Multi-attention Network for One Shot Learning

Peng Wang[1]*, Lingqiao Liu[1]*, Chunhua Shen[1], Zi Huang[2], Anton van den Hengel[1], Heng Tao Shen[3]
[1]The University of Adelaide, SA, Australia [2]The University of Queensland, QLD, Australia
[3]University of Electronic Science and Technology of China, Chengdu, China

Abstract

One-shot learning is a challenging problem where the aim is to recognize a class identified by a single training image. Given the practical importance of one-shot learning, it seems surprising that the rich information present in the class tag itself has largely been ignored. Most existing approaches restrict the use of the class tag to finding similar classes and transferring classifiers or metrics learned thereon. We demonstrate here, in contrast, that the class tag can inform one-shot learning as a guide to visual attention on the training image for creating the image representation. This is motivated by the fact that human beings can better interpret a training image if the class tag of the image is understood. Specifically, we design a neural network architecture which takes the semantic embedding of the class tag to generate attention maps and uses those attention maps to create the image features for one-shot learning. Note that unlike other applications, our task requires that the learned attention generator can be generalized to novel classes. We show that this can be realized by representing class tags with distributed word embeddings and learning the attention map generator from an auxiliary training set. Also, we design a multiple-attention scheme to extract richer information from the exemplar image and this leads to substantial performance improvement. Through comprehensive experiments, we show that the proposed approach leads to superior performance over the baseline methods.

1. Introduction

The volume of data required to train current machine learning technologies is one of the major limitations on the range of problems they can usefully be applied to. Human beings, in contrast, are often able to learn to identify a class from a single training instance. One-shot learning is the machine learning problem which aims to mimic this human ability. One of the primary difficulties in one-shot learning is to generalize beyond the specific, single, training instance, which inevitably requires identifying which parts of the training image are important. The class tag is a useful source of information which can help to identify the essential features of the class. However, most, if not all, existing approaches use the class tag in a very restricted way, that is, they turn to the class tag only when seeking other classes from which to source exploitable classifiers [7, 19].

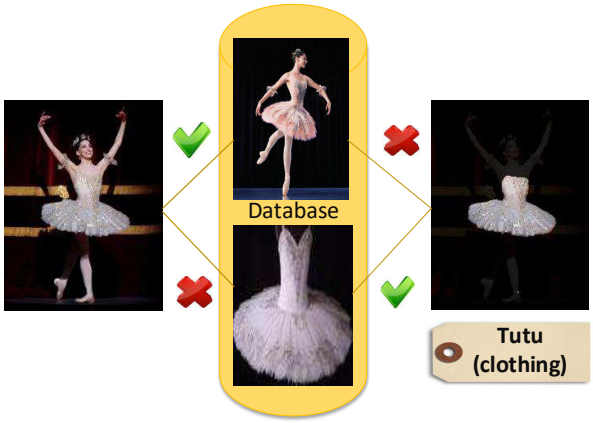

Figure 1. Given an exemplar image of a novel class, the objective of one-shot learning is to identify the images belonging to the same class from a database. The left image is an exemplar image of class **tutu**. Using the image solely can result in ambiguity in recognition. If the class tag is understood that **tutu** is a kind of clothing, it can help to focus the attention on the **tutu** parts and consequently make more accurate decision.

The approach we propose here uses the class tag to guide an attention mechanism able to identify which parts of the training image are most relevant. Our method is motivated by the observation that human beings can better interpret an exemplar image if its class tag is well understood. For example, as illustrated in Fig. 1, from a single exemplar it is difficult to understand which part of the image is relevant to the class, which leads to ambiguity in recognition. But if we understand that the class tag "tutu" implies a kind of clothing, we can infer that the region around the human body is most relevant. Mimicking this process, in this paper we

*The first two authors contributed to this work equally. L. Liu's participation was in part supported by ARC DECRA Fellowship (DE170101259).

propose to use an attention mechanism to establish a correspondence between the class tag and the visual content to enable a better image representation for one-shot learning.

In contrast to the application of attention modules in other tasks, such as visual question answering [27], our method requires that the attention map generator can be transferred to novel classes. In our work, we leverage distributed word embeddings [16] to represent the class tag to capture the semantic relationships between different concepts. Thus the attention map generator becomes a mapping function from the word embedding and image features to attention maps. We learn such a mapping from an auxiliary dataset and we show that this mapping function does in fact generalize to the exemplars of novel classes.

Another key novelty of our proposed network is that we propose the idea of generating and using multiple attention maps. There exists various clues that can help to recognize a class e.g., different object parts and contextual scenes, the visual appearance of which can vary significantly. A single attention map can be insufficient to explore this information and consequently may risk losing important cues. Multiple attention maps can alleviate this problem by offering additional opportunities to extract useful information, thus helping to create a more robust image representation.

