Dynamic Feature Combination by Agreement for Image Classification

Dynamic Feature Combination by Agreement for Image Classification

Chenghua Li[1], Wanguo Wang[2], Linzhi Liu[1,3] and Tian Liang[1,3]

[1] NLPR & AIRIA, Institute of Automation, Chinese Academy of Sciences, No. 95 Zhongguancun East Rd., Beijing 100190, China
[2] State Grid Intelligence Technology Co., Ltd., China
[3] University of Chinese Academy of Sciences, Beijing 100190, China
Email: lichenghua2014@ia.ac.cn

Abstract. Image classification is a basic and important task in computer vision. Recently, various neural networks have been designed and proved to be very powerful models for image classification. It is natural for thinking of how to gather their strengths together, which refers to feature combination tasks. Traditional combination methods mainly focus on designing specific combination algorithms to achieve higher performance. However, few works consider how to utilize their agreement on a given target (for example, a specific class) to achieve better combinations. This paper presents a novel dynamic feature combination method (DFCA) for image classification problems based on the agreement of the individual features. DFCA promisingly takes the agreement of not only the commonalities, but also the individualities of different features by dynamically updating the weighting coefficients of given features using a routing module. Experiment and extensive analysis on CIFAR-10 prove the effectiveness and promising characteristics of the proposed method.

1. Introduction

The image is categorized according to its visual content. Recently, CNNs have been widely used in visual tasks and various structures have been designed and verified in image classification, for example, AlexNet [1], VGG [2], GoogleNet [3], ResNet [4], DenseNet [5], SENet [6], etc. These CNNs can be viewed as strong feature descriptors compared to the traditional hand-crafted ones.

After generating multiple individual models, it is natural to find a best combination policy of them because different features may reinforce each other. Traditional combination methods like averaging or voting simply try to combine all the individual learners to make predictions. However, these "strong combination" strategies have a drawback on the complementarity issue, that is there may be different best combinations for different test samples. Therefore, a dynamic selection scheme is needed to solve this issue.

Dynamic Classifier Selection (DCS) method can selectively adopt one learner for each test instance, which is efficient in terms of computation and the utilization of individual learners. However, this "soft combination" actually dodges the complementary issue, rather than solving it. Hopefully, the mixture-of-experts posses the ability to solve the complementary issue by employing gating module to decide different combinations for different test samples. Unfortunately, the mixture-of-experts normally works in a divide-and-conquer strategy where different learners are trained for different sub-tasks divided from a complex task, which would largely departure from our original intention to utilize the experts' strengths together.

In this work, we propose an effective Dynamic Features Combination method using Agreement, named DFCA, which takes the agreements into consideration and can also well address the complementary issue. We choose CNNs as the target individual learners and design specific DFCA methods. Different from DCS, our method dynamically selects a subset of the features for each test instance, which provides the possibility to find the combinations of learners with the maximum complementarity. Different from ME, the mechanism behind DFCA is to iteratively update the combinational weights of each individual features according to their agreements with no complex constraints.

We evaluate DFCA using different groups of frequently-used CNNs on the CIFAR-10 dataset. The performances show the effectiveness and promising ability of the proposed method. The paper is organized as follows. In section 2 briefly introduce the related work. Our approach is presented in section 3. Section 4 gives our experimental results. The last section contains some conclusions and some future works.

2. Related Work

Deep learning architectures such as deep neural networks, deep belief networks and recurrent neural networks have been applied to the fields including computer vision, speech recognition, natural language processing, audio recognition, social network filtering, machine translation, bioinformatics and drug design, where they have produced results comparable to and in some cases superior to human experts. In the past several years, researchers design at least ten more widely used deep neural networks [1-5] for image classification problems. This paper aims to learn dynamic combinations of individual CNNs for image classification.

