Learning for Fine-grained Image Recognition

Yue Chen[1]* Yalong Bai[2]* Wei Zhang[3] Tao Mei[4]
JD AI Research, Beijing, China
[1]chenyue21@jd.com, [2]ylbai@outlook.com, [3]wzhang.cu@gmail.com, [4]tmei@live.com

Delicate feature representation about object parts plays a critical role in fine-grained recognition. For example, experts can even distinguish fine-grained objects relying only on object parts according to professional knowledge. In this paper, we propose a novel "Destruction and Construction Learning" (DCL) method to enhance the difficulty of fine-grained recognition and exercise the classification model to acquire expert knowledge. Besides the standard classification backbone network, another "destruction and construction" stream is introduced to carefully "destruct" and then "reconstruct" the input image, for learning discriminative regions and features. More specifically, for "destruction", we first partition the input image into local regions and then shuffle them by a Region Confusion Mechanism (RCM). To correctly recognize these destructed images, the classification network has to pay more attention to discrimina- tive regions for spotting the differences. To compensate the noises introduced by RCM, an adversarial loss, which distinguishes original images from destructed ones, is applied to reject noisy patterns introduced by RCM. For "construction", a region alignment network, which tries to restore the original spatial layout of local regions, is followed to model the semantic correlation among local regions. By jointly training with parameter sharing, our proposed DCL injects more discriminative local details to the classification network. Experimental results show that our proposed framework achieves state-of-the-art performance on three standard benchmarks. Moreover, our proposed method does not need any external knowledge during training, and there is no computation overhead at inference time except the standard classification network feed-forwarding. Source code: https://github.com/JDAI-CV/DCL.

1. Introduction

In the past decade, generic object recognition has achieved steady progress with efforts from both large-scale annotated dataset and sophisticated model design. However, recognizing fine-grained object categories (e.g., bird species [3], car models [14] and aircraft [18]) is still a challenging task, which attracts extensive research attention. Although fine-grained objects are visually similar by a rough glimpse, they can be correctly recognized by details in discriminative local regions.

Learning discriminative feature representations from discriminative parts plays the key role in fine-grained image recognition. Existing fine-grained recognition methods can be roughly grouped into two categories, as illustrated in Figure 1. One group (a) first locates the discriminative object parts and then classifies based on the discriminative regions. These two-steps methods [21, 11, 1] mostly need additional bounding box annotations on objects or parts, which are expensive to collect. The other group (b) tries to automatically localize discriminative regions by attention mechanism in an unsupervised manner, and thus does not needs extra annotations. However, these methods [7, 42, 41, 22] usually need additional network structure (e.g., attention mechanism), and thus introduce extra computation overhead for both training and inference stages.

In this paper, we propose a novel fine-grained image recognition framework named "Destruction and Construction Learning" (DCL), as shown in Figure 1 (c). Besides the standard classification backbone network, we introduce a DCL stream to learn from discriminative regions automatically. An input image is first carefully *destructed* to emphasize discriminative local details, and then *reconstructed* to model the semantic correlation among local regions. On one hand, DCL automatically localizes discriminative regions, and thus does not need any extra knowledge while training. On the other hand, the DCL structure is only adopted at the training stage, and thus introduces no computational overhead at inference time.

For "Destruction", we propose a Region Confusion Mechanism (RCM) to deliberately "confuse" the global structure, which partitions the input image into local patches and then shuffles them randomly (Figure 3). For fine-grained recognition, local details play a more important role than global structures, since images from different fine-

*Equal contribution.

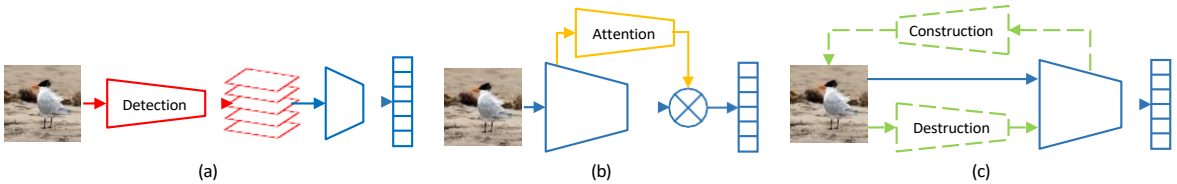

Figure 1. Illustrations of two previous general frameworks (a,b) and our proposed framework (c) for fine-grained classification. (a) Two-stage part detection based framework. (b) Attention based framework. (c) Our proposed destruction and construction learning framework. The network structures in dashed lines are disabled during inference.