To evaluate the proposed attention scheme for one-shot learning we construct two datasets, one focusing on different animal classes as in [12], the other containing a larger number of generic object classes. These datasets are not limited to this work, and can be used as benchmark for one-shot learning. Comprehensive experiments conducted on these two datasets demonstrate the advantage of our method over baseline methods. In summary, the contributions of this work are as follows:

- We show that class tag information can contribute one-shot learning, and devise a novel method which is capable of exploiting this information.
- We propose an attention network that can generate attention maps for creating the image representation of an exemplar image in novel class based on its class tag.
- We further propose a multi-attention scheme to boost the performance of the proposed attention network.
- We collect two new datasets and establish an experimental protocol for evaluating one-shot learning.

2. Related Work

One Shot Learning. A variety of methods have approached the one-shot learning problem by transferring the classifiers or metrics learned in previous categories resorting to the class tag [7, 19, 15]. In [7], the authors represent the object categories by a probabilistic model. They model the knowledge learned in other classes as a prior probability function w.r.t. the model parameters, and given an exemplar of a novel class, they update the knowledge and generate a posterior density to recognize novel instances. They learn a single prior for all categories, however, and only three categories of models are employed to form the prior. This impedes the generalization ability of the method. The method in [19] takes a step further. They group categories into super-categories and learn a prior for each of these super-groups. Given a single training image of a new class, they first assign it to a super-category and then estimate the parameters of the class resorting to the corresponding super-category. Another type of works addressing the one-shot learning problem using the essential idea of metric learning, which tries to map the image features into a space where images of the same class are close to each other while instances belonging to different classes are separated. The authors in [10] present a typical method of this kind. They train a Siamese network to identify the positive/negative training pairs and apply the learned feature maps to novel classes to verity whether two instances belong to the same class. Recently, to overcome deep neural networks' need for a large amount of data to train a class, some efforts [20, 23] have exploited the memory-augmented model to quickly encode and retrieve sufficient information for the new task. Note that many of the above existing methods are orthogonal to our approach in the sense that they can be applied on top of the image representation generated by our networks.

Attention Models. Attention models have been applied to a variety of computer vision problems including image classification [25, 1, 8], semantic segmentation [3], visual tracking [5], person identification [9], image captioning [4, 6, 26] and question answering [27]. The focus of the attention in each case varies with the application. The focus can be image parts [25, 1, 27], different scales [3] or spatio-temporal regions [9]. Despite the different application scenarios, the essential schemes of some attention models are similar. They use the training data to learn a network that can adaptively locate the relevant information. In some sense this is akin to implicitly learning a classifier or detector. Considerable training data is thus required to guarantee the generalization of the network. In this paper, we generate the attention map in a completely different way, which uses the embeddings of class tags to emphasize the class-relevant content. By exploiting the underlining semantic relationship between the semantic representations [16] of class tags, our attention model can work on novel classes even from only a single training instance. Our method is also related to saliency detection [11]. The objectives, however, are different. Saliency detection aims at identifying salient objects within an image and segmenting the object boundaries. Our goal is rather to place the attention on visual content that is relevant to the class tag without the requirement of accurate localization.