Mixture-of-experts [7] provides a much better way of finding different combinations of features dynamically. In contrast to typical ensemble methods where individual learners are trained for the same problem, the mixture-of-experts works in a divide-and-conquer strategy where a complex task is broken up into several simpler and smaller subtasks, and individual learners (called experts) are trained for different subtasks. Gating is usually employed to combine the experts.

Routing mechanism always plays an important role in human brain [8] and further motivates Hinton to propose “capsule” [9]. He found that most neuroanatomy supports the existence of cortical minicolumn (most mammals, especially primates). A cortical minicolumn works in a dynamic routing mode, where the lower neurons “select” the upper neurons with the high agreement on the specific information and combine them together to activate a lower neuron. This way dynamically selects useful information to combine.

3. Our Approach

In this paper, we design a novel dynamic features combination method, named DFCA, by adopting the agreement of individual features. DFCA has three key steps (as show in figure 1), i.e., extracting features and projecting them into a common space, iteratively updating the combination coefficients using a dynamic routing module, and making prediction. The projecting step is necessary because different CNNs may output features with various dimension. We use averaging pooling and 1*1 convolution to map all these features in different space to a common space. The second step is the key of addressing the complementarity issue, which is accomplished by a dynamic routing module. The prediction decision is based on the length of the final encoded vector of different classes.

3.1. Feature Mapping

As CNNs have been proved to be powerful for vision tasks and can be trained end-to-end, we choose various CNNs as our base feature extractors, for example E1, E2, E3 in figure 1. Firstly, we trained them independently, and then use the pre-trained model directly for feature extraction.

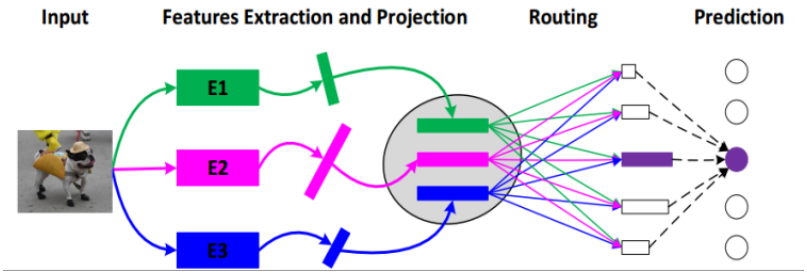

Figure 1. Structure of DFCA.

The feature extraction step extracts features with different sizes and channels using the pre-trained CNNs. The projection step aims to project these features into a common space, where average pooling and 1*1 convolution are adopted. Firstly, average pool is applied to the features if the size the feature map is not 1*1. Then, features with different channels are projected to a common space by applying 1*1 convolution to obtain output channels with the same dimension. As shown in figure 1, three feature vectors with different sizes are projected to a common space (gray circle).

3.2. Target Encoding

Common CNNs for image classification finally output a large vector and feed it to a softmax layer. Differently, we encode information into vectors at the final step before prediction in our approach. This is done by applying a mapping matrix to the feature vectors in the common space.

Just like conventional vector representations, we use the encoded vector to characterize two key parts: (1) using the length to characterize the probability that the entity (object, visual concept, or part of it) represented by the vector is present in the input image. Thus, the length of these vectors will be the prediction rules; (2) using the direction (length-independent) to characterize some of the objects graphic properties like position, color, orientation, shape, etc.

3.3. Routing by Agreement

Given the prepared features in the common space and defined target vectors, we use a routing module to link them together. The routing module is the key to address the complementary issue according to their agreement.

Commonly, features extracted from different CNNs have different scales, which would bring unbalance in combination. To solve this imbalance, we use a non-linear “squashing” function [10] to ensure that short vectors get shrunk to almost zero length and long vectors get shrunk to a length slightly below 1. The prepared feature vectors should be squashed by and then mapped independently using. Then the routing procedure is conducted to update c_{j1}, c_{j2}, c_{j3} for this target class.

