Attend in groups: a weakly-supervised deep learning framework for learning from web data

Bohan Zhuang*, Lingqiao Liu*, Yao Li, Chunhua Shen†, Ian Reid
The University of Adelaide, Australia; and Australian Centre for Robotic Vision

Abstract

Large-scale datasets have driven the rapid development of deep neural networks for visual recognition. However, annotating a massive dataset is expensive and time-consuming. Web images and their labels are, in comparison, much easier to obtain, but direct training on such automatically harvested images can lead to unsatisfactory performance, because the noisy labels of Web images adversely affect the learned recognition models. To address this drawback we propose an end-to-end weakly-supervised deep learning framework which is robust to the label noise in Web images. The proposed framework relies on two unified strategies – random grouping and attention – to effectively reduce the negative impact of noisy web image annotations. Specifically, random grouping stacks multiple images into a single training instance and thus increases the labeling accuracy at the instance level. Attention, on the other hand, suppresses the noisy signals from both incorrectly labeled images and less discriminative image regions. By conducting intensive experiments on two challenging datasets, including a newly collected fine-grained dataset with Web images of different car models, [1]*, the superior performance of the proposed methods over competitive baselines is clearly demonstrated.*

1. Introduction

Recent development of deep convolutional neural networks (CNNs) has led to great success in a variety of tasks including image classification [17, 19, 38], object detection [15, 22, 33], semantic segmentation [20, 23] and others. This success is largely driven by the availability of large-scale well-annotated image datasets, e.g. ImageNet [36], MS COCO [21] and PASCAL VOC [13]. However, annotating a massive number of images is extremely labor-intensive and costly. To reduce the annotating labor cost, an alternative approach is to obtain the image annotations directly from the image search engine from the Internet, e.g. Google image search or Bing images.

Web-scale image search engine mostly uses keywords as queries and the connection between keywords and images is established by the co-occurrence between the Web image and its surrounding text. Thus, the annotations of Web images returned by a search engine will be inevitably noisy since the query keywords may not be consistent with the visual content of target images. For example, using "black swan" as a query keyword, the retrieved images may contain "white swan," "swan painting" and some other different categories. These noisy labels can be misleading if we use them to train a classifier to learn the corresponding visual concept.

To overcome this drawback, we propose a deep learning framework designed to be more robust to the labeling noise and thus better able to leverage Web images for training. There are two key strategies in our framework: random grouping and attention. As will be shown later, these two strategies seamlessly work together to reduce the negative impact of label noise.

Specifically, the random grouping strategy randomly samples a few images and merges them into a single training instance. The idea is that although the probability of sampling an incorrectly labeled Web image is high, the probability of sampling an incorrectly labeled group is low because as long as one image in the group is correctly labeled, the label of the group is deemed correct (bag label as in multi-instance learning). In the proposed approach, each image is represented by the extracted contextual features depicting the visual patterns of local image regions. After the random grouping, a training instance is represented as the union of convolutional feature maps extracted from each image in the group. If there are any incorrectly labeled images in the group, the unified feature maps of an instance will contain a substantial amount of local features which are irrelevant to the group-level class annotation. To avoid the distraction of those local features, we apply the second strategy of our framework, the attention mechanism, to encourage the network not to focus on the irrelevant features.

*The first two authors contributed equally to this work.

†C. Shen is the corresponding author.

[1] The dataset and code are available at https://bitbucket.org/jingruixiaozhuang/cvpr2017_code_dataset/

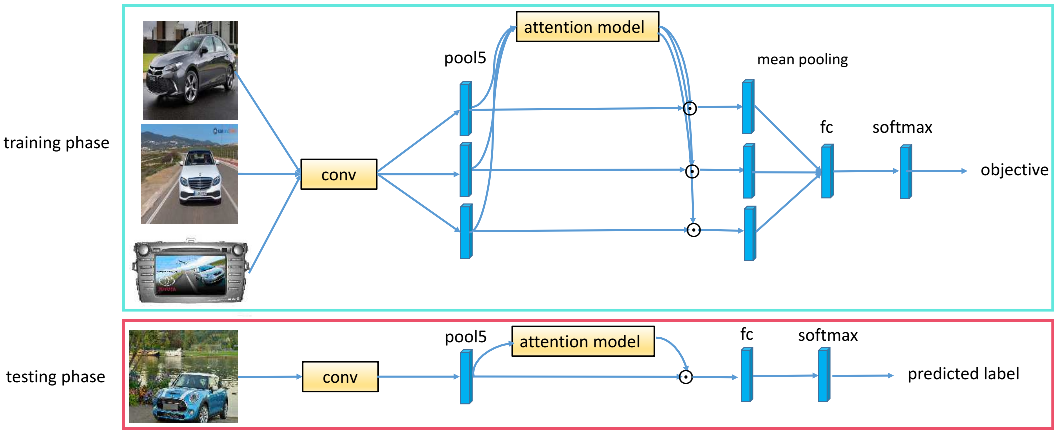

Figure 1: Overview of our "webly"-supervised learning pipeline. For the training phase, inputs are a group of images, including one correctly labeled image and two noise images from top to bottom. The convolutional layers are shared. The attention model is added on each training data and followed by a global average pooling layer to get the aggregated group-level representation, followed by a softmax layer for classification. For the testing phase, the input is a single image and output is the predicted class label.