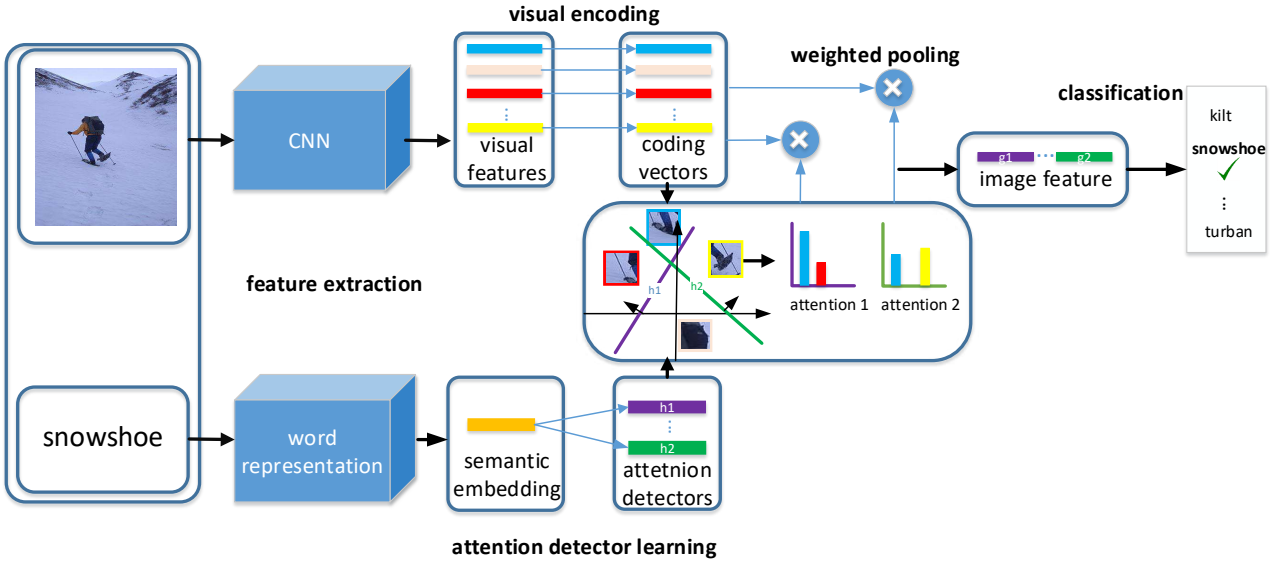

Figure 2. An illustration of the overview architecture of the proposed attention network.

3. Proposed Approach

In this section, we will elaborate on the proposed attention network for one-shot learning. We firstly give a formal definition of the problem we study, and then illustrate the overall architecture of the proposed network which is followed by the detailed depiction of the key modules.

3.1. Problem Definition and Notations

Given an exemplar image I_e belonging to class c, our task is to predict whether an image in a test set belongs to class c. Without loss of generality, we represent each image as a set of local features $\mathcal{X} = \{\mathbf{x}_i\}$, $\mathbf{x}_i \in \mathbb{R}^{d_v}$, where d_v is the dimensionality of the local features. For each class c, we define a vector $\mathbf{c} \in \mathbb{R}^{d_w}$, representing the semantic embedding of the class' tag where d_w is the dimensionality of the tag embedding. A mapping function is learned from image and class tag pairs, to attention maps. The mapping is trained on an auxiliary dataset consisting of a category set $\mathcal{C}^N$ which does not overlap with $\{c\}$ ($\{c\} \cap \mathcal{C}^N = \emptyset$).

3.2. The Attention Network

3.2.1 Overview

The architecture of the proposed attention network is illustrated in Fig. 2. The input of the network is an image and its associated class tag. The image is fed into a CNN to extract the local visual features and the class tag is represented by its distributed semantic embedding e.g., word2vector [14] or GloVe [16]. We propose to use this embedding to guide the visual attention in the image. Firstly we apply an encoder to map the local visual features into a set of coding vectors. Then we generate an attention detector (or attention detectors) from the semantic embedding of the class tag and apply this detector on the coding vectors to generate the attention map (or attention maps). The attention map is then used to perform weighted pooling on the coding vectors to obtain the image-level representation. In the following sections, we elaborate the modules in our network.

3.2.2 Feature Extraction

Our method applies to any scenario where the input image can be represented by a set of local features. In this work, we extract the convolutional feature maps of a CNN and view them as an array of local features. Other representations, such as extracting features from multiple object proposals [22], could also be used. For the class tag, we use the GloVe [16] pre-trained from a large-scale corpus as the word embedding. If the class tag contains a phase with more than one word, we average the embedding of each word in the phase as the semantic representation of the class tag.

3.2.3 Visual Feature Encoding and Weighted Pooling

We apply a local feature encoder to each of the local feature. Formally, the encoder is a mapping function defined as follows:

$$\mathbf{v}_i = f(\mathbf{W}_v \mathbf{x}_i + \mathbf{b}_v), \tag{1}$$

where $f(a) = \max(0, a)$ is a rectified linear unit (ReLU), $\mathbf{x}_i$ is a local feature, $\mathbf{W}_v \in \mathbb{R}^{d \times d_v}$ and $\mathbf{b}_v \in \mathbb{R}^d$ are the model parameters.

Instead of directly aggregating these local features to create the image representation, we pool these features via the guidance of a set of attention maps. The basic form of this pooling operation is as follows:

$$\mathbf{g} = \sum_{i=1}^{|\mathcal{X}|} \mathbf{v}_i a_i, \tag{2}$$

where a_i indicates the attention value on the i-th coding vector. In the following section, we introduce the details of the calculation of a_i.