grained categories usually share the same global structure or shape, but only differ in local details. Discarding global structure and keeping local details could force the network to identify and focus on the discriminative local regions for recognition. After all, the devil is in the details. Shuffling is also adopted in natural language processing [15] to let the neural network focus on discriminative words. Simi- larly, if local regions in an image are "shuffled", irrelevant regions that are non-critical to fine-grained recognition will be neglected, and the network will be forced to classify images based on the discriminative local details. With RCM, the visual appearance of the image has been substantially changed. As shown in the bottom row of Figure 3, though it becomes more difficult for recognition, bird experts can still spot the difference easily. Car enthusiasts can distinguish car models by only examining parts of car [34]. Similarly, the neural network also needs to learn expert knowledge to classify the destructed images.

It is worth noting that "destruction" is not always beneficial. As a side effect, RCM introduces several noisy visual patterns as in Figure 3. To offset the negative impact, we apply an adversarial loss to distinguish original images from destructed ones. As a result, the effect of noisy patterns can be minimized, keeping only beneficial local details. Conceptually, the adversarial and classification losses work in an adversarial manner to carefully learn from "destruction".

For "Construction", a region alignment network is introduced to restore the original region arrangement, which acts in the opposite way of RCM. By learning to restore the original layout as in [19, 6], the network needs to understand the semantics of each region, including those discriminative ones. Through "construction", the correlation between different local regions can be modeled.

The main contributions are summarized as follows:

- A novel "Destruction and Construction Learning (DCL)" framework is proposed for fine-grained recognition. For destruction, the region confusion mechanism (RCM) forces the classification network to learn from discriminative regions, and the adversarial loss prevents over-fitting the RCM-induced noisy patterns. For construction, the region alignment network restores the original region layout by modeling the semantic correlation among regions.
- State-of-the-art performances are reported on three standard benchmark datasets, where our DCL consistently outperforms existing methods.
- Compared to existing methods, our proposed DCL does not need extra part/object annotation and introduces no computational overhead at inference time.

2. Related works

Researches for fine-grained image recognition task mainly proceed along two dimensions. One is learning better visual representations from the original image directly [26, 25, 28] and the other is using part/attention based methods [41, 42, 7, 13] to obtain discriminative regions in images and learn region-based feature representations.

Due to the success of deep learning, fine-grained recognition methods have shifted from multistage frameworks based on hand-crafted features [39, 36, 23, 10] to multi-stage frameworks with CNN features [13, 31, 29]. Sec- ond order bilinear feature interactions were shown to have a significant improvement for visual representations learning [16, 30]. This method was later extended to a series of related works with further improvements [12, 4, 8]. Deep metric learning is also used to capture subtle visual differences. Zhang *et al*. [40] introduced label structures and a generalization of triplet loss to learn fine-grained feature representations. Chen *et al*. [27] investigate simultaneously predicting categories of different levels in the hierarchy and integrating this structured correlation information into the network by an embedding method. However, these pairwise neural network models often bring complex network computing.

There is also a large amount of part localization based methods proposed regarding the theory that the object parts are essential to learning discriminative features for fine-grained classification [32]. Fu *et al*. [7] proposed a reinforced attention proposal network to obtain discriminating attention regions and region-based feature representation of multiple scales. Sun *et al*. [20] proposed a one-squeeze multi-excitation module to learns multiple attention region features of each input image, and then apply a multi-attention multi-class constraint in a metric learning framework. Zheng *et al*. [42] adopted a channel grouping network to generate multiple parts by clustering, then classi-