To experimentally validate the robustness of the proposed method, we collect a large-scale car dataset using a Web image search engine. This dataset is particularly challenging due to its fine-grained nature. By conducting an experimental comparison on this dataset, we demonstrate that the proposed method achieves significantly better performance than competitive approaches.

2. Related Work

Our work is closely related to learning from web-scale data and noisy labels [8,9,12,14,18,29,30,32,37,41,45,48]. In terms of learning from Web data, in [8, 9], Chen *et al.* propose to pretrain CNN on simple examples and adapt it to harder images by leveraging the structure of data and categories in a two-step manner. In contrast, we propose a simply-yet-effective end-to-end learning framework without pretraining. To better dealing with noise, some approaches [40, 45] propose to add an extra noise layer into the network which adapts the network outputs to match the noisy label distribution. On the other hand, some approaches attempt to remove or correct noisy labels [7, 27], but because of the difficulty of separating correctly labeled hard samples from mislabeled ones, such a strategy can result in removing too many (correct) instances. Moreover, several label noise-robust algorithms [5, 26] are proposed to make classifiers robust to label noise. However, noise-robust methods seem to be adequate only for simple cases of label noise that can be safely managed by regularization. In this paper, we instead propose to suppress label noise by unified two strategies without any strong assumptions.

Our work is also related to weakly-supervised object localization [6, 10, 11, 31, 34, 35, 39, 42]. The objective of these methods is to localize object parts that are visually consistent with the semantic image-level labels across the training data. A major approach for tackling this task is to formulate it as a multiple instance learning problem. In these methods [10, 34, 44], each image is modeled as a bag of instances (region features) and the classifier is learned to select the foreground instances. Further in [31], a weakly-supervised deep learning pipeline is proposed to localize objects from complex cluttered scenes by explictly searching over possible object locations and scales in the image. In light of the above methods, we convert the problem of learning from noisy labels to a weakly-supervised problem that in spirit is similar to the multiple instance learning assumption. What's more, we further propose to incorporate attention strategy to reduce the adverse effect of noise.

Related to our work, the attentive mechanisms have been applied to many computer vision tasks [1–3, 16, 24, 28, 43, 46, 47, 50] to help improve the performance. To guide the models' focus on the objects specified by the question or caption, attention models are designed to pay attention to local CNN features in the input image [1,24,46,47,50]. The attentive mechanism has also been used to handle sequential problems in neural machine translation [3, 25] and manage memory access mechanisms for memory networks [43] and neural turing machines [16]. Different from the above methods, we are the first to apply the attention mechanism to cope with noisy labels. It can not only detect discriminative local feature regions, but also serves to filter out noisy signals from the mislabeled samples in the training instance.

3. Method

In our task, we intend to distill useful visual knowledge from the noisy Web data. It consists of correctly labeled samples and mislabeled samples on the Web. To make the classifier robust to noisy labels, we propose a deep learning framework by incorporating two strategies, random group training, and attention. The overview of our method is shown in Figure 1. At the training stage, we randomly group multiple training images into a single training instance as the input of our neural network. The proposed neural network architecture has two parts. The first part is similar to a standard convolutional neural network which is comprised of multiple convolutional layers and pooling layers. The second part is an attentional pooling layer which selects parts of the neuron activations and pools the activations into the instance-level representation. Once the neural network is trained, we can drop off the random grouping module and takes a single image as input at the test stage.

In the following sections, we will elaborate the random grouping training and the attention module and discuss their benefits for reducing the impact of noisy labels.

3.1. Random grouping training

Random grouping training (RGT) aims at reducing the probability of sampling an incorrectly labeled instance and thus mitigate the risk confusing a neural work with wrong annotations. The idea of RGT is to stack multiple images of one class into a single grouped training instance of the same class. In practice, we implement this idea by stacking the last layer convolutional feature maps obtained from each image into a unified convolutional feature map and perform (attention based) pooling on this feature map to obtain the instance-level representation. In this sense, we can view the input of a grouped instance as a "merged image" and as long as one image is correctly labeled as containing the object-of-interest, the "merged image" indeed contains it. In other words, the grouped training instance is correctly labeled as long as one image within is correctly labeled.

Consequently, if the probability of sampling an incorrectly labeled image is ξ, then the probability of sampling a correctly labeled grouped instance will become

$$p = 1 - \xi^K \tag{1}$$

where K is the group size and when K becomes larger, the probability of sampling a correctly labeled instance will become very high. For example, if $\xi = 0.2$ and $K = 3$, p will be greater than 99%. However, when K becomes larger, the independence between multiple training instances will reduce and this tends to undermine the network training. Thus in practice, we choose K as a small value (2 to 5). We have conducted an experimental study on the impact of K with respect to different level of labeling noise at Section 4.4.