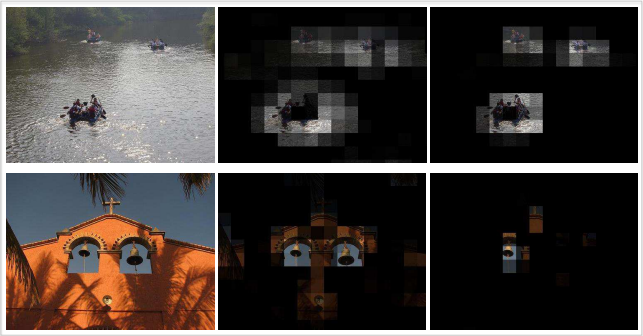

Figure 3. Examples of generated attention maps, where the intensity indicates the activeness of different regions. Two attention maps are generated for each exemplar image and these two categories, "raft" and "church bell", do not appear in the network training phase.

3.2.4 Attention Map Generation

To generate the attention map, we first create an attention detector from the semantic embedding of the class tag, that is, we let:

$$\mathbf{h} = \mathbf{W}_s \mathbf{c} + \mathbf{b}_s, \tag{3}$$

where $\mathbf{c}$ is the semantic embedding of the class tag, $\mathbf{W}_s \in \mathbb{R}^{d \times d_w}$ and $\mathbf{b}_s \in \mathbb{R}^d$ are the model parameters to be learned.

The generated attention detector is then applied to each coding vector to obtain its initial attention confidence score a'_i:

$$a'_i = \mathbf{h}^\top \mathbf{v}_i. \tag{4}$$

Higher scores are expected on local regions that are relevant to the class tag. This is driven by the objective function i.e., emphasizing on irrelevant content results in an image feature not discriminative which will be penalized by the loss function. This confidence score is then normalized to obtain the final attention value a_i:

$$a_i = \frac{b(a'_i)}{\sum_i b(a'_i)}, \tag{5}$$

where $b(\cdot)$ is a function guaranteeing that the attention scores are positive. We design two normalization strategies here. For the first, we use a ReLU function:

$$b(a'_i) = \max(a'_i, 0), \tag{6}$$

and for the second, we use a Score Shifting scheme:

$$b(a'_i) = a'_i - \min_j(\{a'_j\}). \tag{7}$$

The difference between these two strategies is that the former strategy completely ignores the negative-scored parts which are identified as irrelevant by the attention detector while the latter will consider both positive and negative detection scores. In section 4.3.4 we conduct experiments to compare these two normalization strategies.

Fig. 3 shows the examples of the attention maps generated from our attention module. As can be seen, the attention maps emphasize more on the parts that are relevant to the class tag. Thus by applying weighted pooling with the attention maps, the distraction from the irrelevant content of an image class can be largely avoided. Note that the classes in these examples are not seen at the training stage.

3.2.5 Multi-Attention Scheme

As one of the key novelties in our approach, we propose to use a multiple-attention scheme to generate multiple attention maps. The advantage of using multiple attention maps over a single attention map are twofold: (1) it can depict various aspects of an exemplar image, e.g., different attention map highlights different parts of an object or highlights the object and its visual context; (2) it reduces the risk of having an incorrect attention map since more attention maps means more chances of having at least one attention map correctly focus on the relevant content. Fig. 4 shows such an example. As can be seen, the single attention map fails to focus on the "signboards" but they are captured when two attention maps are used.

We realize the multiple-attention scheme by creating multiple attention detectors $\{\mathbf{h}_k\}$. This is achieved by using t sets of $(\mathbf{W_s}^k, \mathbf{b_s}^k)$ as in Eq. 3. By applying the same aforementioned normalization and weighted pooling scheme, we finally create t pooling vector $\{\mathbf{g}_k\}$, which we concatenate together to obtain the final image-level representation:

$$\mathbf{G} = [\mathbf{g}_1, \mathbf{g}_2, \cdots, \mathbf{g}_t]. \tag{8}$$

3.3. Network Training

The purpose of the training stage is to learn an image feature generator with the function form $F(\mathbf{I}, c)$, where $\mathbf{I}$ is the input image and c is its associated class tag. At the testing stage, we generate the image feature for an exemplar image $\mathbf{I}_e$ by using $F(\mathbf{I}_e, c_e)$, where c_e is its class tag.

Then the distance between a test image $\mathbf{I}_t$ and $\mathbf{I}_e$ is calculated via $dist(\mathbf{I}_t, \mathbf{I}_e) = dist(F(\mathbf{I}_e, c_e), F(\mathbf{I}_t, c_e))$. With this distance, we can perform either image retrieval or image classification task.

We train the image feature generator on an auxiliary dataset which does not contain training samples from the testing classes. To train the function $F(\cdot, \cdot)$, we apply a linear classifier on top of its generated features and use cross-entropy loss to jointly train $F(\cdot, \cdot)$ and the classifier in an end-to-end fashion. In practice, we use Stochastic Gradient Descent (SGD) to train the network and apply weight decay to regularize the model parameters.