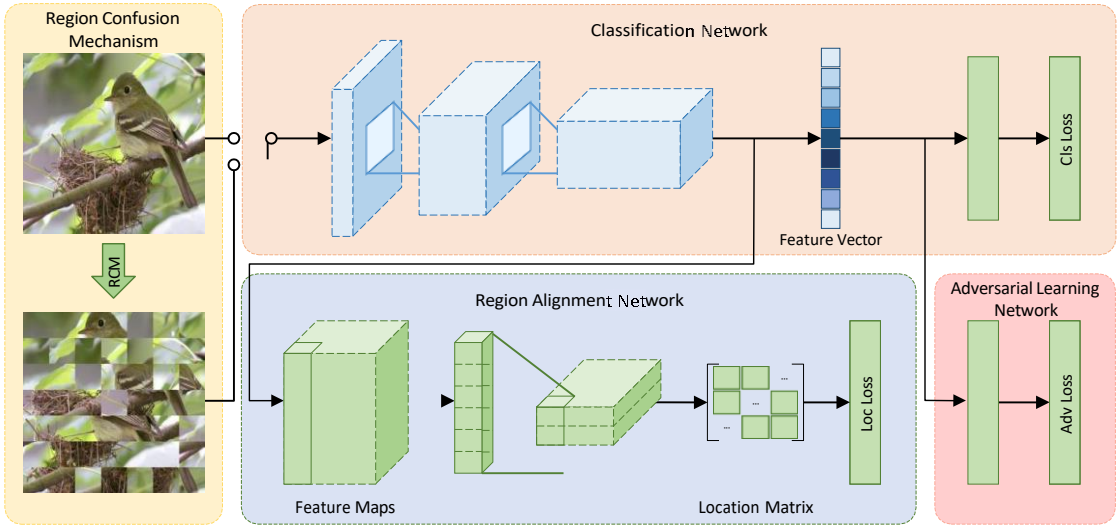

Figure 2. The framework of the proposed DCL method which consists of four parts. (1) Region Confusion Mechanism: a module to shuffle the local regions of the input image. (2) Classification Network: the backbone classification network that classifies images into fine-grained categories. (3) Adversarial Learning Network: an adversarial loss is applied to distinguish original images from destructed ones. (4) Region Alignment Network: appended after the classification network to recover the spatial layout of local regions.

fied these parts features to predict the categories of input images. Compared with earlier part/attention based methods, some of the recent methods tend to be weak supervised and do not require the annotations of parts or key areas [21, 35]. In particular, Peng *et al*. [21] proposed a part spatial constraint to make sure that the model could select discriminative regions, and a specialized clustering algorithm is used to integrate the features of these regions. Yang *et al*. [35] introduced a method to detect informative regions and then scrutinizes them to make final predictions. However, the correlation among regions is helpful to build deep understanding about the objects, it is usually ignored by previous works. The research [19] also shows that utilizing the location information of regions can enhance the visual representation ability of the neural network and result in improving performance on classification and detection tasks.

Our proposed method differs previous works in three aspects: First, by training classifier with our proposed RCM, the discriminative regions can be automatically detected without using any prior knowledge except object labels. Second, our formulation considers not only the fine-grained local region feature representations but also the semantic correlation among different regions in the whole image. Third, our proposed method is highly efficient, that there is no additional overhead except backbone network feed-forward in prediction time.

3. Proposed Method

In this section, we present our proposed Destruction and Construction Learning (DCL) method. As shown in Figure 2, the whole framework is composed of four parts. Please note that only the "classification network" is needed during inference time.

3.1. Destruction Learning

The devil is in the details. For fine-grained image recognition, local details are much more important than the global structure. In most cases, different fine-grained categories usually share a similar global structure and only differ in certain local details. In this work, we propose to carefully destruct the global structure by shuffling the local regions for better identifying discriminative regions and learning discriminative features (Section 3.1.1). To prevent the network learning from noisy patterns introduced by destruction, an adversarial counterpart (Section 3.1.2) is proposed to reject RCM-induced patterns that are irrelevant to fine-grained classification.

3.1.1 Region Confusion Mechanism

As an analogy [15] to natural language processing, shuffling words in a sentence would force the neural network to focus on discriminative words and neglect irrelevant ones. Similarly, if local regions in an image are "shuffled", the neural network would be forced to learn from discriminative region details for classification.

As shown in Figure 3, our proposed Region Confusion Mechanism (RCM) is designed to disrupt the spatial layout of local image regions. Given an input image I, we first uniformly partition the image into $N \times N$ sub-regions denoted by $R_{i,j}$, where i and j are the horizontal and vertical indices respectively and $1 \leq i, j \leq N$. Inspired by [15], our proposed RCM shuffles these partitioned local regions

in their 2D neighbourhood. For the j^{th} row of R, a random vector q_j of size N is generated, where the i^{th} element $q_{j,i} = i + r$, where $r \sim U(-k, k)$ is a random variable following a uniform distribution in the range of $[-k, k]$. Here, k is a tunable parameter ($1 \leq k < N$) defining the neighbourhood range. Then we can get a new permutation σ_j^{row} of regions in j^{th} row by sorting the array q_j, verifying the condition:

$$\forall i \in \{1, ..., N\},\ \sigma_j^{row}(i) - i\ < 2k. \tag{1}$$

Similarly, we apply the permutation σ_i^{col} to the regions column-wisely, verifying the condition:

$$\forall j \in \{1, ..., N\},\ \sigma_i^{col}(j) - j\ < 2k. \tag{2}$$

Therefore, the region at (i, j) in original region location is placed to a new coordinate:

$$\sigma(i, j) = (\sigma_j^{row}(i), \sigma_i^{col}(j)). \tag{3}$$

This shuffling method destructs the global structure and ensures that the local region jitters inside its neighbourhood with a tunable size.