3.2. Attention

3.2.1 Attention formulation

After random grouping, each instance is now represented as an array of activations. These activations come from both correctly labeled images and mislabeled images. Although containing activations from the correct region of interest, many of the activations are noisy signals and will negatively impact the learning process. To mitigate this issue, we propose to use an attention model to focus processing only on the attended activations. Let $\mathbf{x}^n_{ijk} \in \mathbb{R}^c$ denote the last convolutional layer activations from the k-th image of the n-th instance at the spatial location (i, j), where $i = 1, 2, ..., d$ and $j = 1, 2, ..., d$ are the coordinates of the feature map and d is the height or width of the feature map.

The unnormalized attention score $s^n_{ijk} \in \mathbb{R}$ can be formulated as

$$s^n_{ijk} = f(\mathbf{w}^{\mathrm{T}}\mathbf{x}^n_{ijk} + b), \tag{2}$$

where $\mathbf{w} \in \mathbb{R}^c$, $b \in \mathbb{R}^1$ denote the weight and bias of the attention detector respectively, which are parts of the model parameters and will be learned in an end-to-end manner. $f(\cdot)$ is the softplus function $f(x) = \ln(1 + \exp(x))$. Since we are only concerned with the relative importance of the local features within an image, we propose to normalize the attention scores to $[0, 1]$ for aggregating the local features:

$$a^n_{ijk} = \frac{s^n_{ijk} + \varepsilon}{\sum\limits_i \sum\limits_j (s^n_{ijk} + \varepsilon)}, \tag{3}$$

where a^n_{ijk} is the normalized attention score, ε is a small constant and quite important to make the distribution reasonable.

If the element s^n_{ijk} is low but there is no ε, then the corresponding a^n_{ijk} can be large even though s^n_{ijk} is small. The constant ε can solve this problem effectively. If it is properly set, a small s^n_{ijk} (approaching zero) will result in $a^n_{ijk} = \frac{1}{d^2}$. In our work, we set it to 0.1.

After obtaining the normalized attention scores, we can get the attended feature representation by applying a^n_{ijk} to $\mathbf{x}^n_{ijk}$ as follows:

$$\widehat{\mathbf{x}}^n_{ijk} = a^n_{ijk} \odot \mathbf{x}^n_{ijk}, \tag{4}$$

where $\odot$ is the element-wise multiplication, $\widehat{\mathbf{x}}^n_{ijk}$ is the attended feature representation.

Then the representation of a grouped training instance can be obtained by a global average pooling over all the feature dimensions except for the channel-wise dimension:

$$\mathbf{h}_n = \frac{1}{d^2 k} \sum_i \sum_j \sum_k \widehat{\mathbf{x}}^n_{ijk}, \tag{5}$$

where $\mathbf{h}_n \in \mathbb{R}^c$ is the group-level representation of the n-th training instance.

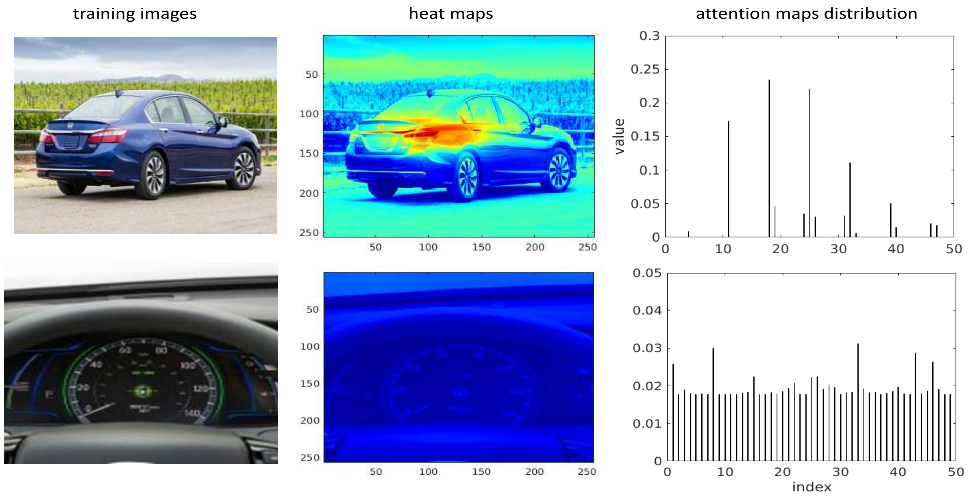

Figure 2: This figure illustrates the effectiveness of the group-wise attention model used in the proposed method. The left column shows the original training images. The middle column is the images plus its corresponding attention heat maps. The right column shows the distribution of the attention maps. The upper row relates to the correctly labeled sample and the bottom row corresponds to the mislabeled sample. We can see that for the correctly labeled sample, the normalized attention model only focus on the discriminative local parts and the score distribution is sparse. In contract, for the mislabeled sample, the normalized attention model fails to concentrate on any local regions and the score distribution is dense.

Then we apply a linear classifier layer to predict the class label of each grouped instance and use the multi-class cross-entropy loss to train the network:

$$L_{class} = -\sum_{n} y_n \log(\frac{\exp(\mathbf{F}_n)}{\sum_{n} \exp(\mathbf{F}_n)}) \tag{6}$$

where $\mathbf{F}_n$ and y_n are the last linear classification layer and the class label for the n-th training instance, respectively.

