Multi-attention Network for One Shot Learning

Zhang, Liqiang

May 9,2018

1 Multi-attention Network

The author propose the approach which use the class tag to guide an attention mechanism able to identify which parts of the training image are most relevant. This method is motivated by the observation that human beings can better interpret an exemplar image if its class tag is well understood. [1]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.

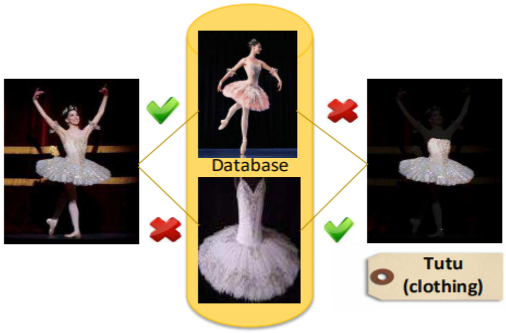

Figure 1: Given an exemplar image of a novel class, the objective of one-shot learning is to identify the images belonging to the same class from a database.

The contributions of this work are as follows:

- Showing that class tag information can contribute oneshot learning, and devise a novel method which is capable of exploiting this information.
- Proposing 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.
- Furthering propose a multi-attention scheme to boost the performance of the proposed attention network.
- Collecting two new datasets and establish an experimental protocol for evaluating one-shot learning.

2 Visual Feature Encoding and Weighted Pooling

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

$$V_i = f(W_v x_i + b_v) \tag{1}$$

where $f(a) = max(0, a)$ is a rectified linear unit (ReLU), x_i is a local feature, $W_v \in R^{d \times d_v}$ and $b_v \in R^d$ 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:

$$g = \sum_{i=1}^{|X|} \tag{2}$$

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

3 Results on Artifact Dataset

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

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

Table. 1 demonstrates the results on the Artifact Dataset. [2]This means simply increasing the dimensionality of the coding vector cannot help to capture more useful information.

References

[1] Chunshui Cao, Wei Xu, Deva Ramanan, Thomas S. Huang, Xianming Liu, Yi Yang, Yinan Yu, Jiang Wang, Zilei Wang, and Y-ongzhen Huang. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In *IEEE International Conference on Computer Vision*, pages 2956–2964, 2015.

[2] Dong Chen, Xudong Cao, Liwei Wang, Fang Wen, and Jian Sun. Bayesian face revisited: A joint formulation. In *European Conference on Computer Vision*, pages 566–579, 2012.