The original image I, its destructed version $\varphi(I)$ and its ground truth one-vs-all label l indicating the fine-grained categories are coupled as $\langle I, \varphi(I), l \rangle$ for training. The classification network maps input image into a probability distribution vector $C(I, \theta_{cls})$, where θ_{cls} is all learnable parameters in the classification network. The loss function of the classification network L_{cls} can be written as:

$$L_{cls} = -\sum_{I \in \mathbf{I}} l \cdot \log [C(I)\, C(\varphi(I))], \tag{4}$$

where $\mathbf{I}$ is the image set for training.

Since the global structure has been destructed, to recognize these randomly shuffled images, the classification network has to find the discriminative regions and learn the delicate differences among categories.

3.1.2 Adversarial Learning

Destructing images with RCM does not always bring beneficial information for fine-grained classification. For example in Figure 3, RCM also introduces noisy visual patterns as we shuffle the local regions. Features learned from these noise visual patterns are harmful to the classification task. To this end, we propose another adversarial loss L_{Adv} to prevent overfitting the RCM-induced noise patterns from creeping into the feature space.

Considering the original images and the destructed ones as two domains, the adversarial loss and classification loss work in an adversarial manner to 1) keep domain-invariant patterns, and 2) reject domain-specific patterns between I and $\varphi(I)$.

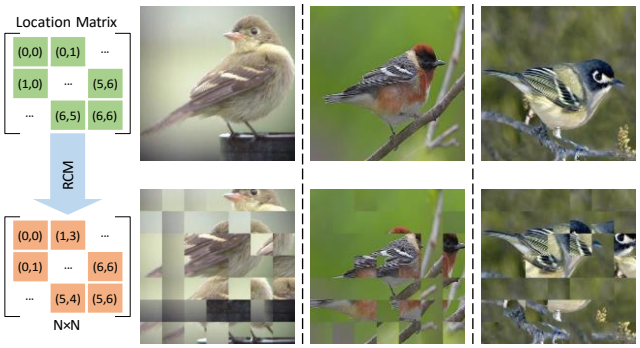

Figure 3. Example images for fine-grained recognition (top) and the corresponding "destructed" images by our proposed RCM (bottom).

We label each image as a one-hot vector $d \in \{0, 1\}^2$ indicating whether the image is destructed or not. A discriminator can be added as a new branch in the framework to judge whether an image I is destructed or not by:

$$D(I, \theta_{adv}) = \text{softmax}(\theta_{adv} C(I, \theta_{cls}^{[1,m]})), \tag{5}$$

where $C(I, \theta_{cls}^{[1,m]})$ is the feature vector extract from the outputs of the m^{th} layer in backbone classification network, $\theta_{cls}^{[1,m]}$ is the learnable parameters from the 1^{st} layer to m^{th} layer in the classification network, and $\theta_{adv} \in \mathbb{R}^{d \times 2}$ is a linear mapping. The loss of the discriminator network L_{adv} can be computed as:

$$L_{adv} = -\sum_{I \in \mathbf{I}} d \cdot \log [D(I)] + (1 - d) \cdot \log [D(\varphi(I))]. \tag{6}$$

Justification. To better understand how the adversarial loss tunes feature learning, we further visualize the features of backbone network ResNet-50 with and without the adversarial loss. Given an input image I, we denote the k^{th} feature map in m^{th} layer by $F_m^k(I)$. For ResNet-50, we extract feature from the outputs of the convolutional layer with average pooling next to the last fully-connect layer for adversarial learning. Thus, the response of k^{th} filter in the last convolutional layer for ground truth label c can be measured by $r^k(I, c) = \bar{F}_m^k(I) \times \theta_{cls}^{[m+1]}[k, c]$, where $\theta_{cls}^{[m+1]}[k, c]$ is the weight between the k^{th} feature map and the c^{th} output label.

We compare the responses of different filters for original image and its destructed version in scatter plot shown as Figure 4, where every filter with positive response is mapped to the data point $(r(I, c), r(\varphi(I), c))$ in the scatter plot. We can find that the distributions of feature maps trained by L_{cls} is more compact than those trained by $L_{cls} + L_{adv}$. It means that the filters have large responses on the noise patterns introduced by RCM may also have large responses on the original image (as the visual patterns visualized in A, B and C, there are lots of filters responding