3.2.2 Attention module regularization

Ideally, for the correctly labeled image, the attention scores should have large values on one or few image regions; for the mislabeled image, none of the image regions should correspond to large attention values. In the above framework, we expect this situation can happen after the end-to-end training of the network. In this section, we devise a regularization term to further encourage this property. To apply this regularization, we assume that a set of negative class images belonging to none of to-be-learned image categories is available. Then we can apply the attention detector on those negative class images and require that the obtained normalized attention values are as small as possible since those images do not contain the object-of-interest. Define $u^n_{ijk} = \mathbf{w}^{\mathrm{T}}\mathbf{x}^n_{ijk} + b$ to be the linear attention scores for the sample $\mathbf{x}^n_{ijk}$; then the above requirement is equivalent to expecting $\max_{ijk} u^n_{ijk} < 0$. On the other hand, for a grouped training instance generated from each class, we expect that the attention detector identifies at least one relevant region and this leads to the objective $\max_{ijk} u^n_{ijk} > 0$. In this paper, we propose to use the following objective function to impose the aforementioned two requirements:

$$R(\mathbf{w}, b) = \sum_{n} \max(0,\ 1 - \delta_n \max_{ijk}(u^n_{ijk})) \tag{7}$$

where $\delta_n = \{1, -1\}$ indicates whether the instance is sampled from the classes of object-of-interest or from the negative class. We then use the weighted sum of L_{class} and R as the final objective function:

$$L = L_{class} + \lambda R. \tag{8}$$

The effect of the attention module is illustrated in Figure 2. The input is an instance including a correctly labeled car sample and a mislabeled noise sample. We can observe that for the correctly labeled sample, the normalized attention scores are pushed high at the region-of-interest, which corresponds to the back of the car in the example. In contrast, for the mislabeled sample, the normalized attention scores are all pushed approaching zero, resulting in no parts to be concentrated on for the attention model. In terms of this observation, we can explore that the attention model can not only filter out the contextual features of the mislabeled samples in the training instance, but also help detect the discriminative parts of the correctly labeled samples.

4. Experiments

In this section, we test our weakly-supervised learning framework on two datasets collected from the Web. One is a fine-grained dataset and the other one is a conventional classification dataset. The training data for both tasks are obtained via search results freely available from Google image search, using all returned images as training data. It's worth noticing that fine-grained classification is quite challenging because categories can only be discriminated by subtle and local differences.

4.1. Datasets

WebCars: We collect a large-scale fine-grained car dataset from the internet, named WebCars, using the categories of the clean CompCars dataset [49]. We treat the car model names as the query keywords and automatically retrieve images for all the 431 fine-grained categories. We collect 213,072 noisy Web images in total and still use the test set of the original clean dataset for testing. We sample a few categories from WebCars and manually annotate the ground-truth labels, noting in the process that approximately 30% of images are outliers. We further collect 10,000 images that doesn't belong to the training categories as the negative class.

Web data + ImageNet: We randomly sample 100 classes used in ImageNet and use the category names for collecting a noisy Web image dataset. All the images are automatically downloaded and the ones that appear in the original ImageNet dataset are manually removed. This dataset contains 61,639 images in total. The noise gradually increases from the highly ranked images to the latter samples. We estimate the percentage of mislabeled samples is approximately 20 %. We also collect 5,000 negative class Web images.

4.2. Implementation details

We use Theano [4] for our experiments. We use the pretrained VGG-16 model trained on the ImageNet dataset [36] to initialize the convolutional layers of our framework. The learning rate is set to 0.001 initially, and divided by 10 after 5 epoches. The regularizer λ is set to 0.1. Training samples are randomly grouped online.

To investigate the impact of the various elements in our end-to-end framework, we analyse the effects of the attention model, group-wise training approach and the attention regularization described in Section 3.2.2 independently.

1. "Average pooling without attention (AP)": We employ the average-pooling method as an important baseline here since it's commonly used for image classification on clean images without any noise-robust strategy. The average pooling structure simply replaces the two 4096 dimensions fully-connected layers in VGG-16 model with an average pooling layer, followed by a softmax layer for classification.

2. "Random grouping training without attention (RGT)": In this method, samples are randomly grouped during training, with the mean-pooling operation in Eq. 5 to get the instance-level representation.

3. "Average pooling with attention (AP+AT)": Based on AP, the attention model is embedded in the network to test its ability to localize discriminative feature regions.

4. "Random grouping training with attention (RGT+AT)": Attention is added to RGT.

5. "Average pooling with attention and regularizer (AP+AT+R)": We add the regularizer to AP+AT to evaluate its influence to cope with noisy labels.

6. "Random grouping training with attention and regularizer (RGT+AT+R)": We test its performance on filtering out incorrectly labeled samples in each group as well as noisy local feature parts by adding the regularizer to RGT+AT.

4.3. Evaluation on the WebCars

We quantatively compare the methods described in Section 4.2 and report the results in Table 1. For RGT based methods, the group size is set to 2.

methods	accuracy
AP	66.86%
RGT	69.83%
AP+AT	73.64%
RGT+AT	76.58%
AP+AT+R	70.77 %
RGT+AT+R	**78.44**%

Table 1: Comparison of classification results on the Compcars test set.