At the testing stage, the classifier is discarded and only $F(\cdot, \cdot)$ is used. This is because the class tag used in the testing stage does not appear in the training stage.

Figure 4. Single v.s. two attention maps on class "signboard".

4. Experiments

In this section, we will first introduce some implementation details, which is followed by the illustration of the datasets constructed to evaluate one-shot learning task. Then we will present some quantitative comparison between our method and some baseline methods. Finally, the visualization of the attention maps will be given to qualitatively analyse the effectiveness of the proposed attention scheme.

4.1. Implementation Details

We use the activations of the last convolutional layer of VGG network [21] as visual features for images. Note that we do not rescale the images into fixed size before feeding them into the network but preserve their original sizes (or aspect ratios). For class tag representation, we use GloVe [16] to extract the 300-dimensional semantic embedding. In the visual feature encoding stage, we encode the features into 256 dimensionality.

4.2. Datasets

To evaluate the proposed attention network, we construct two datasets from ImageNet [18]. For each dataset, the images are divided into two subsets and the classes of these two subsets do not overlap. One subset is used as auxiliary dataset for network training and the other to evaluate the one-shot learning task. For simplicity, we name them the auxiliary set and the evaluation set. In this paper, we have two types of experimental settings. While one setting uses only the evaluation set as the database from which to identify the images of a target class, the other setting is more challenging in that it uses both the evaluation set and auxiliary set as the database. Through out this paper, we name the former setting "close-word" setting and the latter setting "open-world" setting.

4.2.1 Animal Dataset

We construct the Animal Dataset based on a benchmark dataset for zero-shot learning, Animals with Attributes [12]. That dataset consists of 50 animal classes and provides a real-valued attribute vector for each class. Since the dataset does not provide the raw images, we collect the images under each animal class from ImageNet. We use the same split protocol as in [12], that is, we use 40 categories for network training and the other 10 classes for one-shot learning evaluation. We employ the attribute vectors provided in [12] as one of the semantic representations of the class tags.

4.2.2 Artifact Dataset

The Animal Dataset tries to transfer the knowledge learned from 40 categories to 10 novel categories, where all the categories are animals. To verify the tansferability in a more general range of classes, we collect another dataset called the Artifact Dataset. In this dataset, we use the 1000 classes in the classification task of ImageNet as network training data and sample another 100 classes that do not appear in the aforementioned 1000 classes from the synset "Artifact" of ImageNet for one-shot learning task. Considering the training and evaluation efficiency, instead of putting all the images under a class into our dataset, we randomly sample 50 images per class.

4.3. Experimental Results

In this section, we demonstrate the experimental results. Given an exemplar image of a class, the task is to identify the images having the same class label from a database, which is similar to image retrieval. We use the images in a class as the exemplar image in turn and the Mean Average Precision (mAP) in image retrieval is employed as the evaluation metric.[1]

We will first show the performance comparison between our method and some baseline methods. Then we delve into the network structure to study the affect of the two aforementioned attention score normalization strategies on the one-shot learning performance. Finally, the visualization of

[1]Not that, as illustrated in section 3.3, we can evaluate classification performance using nearest-neighbour classifier as well.

some example attention maps will be shown to qualitatively evaluate the proposed attention scheme.

4.3.1 The Comparison Methods

Note that the focus of the experiments is to show the advantage of our attention based feature generation scheme for one-shot learning task. Most existing approaches for one-shot learning are orthogonal to our method in the sense that they can be applied on top of the image feature generated by our method. Here we compare the following methods.

- Global FC representation: we feed the whole image into VGG network [21] and extract the activations of the last fully-connected layer as the global image representation. We perform PCA to decorrelate the dimensions and conduct ℓ_2 normalization to normalize the features. Cosine similarity is used to measure the distance between images.

- Supervised encoding (SE): to verify the effectiveness of the attention scheme, we choose another end-to-end learning baseline, namely supervised encoding [24, 13]. It first encodes each local visual feature into a coding vector as in our method and then directly aggregates these coding vectors to form an image representation via sum pooling. The pooled feature is fed into a classification layer for classification. The number in the parenthesis (if apply) indicates the dimensionality of the coding vectors. The supervised encoding method as well as the joint Bayesian [2] introduced later are in some sense similar to the metric learning fashion methods [10], the essence of which is to map the images of the same class together and separate the instances from different classes apart. Both supervised encoding method and our method use the same auxiliary dataset to train the network.

- Supervised encoding + Joint Bayesian: the joint Bayesian [2] is a very effective method for face verification and can be applied to other scenarios as a general metric learning method. We use the image feature generated from supervised ecoding to learn the parameters and use a probabilistic measure of similarity between two images, introduced in [2], to verify whether they belong to the same class.