The routing procedure is shown as below. The first step is the target encoding, is the encoding matrix. The second step is the combination of all the given features with the initial condition, which will obtain, where and C is the number of target classes. Then, we define the agreement of with the target class as using their dot product. For each target class, the coupling coefficients iteratively update N times according to the agreement of the individual features with the target class.

Assume that there are C classes in all, we will get C output vectors. Based on the assumption in Section, the length of the target vectors represents the probability of the existence of an object in the input sample. For the end-to-end training, we adopt a margin loss defined in, which is similar to Ref. [10].

4. Experiments

We conduct extensive experiments on CIFAR-10 in our experiments and compare with two strong baselines.

The CIFAR-10 dataset consists of colored natural scene images, with 32*32 pixels each and totally 10 classes. The training and test sets contain 50,000 and 10,000 images, respectively. We adopt a

similar data augmentation scheme that is widely used in Refs. [4, 5, 11]. For preprocessing, we normalize the data using the channel means and standard deviations. For the final run we use 50,000 training images and report the final test error at the end of training.

We take two strong baselines: Majority Voting and Overall Averaging. Policy chooses the prediction with the most agreement of all the experts. For example, if there are 4 experts make the same prediction (the agreement is 4 for this prediction) and others experts' agreement is less than 4, Majority Voting will take the prediction as the final decision.

4.1. Individual CNNs

We choose 10 widely used CNNs for CIFAR-10 dataset. As shown in table 1, the performances of all the CNNs are different although they are trained on the same training dataset. That is to say, CNNs with different structures learns different features.

To better understand the diversity of these models, table 1 presents the top-1 errors of the models (0-9) on CIFAR-10 dataset. It is easy to see that different CNNs achieve different performances on different classes. For example, the MobileNet's performance varies from class to class. Thus, it is believed that our method would achieve better performances.

4.2. Performances and Comparation with Baselines

The results are shown in table 1. The first column is the ID of CNNs, for example, "0" indicates VGG19. For combinations, "9160" means we take four CNNs SENet, ResNet18, MobileNet, and VGG19 as the base feature extractors. The above part of table 1 lists the top-1 accuracies of the given 10 CNNs. They are all trained from scratch under totally same settings of the training process.

Table 1. Top-1 Accuracy (%) on CIFAR-10.

ID	Model Name	Top-1 Accuracy
0	VGG19	93.48
1	ResNet18	94.83
2	PreActResNet	94.85
3	GoogleNet	95.06
4	DenseNet121	95.05
5	ResNeXt29_2x64d	95.27
6	MobileNet	88.91
7	DPN92	94.81
8	ShuffleNet	90.71
9	SENet	94.66
	Majority Voting	94.79
9160	Overall Averaging	95.02
	DFCA-32	**95.37**
	Majority Voting	95.53
6809125	Overall Averaging	95.66
	DFCA-32	**95.71**

Majority voting just chooses the class with more voters. Overall averaging applies averaging to the given classifiers. These are two strong baselines for model combination. However, the majority voting performance may be affected by bad classifiers, because the prediction decision is made on the number of classifiers. Overall averaging is simple and effect way for combination.

As shown in table 1, the accuracy of the proposed DFCA is much better than the individual CNNs, and is always slightly better than the two baselines. This proves the effectiveness of the proposed

DFCA, as it is based on the dynamic routing mechanism. The two baselines are both direct ensemble ways, whereas our method is trained end-to-end, including the mapping layer, the encoded layer and the weighted coefficients.

Furthermore, the proposed DFCA can dynamically select features with high agreement and should perform better for combinations with large diversity. The performance of DFCA also proves the agreement works well for the combinations of CNN features.

Table 2. Top-1 Accuracy (%) of the DFCA on CIFAR-10.

ID	8	32	128	512	1024
9160	95.27	95.37	95.14	95.25	95.02
6809125	95.56	95.71	95.63	95.6	95.68

The dimension of the common space is a major hyper parameter of DFCA. According to the performance of different settings of this parameter in table 2, its effect varies. As larger dimension means higher cost of memory and computation, it is better to choose a small value of this parameter, for example, 32 or even 8 both works very well.