Average pooling vs. Random grouping training

By comparing the results of AP and RGT, we can see that the group-wise training can effectively suppress the influence of noise due to the improved labeling accuracy at the instance level. For this reason, the model can always learn some useful information from the correctly labeled samples in each group. In contrast, for training at the image level with no attention, the noisy labels will give networks misleading information that will harm the learning process.

Attention vs. without attention

For AP+AT and RGT+AT, the accuracy all improves by a large margin compared to AP and RGT respectively, which proves the effectiveness of the attention model employed. The attention model filters out uninformative parts of the feature maps for each sample and only let the useful parts flow through the latter network for classification. In this way, it works like a gate that can prevent the noisy regions of the feature representation from misleading the classifiers.

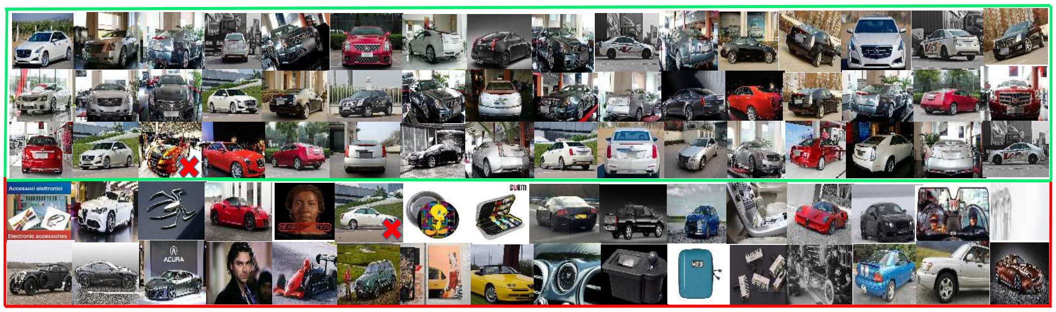

Figure 3: Examples of the image re-ranking performance on one sampled car category ("cadillac"). The red crosses indicate the images that are classified incorrectly. The images are sorted according to the rank of the classification scores in descending order. The images in the green rectangle and red rectangle are correctly labeled samples and mislabeled samples, respectively. The noise level is 0.4.

A similar strategy is found effective on clean images for multi-label image classification [51].

With vs. without regularizer

An interesting phenomenon we observe is that the accuracy for AP+AT drops significantly when using the noise regularizer, AP+AT+R. The reason is that the noise presents in both classes of object-of-interest and negative class, and consequently the image-level learning strategy confuses the network with how to classify the noise. But this confusion doesn't exist in the group-level training approach, since very few training instances have incorrect labels after random grouping. The reasons for adding noise regularizer is helpful for group-wise training are two-fold: First, the hinge loss regularizer forces the attention map not to concentrate on any feature regions of mislabeled samples, which results in a much cleaner group-level feature representation; Second, it helps the classifiers to distinguish the correctly labeled samples from the noise [15]. It's worth noticing that compared to utilizing clean images as constraint [45], the negative samples are much easier to collect.

We consider two types of label noise defined in [18], which are called *cross-domain* noise and *cross-category* noise. The cross-domain noise is defined to be the portion of images that are not of any category in the fine-grained domain, *i.e.* for cars, these images don't contain a car. In contrast, the cross-category noise is the mislabeled images within a fine-grained domain, *i.e.* a car example with the wrong model label.

We also provide qualitative examples in Figure 5. We see that the attention model mostly focuses on the discriminative parts in the front of or at the end of the cars. For some challenging examples, the correctly labeled car appears simultaneously with the cross-domain noise or cross-category noise in the same image. In this case, the attention model still successfully localizes to the correct parts. For the mislabeled samples, there's no object-of-interest to be concentrated on.

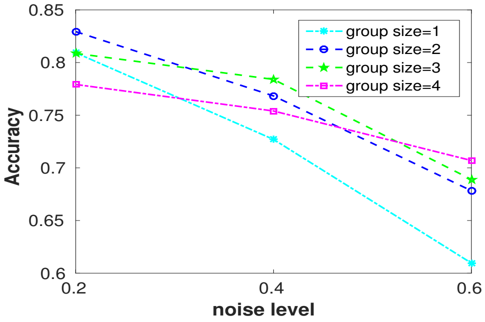

Figure 4: The classifcation accuracy under different group sizes of the proposed method.

4.4. Analysis of group size

In this section, we conduct a toy experiment to investigate the impact of the group size on our method (RGT+AT+R)[2]. We randomly sample 100 car categories of the Compcars dataset and deliberately pollute the clean training data by adding cross-category noise and cross-domain noise in a proportion of 1:1. The total number of training images doesn't change. We then gradually increase the noise level from 0.2 to 0.6 and report the classification accuracy on the test set of Compcars using different group sizes. The results are shown in Figure 4. From Figure 4, we could make the following observations: (1) using group size ≥ 2 makes the network training more robust to noise. As can be seen, when the dataset contains a substantial amount noise label e.g. noise level = 0.6, the performance gap between group size = 1 and group size ≥ 2 can be larger than 10%. (2) the optimal group size changes with the noise level. For example, when the noise level = 0.2, the optimal group size is 2 but when the noise level = 0.6, the optimal

[2] When group size equals 1, the method is equivalent to AP + AT + R. We empirically find adding the regularization term in this case will lead to inferior performance so we do not use the regularization term when group size equals 1.

group size becomes 4. This observation could be partially explained by the analysis in section 3.1, that is, the larger group size reduces the chance of having an incorrect label at the group-level. (3) Finally, we observe that larger k does not always lead to better performance. As also mentioned in section 3.1, we speculate that this is because having a larger group will reduce the independency of grouped instance. For example, when having a larger k, the chance of two groups sharing one common image will grow significantly.