- Zero shot learning: to demonstrate the importance of having an image exemplar to learn a class, we implement a state-of-the-art zero-shot learning approach [17]. zero-shot learning tries to recognize a novel class by just having a description of it. Here we use the attributes provided in [12] as the class description. The method consists of training and inference. At the training step, we use the descriptions and instances of the training classes in the auxiliary dataset to learn a projection matrix V which maps from the visual feature space to the attribute space. At inference step, we use that matrix V to map the attribute of a novel class c_e into a linear model. This linear model can be used as a classifier applicable to the images in the evaluation dataset, where the images corresponding to class c_e are intended to have higher classification scores. We tune the hyper-parameters via cross validation.[2]

- Attention: this denotes our network with a single attention map. The content in the parenthesis (if apply) indicates how to represent the class tag, attributes [12] or word embedding [16]. The generated image feature is ℓ_2 normalized and cosine distance is employed as the measure of similarity.

- Attention + Joint Bayesian: for fair comparison we also apply the joint Bayesian [2] to the image representation generated by our attention model. And the same probabilistic measure in [2] is employed to identify whether two images belong to the same class.

- Multi-Attention: we generate and use multiple attention maps and the number in the parenthesis denotes the number of attention maps adopted. For multi-attention, we use word embedding [16] as the class tag representation.

4.3.2 Results on Animal Dataset

Table 1 shows the experimental results on the Animal Dataset. Note that the performance of FC is partially due to the reason that it is trained to classify 1000 classes which cover a subset of the testing animal classes. Supervised encoding is most comparable to our method. To show the detailed comparison, we also give the comparison by class in Table 3. It treats the local visual features equally and can, to a large extent, suffer from the distraction influence of the irrelevant content. Benefiting from the guidance of the attention maps, our method can generate a more discriminative image representation which focuses mainly on the relevant content. A significant performance jump is observed when multiple attention maps are adopted. As seen, two and five attention maps both boost the recognition performance obviously. Note that in the close-world scenario, two attention maps achieves better performance.

Zero-shot learning method [17] has the worst performance especially when distinguishing a novel class from a wider range of classes in the open-world scenario. This observation reveals the limitation of zero-shot learning and

[2]Using the tuned parameters we achieve 79.9% accuracy on Animal with Attributes Dataset [12] which represents the state-of-the-art performance.

Table 1. Comparison of the attention network to alternative solutions on the Animal Dataset.

close-world	Global FC	68.9%
	SE	67.7%
	Zero shot learning	55.2%
	Attention (attribute)	**72.4**%
	Attention (word vector)	**74.0**%
	Multi-Attention (2)	**82.4**%
	Multi-Attention (5)	**77.7**%
open-world	Global FC	42.8%
	SE	42.1%
	Zero shot learning	15.3%
	Attention (attribute)	**47.4**%
	Attention (word vector)	**49.0**%
	Multi-Attention (2)	**55.7**%
	Multi-Attention (5)	**56.8**%

highlights the importance of using visual clues when recognizing a new class, even if there is only one example. Another interesting result here is that when using word embedding [16] representing the class tag, our method achieves better performance. This is beneficial for the generalization of our method because defining the attributes for general classes is difficult and labour intensive while the word embedding can cover massive number of concepts, including those people are not familiar with.

Table 2. Comparison of the attention network to alternative solutions on the Artifact Dataset.

close-world	SE (256D)	27.8%
	SE (512D)	28.2%
	SE + Joint Bayesian	31.5%
	Attention	**34.5**%
	Attention + Joint Bayesian	**36.8**%
	Multi-Attention (2)	**50.5**%
open-world	SE (256D)	11.6%
	SE (512D)	12.2%
	SE + Joint Bayesian	11.4%
	Attention	**15.2**%
	Multi-Attention (2)	**28.2**%

4.3.3 Results on Artifact Dataset

Table 2 demonstrates the results on the Artifact Dataset. Again, we observe significant advantage of our method over the comparing methods. For supervised encoding, we encode the local visual features into two different dimensionalities 256 and 512 to see the affect of differnt dimensionalities of coding vectors. We can see doubling the dimensionality only leads to minor performance improvement. This means simply increasing the dimensionality of the coding vector cannot help to capture more useful information. When applying the joint Bayesian [2] to the image representation generated by supervised encoding, the mAP rises about 4% in close-world setting but remains almost the same in open-world setting. When using a single attention map to guide the local feature aggregation, it harvests about 7% and 4% higher recognition performance comparing to supervised encoding in close-world and open-world settings respectively. And when applying the joint Bayesian [2] to our representation, we see further improvement. Again the most significant performance jump happens when we apply multiple attention maps. From Table 2 we can see, when we use two attention maps, the performance is improved by 16% and 13% respectively in close-world and open-world settings comparing to using single attention map. The advantage of having multiple attention maps is that it can preserve more relevant information. An example is shown in Fig. 4, where the "signboards" are ignored by the single attention map but picked out when using two attention maps.