4.3. Discussion

Despite the widely held belief that the ensemble or combination based system has matured, the field seems to be enjoying a growing attention by all the researcher and technicians [12-14], especially in performance-pursuit systems with large computation power. The question “which ensemble generation or combination rule is the best?” continuously inspires us to search better solutions. It shows the ratio of images with at least n experts making identical and right predictions, where experts are listed in table 1. For example, the ratio of images with at least “1” image is right is 99.08%, which means that the top-1 error is less than 1% on CIFAR-10. However, the theorem [15] calls attention to us not to be against “blind optimism” on the one hand, and to be curious about exploring new ideas.

The proposed DFCA is such an idea to explore the agreement between different features and a specific target. We believe that both individuality and commonality exist in all the classifiers. For example, all the CNNs can classify a "car" but can hardly recognize a "cat", which is the commonality of them. Voting is the direct way to make use of this commonality, but cannot solve the mutual flaws of the individual classifiers.

5. Conclusion

This paper presents a novel feature combination method, named DFCA, based on the agreement between the individual features to a specific target. Based on a routing module, DFCA can dynamically update the weighted coefficients of individual features, which can well address the complementary issue of common combination methods. On the other hand, DFCA takes the agreement of not only the commonalities, but also the individualities of different features and achieves promising performance on CIFAR-10 classification task. Future work includes searching better initialization conditions of DFCA and novel agreement mechanisms.

Acknowledgement

This work was supported by the State Grid Corporation Science and Technology Project (No. 5200-201916261A-0-0-00)

References

[1] Krizhevsky A, et al. 2012 Imagenet classification with deep convolutional neural networks *Advances in Neural Information Processing Systems* pp 1097-1105.

[2] Simonyan K 2015 Very deep convolutional networks for large-scale image recognition *International Conference on Learning Representations* pp 1-9.

[3] Szegedy C, et al. 2015 Going deeper with convolutions *IEEE Conference on Computer Vision and Pattern Recognition* pp 1-9.
[4] He K, et al. 2016 Deep residual learning for image recognition *IEEE Conference on Computer Vision and Pattern Recognition* pp 770-778.
[5] Huang G, et al. 2017 Densely connected convolutional networks *IEEE Conference on Computer Vision and Pattern Recognition* pp 21-26.
[6] Hu J, et al. 2017 Squeeze-and-excitation networks *IEEE Transactions on Pattern Analysis and Machine Intelligence* **8** (6) 679-698.
[7] Jacobs R, et al. 2014 Adaptive mixtures of local experts *Neural Computation* **3** (1) 79-87.
[8] Massouh M, et al. 2010 De-routing neuronal precursors in the adult brain to sites of injury: role of the vasculature *Neuropharmacology* **58** (6) 877-83.
[9] Hinton G, et al. 2011 Transforming auto-encoders *International Conference on Artificial Neural Networks* **1** (1) 44-51.
[10] Hinton G, et al. 2017 Dynamic routing between capsules *Conference and Workshop on Neural Information Processing System* **30** (1) 3856-3866.
[11] Huang G, et al. 2016 *Deep Networks with Stochastic Depth* (Springer International Publishing) pp 646-661.
[12] Ghosh J, et al. 2002 *Multiclassifier Systems: Back to the Future* (Berlin Heidelberg: Springer) pp 1-8.
[13] Wang B, et al. 2011 Elite: Ensemble of optimal input-pruned neural networks using trust-tech *IEEE Transactions on Neural Networks* **22** (1) 96-109.
[14] Polikar R 2006 Ensemble based systems in decision making *IEEE Circuits & Systems Magazine* **6** (3) 21-45.
[15] Wolpert D, et al. 1997 No free lunch theorems for optimization *IEEE Transactions on Evolutionary Computation* **1** (1) 67-82.