4.5. Web Images re-ranking

To inspect whether the proposed method utilize the information from the correctly labeled data for training while ignoring the mislabeled ones, we now propose to re-rank the noisy training data used in Section 4.4 according to their classification scores. The ideal case is that the highly ranked images are all correctly labeled ones while the low-ranking samples are mislabeled ones on the Web. We compare three methods here, including AP, AP+AT as well as RGT+AT+R using different group sizes. The ground truth labels for correctly labeled images and mislabeled images are set to $+1$ and -1, respectively. Correctly labeled images are ranked high in the ground truth labels. Based on the learned models in Section 4.4, we first obtain the classification score for each training sample and rank the images in descending order based on their corresponding classification scores to get the predicted labels in each category. We then calculate the mean average precision (MAP) under different noise levels and group sizes. The mean average precision is obtained by averaging the precisions calculated at the total number of samples in different categories.

noise level methods	20 %	40 %	60 %
AP	93.72	85.08	74.42
AP+AT	96.71	92.84	90.56
RGT+AT+R, group size=2	**98.12**	95.81	91.00
RGT+AT+R, group size=3	97.71	**95.93**	91.04
RGT+AT+R, group size=4	97.95	95.33	**91.98**

Table 2: Comparison of mean average precisions % using several methods under different noise levels.

From the table, we can see that for direct average pooling, the precision drops dramatically as the noise level increases. On the contrary, simply adding attention model only, the precision improves considerably especially when the noise level is high enough. For example, at the noise level 60 %, the precision gap is more than 15 %. This result proves that selecting discriminative regions for each sample can effectively prevent noisy parts from impacting the final classification. By incorporating the group-wise training strategy, the performance further improves. This can be attributed to the highly accurate group-level labels used and the attention model for blocking the local features of mislabeled samples to generate the group-level representation. Overall, the proposed method is stable and performs well at different noise levels.

We also randomly select a car category and qualitatively evaluate the re-ranking performance at the noise level 0.4 (see Figure 3). The images are ranked in descending order based on their classification scores. We can see that only a pair of images are ranked incorrectly among the samples. From the results, we can expect that our method can further be used to assist collecting clean datasets or active learning.

4.6. Evaluation on CIFAR-10 with Synthetic Noises

We also conduct synthetic experiments on CIFAR-10 following the setting of [41, 45] and report the test accuracies under different noise levels in Table 3. As seen, the proposed method is more robust to label noise.

noise level methods	30 %	40 %	50 %
Caffe's CIFAR10-quick	65.57%	62.38%	57.36%
[41]	69.73%	66.66%	63.39%
[45]	69.81%	66.76%	63.00%
RGT+AT+R, group size=2	**74.88** %	70.33%	65.87%
RGT+AT+R, group size=3	71.76 %	**72.25**%	**67.15**%
RGT+AT+R, group size=4	70.23 %	70.74%	66.98%

Table 3: Accuracies on CIFAR-10 with synthetic label noises.

4.7. Evaluation on Web Images + ImageNet

Apart from the challenging fine-grained classification task, the proposed method can also be generalized to a conventional classification task. We trained models from scratch using the noisy Web data with respect to ImageNet described in Section 4.1 and test the performance on the ILSVRC2012 validation set.

methods	accuracy
AP	58.81%
AP+AT	67.68%
RGT+AT+R, group size=2	**71.24**%
RGT+AT+R, group size=3	68.89%
RGT+AT+R, group size=4	66.23%

Table 4: Comparison of classification results on ILSVRC2012 test set.

From the results we can see that for the conventional image classification task with Web data, the proposed method still works much better than the directly average pooling baseline. By only applying the attention model on each sample to select discriminative feature regions for classification, the result improves by $\sim 9\%$. By randomly generating groups online using reasonable group size and incorporating the regularizer, we get the best performance at