Table 4. Comparison of two different attention score normalization schemes on the Artifact Dataset.

close-world	Score Shifting	32.0%
	ReLU	34.5%
open-world	Score Shifting	15.2%
	ReLU	15.2%

4.3.4 Two Schemes on Attention Score Normalization

Attention score normalization is an important step towards network training and one-shot learning performance. An important role of it is to preserve the relative importance of different image parts. With this strategy, we alleviate the requirement that the attention values of the useful parts should remain stably high. Instead, we only need these parts to obtain relatively higher attention values compared to the distraction factors. In this part, we compare the performance of the two normalization schemes introduced in Eq. 6 and Eq. 7. The former is a ReLU function that ignores the parts identified as irrelevant and the latter raises all the attention values with the minimum score obtained

Table 3. Comparison between supervised encoding and the proposed attention network on the Animal Dataset. Average precisions (%) for each class are reported. The upper part shows the close-world results and the bottom part shows the open-world results.

Methods	humpback whale	leopard	chimpanzee	rat	persian cat	hippopotamus	giant panda	pig	raccoon	seal
SE	75.7	98.1	82.9	69.1	83.6	49.4	90.6	28.6	62.0	43.4
Attention	**79.8**	**98.5**	79.3	70.5	89.9	60.1	**93.9**	37.1	77.5	55.6
Multi-Attention (2)	73.1	96.4	**86.8**	**82.0**	**90.6**	**77.3**	**93.5**	**67.5**	**80.4**	**69.1**

Methods	humpback whale	leopard	chimpanzee	rat	persian cat	hippopotamus	giant panda	pig	raccoon	seal
SE	12.3	82.9	39.0	27.5	61.4	24.0	80.2	8.8	39.7	17.3
Attention	**23.6**	**86.4**	42.6	36.2	**65.4**	35.5	83.3	10.9	52.8	28.2
Multi-Attention (2)	21.2	80.2	**47.3**	**47.4**	**65.2**	**40.6**	**84.0**	**26.8**	**54.6**	**39.0**

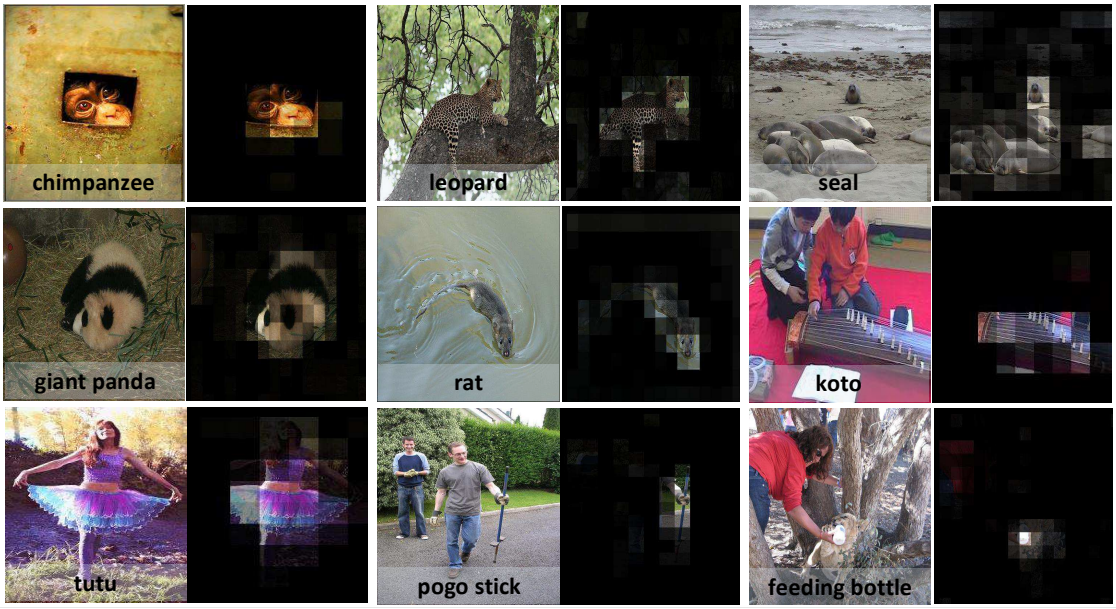

Figure 5. Visualization of attention maps on novel classes. The first five examples are from the Animal Dataset and the remaining examples are from the Artifact Dataset. Note that these classes do not appear in the network training stage.

from an image (Score Shifting). Both schemes can guarantee the attention maps are composed of positive values. Table 4 gives this comparison on the Artifact Dataset. As can be seen, they achieve comparable performance in the open-world setting and ReLU outperforms Score Shifting by 2% in the close-world setting. The results manifest that both schemes can highlight the discriminative information and ReLU may lead to superior performance because it can further remove the distraction of some noisy content.

4.3.5 Attention Map Visualization

To qualitatively evaluate the proposed attention scheme, we visualize some example attention maps generated on novel classes in Fig. 5. Although these classes are not seen in the network training stage, the generated attention maps successfully highlight the content depicted by the class tags. Interestingly, our method can work well in some challenging cases where the object concerned is small w.r.t the size of the image, such as the "pogo stick" and "feeding bottle" examples in Fig. 5. If we create the image feature by directly aggregating all the local features, these small objects tend to be overwhelmed by the distraction content and this can lead to the failure in image recognition. However, with the guidance of the attention maps, our method can create a more discriminative image representation.

5. Conclusion

We propose a novel method to exploit the class tag to benefit one-shot learning. Specially, we design a neural network that can generate attention maps for creating the image representation of exemplar image in novel class based on its class tag. To further boost the performance, a multi-attention scheme is proposed. The framework can be applied to more general settings, e.g., few-shot learning, by generating discriminative image representations resorting to the class tag. This will be investigated in future work.

References

[1] C. Cao, X. Liu, Y. Yang, Y. Yu, J. Wang, Z. Wang, Y. Huang, L. Wang, C. Huang, W. Xu, D. Ramanan, and T. S. Huang. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In *ICCV*, 2015.

[2] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun. Bayesian face revisited: A joint formulation. In *ECCV*, 2012.

[3] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation. In *CVPR*, 2016.

[4] X. Chen and C. Lawrence Zitnick. Mind's eye: A recurrent visual representation for image caption generation. In *CVPR*, 2015.

[5] J. Choi, H. Jin Chang, J. Jeong, Y. Demiris, and J. Young Choi. Visual tracking using attention-modulated disintegration and integration. In *CVPR*, 2016.

[6] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollar, J. Gao, X. He, M. Mitchell, J. C. Platt, C. Lawrence Zitnick, and G. Zweig. From captions to visual concepts and back. In *CVPR*, 2015.

[7] L. Fei-fei, R. Fergus, and P. Perona. One-shot learning of object categories. *TPAMI*, 28, 2006.

[8] K. Gregor, I. Danihelka, A. Graves, D. Rezende, and D. Wierstra. Draw: A recurrent neural network for image generation. In *ICML*, 2015.

[9] A. Haque, A. Alahi, and L. Fei-Fei. Recurrent attention models for depth-based person identification. In *CVPR*, 2016.

[10] G. Koch, R. Zemel, and R. Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML*, 2015.

[11] J. Kuen, Z. Wang, and G. Wang. Recurrent attentional networks for saliency detection. In *CVPR*, 2016.

[12] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by betweenclass attribute transfer. In *CVPR*, 2009.

[13] L. Liu, P. Wang, C. Shen, L. Wang, A. van den Hengel, C. Wang, and H. T. Shen. Compositional model based fisher vector coding for image classification. *TPAMI*, 2017.

[14] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*. 2013.

[15] S. Naha and Y. Wang. Zero-shot object recognition using semantic label vectors. *Conference on Computer and Robot Vision*, 2015.

[16] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

[17] B. Romera-Paredes and P. H. Torr. An embarrassingly simple approach to zero-shot learning. *ICML*, 2015.

[18] 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. *IJCV*, 2015.

[19] R. Salakhutdinov, J. B. Tenenbaum, and A. Torralba. One-shot learning with a hierarchical nonparametric bayesian model. *JMLR Workshop*, 27, 2012.

[20] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap. One-shot learning with memory-augmented neural networks. *arXiv:1605.06065*, 2016.

[21] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[22] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders. Selective search for object recognition. *IJCV*, 2013.

[23] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra. Matching networks for one shot learning. *arXiv:1606.04080*, 2016.

[24] P. Wang, Y. Cao, C. Shen, L. Liu, and H. T. Shen. Temporal pyramid pooling based convolutional neural network for action recognition. *TCSVT*, 2016.

[25] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *CVPR*, 2015.

[26] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

[27] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In *CVPR*, 2016.