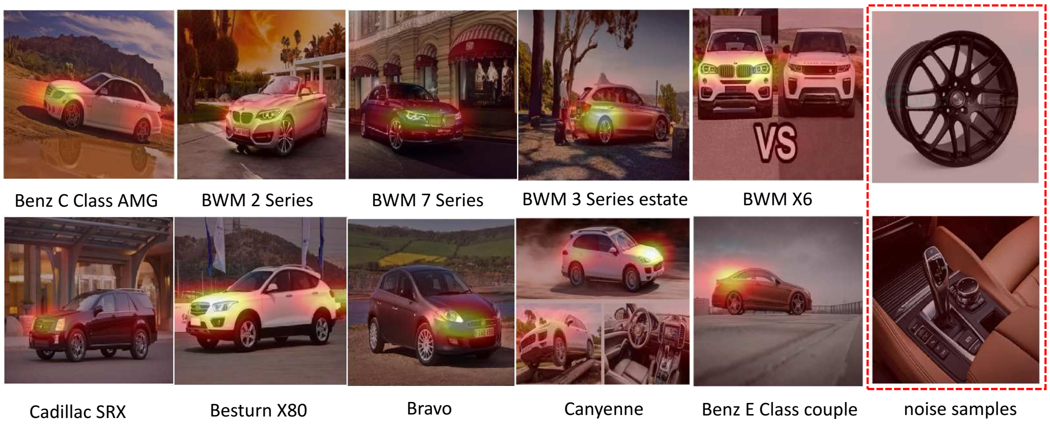

Figure 5: Examples of the attention maps using the large-scale noisy fine-grained dataset described in Section 4.1. The brighter the region, the higher the attention scores. The examples in the red dotted box are mislabeled samples on the Web.

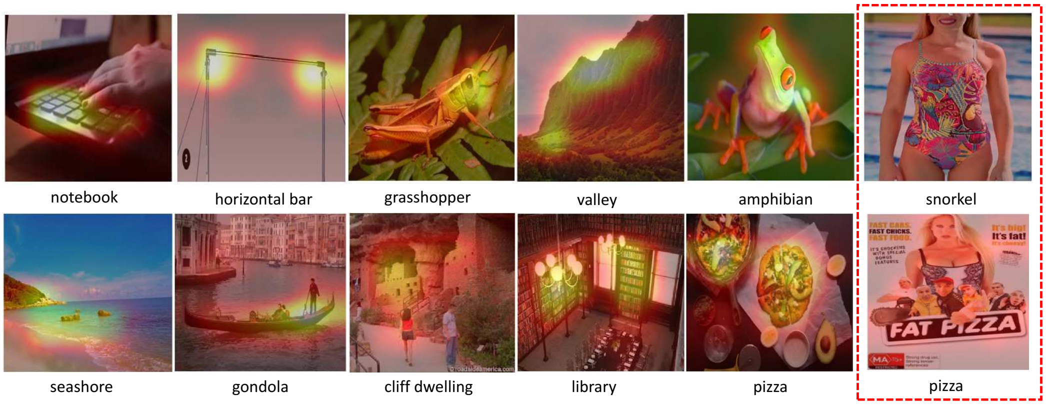

Figure 6: Examples of where attention maps for the collected Web data with respect to ImageNet described in Section 4.1. The brighter the region, the higher the attention scores. The examples in the red dotted box are mislabeled on the Web.

the optimal group size 2, which confirms the conclusions in Section 4.3 and Section 4.4.

We visualize some examples with their attention maps in Figure 6 using the best performed method RGT+AP+R with group size 2. The attention model attempts to localize the most discriminative parts for correctly labeled samples to push them far from the decision boundary. Samples in the red bounding box are mislabeled on the Web and the attention model finds no parts to concentrate on.

5. Conclusion

In this paper, we propose a weakly-supervised framework to learn visual representations from massive Web data with minor human supervision. The proposed method can handle label noise effectively by two unified strategies. By randomly stacking training images into groups, the accuracy of the group-level labels improves. The attention model embedded further localizes the discriminative regions corresponding to correctly labeled samples across the combined feature maps for classification. The efficacy of our methods have been demonstrated by the extensive experiments.

Acknowledgements This work was supported by the Australian Research Council through the ARC Centre for Robotic Vision (CE140100016), and through ARC DECRA Fellowship (DE170101259) to L. Liu, ARC Future Fellowship (FT120100969) to C. Shen and ARC Laureate Fellowship (FL130100102) to I. Reid.

References

[1] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Neural module networks. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, June 2016. 2

[2] J. Ba, V. Mnih, and K. Kavukcuoglu. Multiple object recognition with visual attention. In *Proc. Int. Conf. Learn. Repren.*, 2015. 2

[3] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *Proc. Int. Conf. Learn. Repren.*, 2015. 2

[4] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. Goodfellow, A. Bergeron, N. Bouchard, D. Warde-Farley, and Y. Bengio. Theano: new features and speed improvements. *arXiv preprint arXiv:1211.5590*, 2012. 5

[5] E. Beigman and B. B. Klebanov. Learning with annotation noise. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL*, pages 280–287. ACL, 2009. 2

[6] H. Bilen, M. Pedersoli, and T. Tuytelaars. Weakly supervised object detection with convex clustering. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1081–1089, 2015. 2

[7] C. E. Brodley and M. A. Friedl. Identifying mislabeled training data. *arXiv preprint arXiv:1106.0219*, 2011. 2

[8] X. Chen and A. Gupta. Webly supervised learning of convolutional networks. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1431–1439, 2015. 2

[9] X. Chen, A. Shrivastava, and A. Gupta. Neil: Extracting visual knowledge from web data. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 1409–1416, 2013. 2

[10] R. G. Cinbis, J. Verbeek, and C. Schmid. Multi-fold mil training for weakly supervised object localization. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 2409–2416. IEEE, 2014. 2

[11] T. Deselaers, B. Alexe, and V. Ferrari. Weakly supervised localization and learning with generic knowledge. *Int. J. Comp. Vis.*, 100(3):275–293, 2012. 2

[12] S. K. Divvala, A. Farhadi, and C. Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3270–3277, 2014. 2

[13] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comp. Vis.*, 88(2):303–338, 2010. 1

[14] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from internet image searches. *Proceedings of the IEEE*, 98(8):1453–1466, 2010. 2

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 580–587, 2014. 1, 6

[16] A. Graves, G. Wayne, and I. Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014. 2

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 770–778, 2016. 1

[18] J. Krause, B. Sapp, A. Howard, H. Zhou, A. Toshev, T. Duerig, J. Philbin, and L. Fei-Fei. The unreasonable effectiveness of noisy data for fine-grained recognition. In *Proc. Eur. Conf. Comp. Vis.*, pages 301–320. Springer, 2016. 2, 6

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 1097–1105, 2012. 1

[20] G. Lin, C. Shen, A. van den Hengel, and I. Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, June 2016. 1

[21] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Proc. Eur. Conf. Comp. Vis.*, pages 740–755. Springer, 2014. 1

[22] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *Proc. Eur. Conf. Comp. Vis.*, pages 21–37. Springer, 2016. 1

[23] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3431–3440, 2015. 1

[24] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 289–297, 2016. 2

[25] M.-T. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015. 2

[26] N. Manwani and P. Sastry. Noise tolerance under risk minimization. *IEEE transactions on cybernetics*, 43(3):1146–1151, 2013. 2

[27] A. L. Miranda, L. P. F. Garcia, A. C. Carvalho, and A. C. Lorena. Use of classification algorithms in noise detection and elimination. In *International Conference on Hybrid Artificial Intelligence Systems*, pages 417–424. Springer, 2009. 2

[28] V. Mnih, N. Heess, A. Graves, et al. Recurrent models of visual attention. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 2204–2212, 2014. 2

[29] V. Mnih and G. E. Hinton. Learning to label aerial images from noisy data. In *Proc. Int. Conf. Mach. Learn.*, pages 567–574, 2012. 2

[30] L. Niu, W. Li, and D. Xu. Visual recognition by learning from web data: A weakly supervised domain generalization approach. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 2774–2783, 2015. 2

[31] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 685–694, 2015. 2

[32] S. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014. 2

[33] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 91–99, 2015. 1

[34] W. Ren, K. Huang, D. Tao, and T. Tan. Weakly supervised large scale object localization with multiple instance learning and bag splitting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(2):405–416, 2016. 2

[35] M. Rochan and Y. Wang. Weakly supervised localization of novel objects using appearance transfer. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 4315–4324. IEEE, 2015. 2

[36] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *Int. J. Comp. Vis.*, pages 1–42, 2015. 1, 5

[37] F. Schroff, A. Criminisi, and A. Zisserman. Harvesting image databases from the web. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(4):754–766, 2011. 2

[38] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[39] H. O. Song, R. B. Girshick, S. Jegelka, J. Mairal, Z. Harchaoui, T. Darrell, et al. On learning to localize objects with minimal supervision. In *Proc. Int. Conf. Mach. Learn.*, pages 1611–1619, 2014. 2

[40] S. Sukhbaatar, J. Bruna, M. Paluri, L. Bourdev, and R. Fergus. Training convolutional networks with noisy labels. *arXiv preprint arXiv:1406.2080*, 2014. 2

[41] S. Sukhbaatar and R. Fergus. Learning from noisy labels with deep neural networks. In *Proc. Int. Conf. Learn. Repren. Workshops*, 2015. 2, 7

[42] C. Wang, W. Ren, K. Huang, and T. Tan. Weakly supervised object localization with latent category learning. In *Proc. Eur. Conf. Comp. Vis.*, pages 431–445. Springer, 2014. 2

[43] J. Weston, S. Chopra, and A. Bordes. Memory networks. *arXiv preprint arXiv:1410.3916*, 2014. 2

[44] J. Wu, Y. Yu, C. Huang, and K. Yu. Deep multiple instance learning for image classification and auto-annotation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3460–3469. IEEE, 2015. 2

[45] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang. Learning from massive noisy labeled data for image classification. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 2691–2699, 2015. 2, 6, 7

[46] H. Xu and K. Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *Proc. Eur. Conf. Comp. Vis.*, pages 451–466. Springer, 2016. 2

[47] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proc. Int. Conf. Mach. Learn.*, volume 14, pages 77–81, 2015. 2

[48] Z. Xu, S. Huang, Y. Zhang, and D. Tao. Augmenting strong supervision using web data for fine-grained categorization. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 2524–2532, 2015. 2

[49] L. Yang, P. Luo, C. Change Loy, and X. Tang. A large-scale car dataset for fine-grained categorization and verification. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3973–3981, 2015. 5

[50] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 21–29, 2016. 2

[51] R.-W. Zhao, J. Li, Y. Chen, J.-M. Liu, Y.-G. Jiang, and X. Xue. Regional gating neural networks for multi-label image classification. In *Proc. Brit. Mach. Vis. Conf.*, 2016. 6