Dynamic Conditional Networks for Few-Shot Learning

Fang Zhao[1] [0000−0002−6772−8042] *, Jian Zhao[1,2] [0000−0002−3508−756X] *†,
Shuicheng Yan[1,3], and Jiashi Feng[1] [0000−0001−6843−0064]

[1] National University of Singapore, Singapore, Singapore
elezhf@nus.edu.sg zhaojian90@u.nus.edu {eleyans,elefjia}@nus.edu.sg
[2] National University of Defense Technology, Hunan, China
[3] Qihoo 360 AI Institute, Beijing, China

Abstract. This paper proposes a novel Dynamic Conditional Convolutional Network (DCCN) to handle conditional few-shot learning, i.e, only a few training samples are available for each condition. DCCN consists of dual subnets: DyConvNet contains a dynamic convolutional layer with a bank of basis filters; CondiNet predicts a set of adaptive weights from conditional inputs to linearly combine the basis filters. In this manner, a specific convolutional kernel can be dynamically obtained for each conditional input. The filter bank is shared between all conditions thus only a low-dimension weight vector needs to be learned. This significantly facilitates the parameter learning across different conditions when training data are limited. We evaluate DCCN on four tasks which can be formulated as conditional model learning, including specific object counting, multi-modal image classification, phrase grounding and identity based face generation. Extensive experiments demonstrate the superiority of the proposed model in the conditional few-shot learning setting.

Keywords: Conditional Model · Few-Shot Learning · Deep Learning · Dynamic Convolution · Filter Bank

1 Introduction

A conditional model is a significant machine learning framework which can be exploited in many tasks, such as multi-modal learning and conditional generative models. It usually contains two inputs. One is interest of task, and the other one is conditional input and provides additional information of specific situation. Recently deep conditional models have attracted much attention since deep neural networks have achieved unprecedented advances in many important fields, such as computer vision [13, 15], natural language processing [37, 19] and speech recognition [26, 1]. However, they generally suffer performance decline in the challenging **conditional few-shot learning** scenario, where training samples for each condition are limited due to the high dimension of the condition space although the total number of training samples can be large.

* indicates equal contribution.

† Jian Zhao is the corresponding author, homepage: https://zhaoj9014.github.io/.

Deep learning based methods typically require a huge amount of labelled data for training as well as specialized computational platform and optimization strategies to achieve satisfactory performance. Their performance usually drops severely for learning problems with small training sample size due to severe over-fitting issues. In contrast, humans, even children can grasp a new concept (*e.g.*, a "giraffe") remarkably fast, "sample efficiently" and generalize to novel cases reasonably from just a short exposure to few examples (*e.g.*, pictures in a book) [4, 20]. This phenomenon motivates the research on the problem of *few-shot learning*, *i.e.*, the task to learn a new concept on the fly, from a few or even a single annotated example for each category [3, 36].

Few-shot learning is of great significance both academically and industrially, since 1) models excelling at this task would help alleviate expensive and labour-intensive data collection and labeling as they would not require massive labelled training data to achieve reasonable performance; 2) the target data in practice usually have a large number of different categories but very few examples per category. For instance, when operating in natural environments, robots are supposed to recognize many unfamiliar objects after seeing only few examples for each [17]. The ability of generalizing in such scenarios would be beneficial to modeling the practical data distribution more effectively.

In this paper, we mainly focus on improving two kinds of models in the conditional few-shot learning scenario, *i.e.*, the discriminative one and the generative one. The discriminative models often resort to hand-crafted features with huge human-engineering efforts and then adopt metric learning algorithms or data-driven deep learning solutions from ample labelled data. However, such data-driven methods are too computationally complex to meet practical applications. Moreover, massive labelled training data covering all underlying variations are usually expensive and unavailable. The generative models often leverage data generative models, *e.g.*, Generative Adversarial Networks (GANs) [10], Conditional Generative Adversarial Networks (Conditional-GANs) [24], Boundary Equilibrium Generative Adversarial Networks (BE-GANs) [2], *etc.*, for synthesizing auxiliary training data for data augmentation. However, among current generative methods, the quality of synthesized data is still far from being satisfactory to perform practical analysis tasks.

In order to address the challenging and realistic conditional few-shot learning problems, we explore a novel approach to learn a deep conditional model from a few labeled examples of each condition, which can generalize well to other cases of the same condition. The conditions could be based on category labels, on some part of data, or even on data from different modalities. Moreover, to enable on-the-fly computation with high efficiency, we embody this conditional few-shot learning problem into learning dual subnets jointly in an end-to-end way. One subnet is called DyConNet, which contains a **Dy**namic **Conv**olutional layer with a bank of trainable basis filters. Given any **Condi**tional input, the other subnet, called CondiNet, predicts a set of adaptive weights to linearly combine the basis filters. In this manner, a specific convolutional kernel can be dynamically obtained for each conditional input, as illustrated in Fig. 1. Dur-

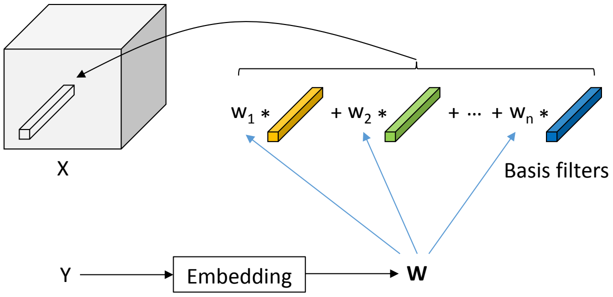


Fig. 1. Dynamic convolutional layer of DyConvNet. It has a filter bank consisting of several basis filters. A set of adaptive weights $\mathbf{W} = \{\mathrm{w}_1, \mathrm{w}_2, \cdots, \mathrm{w}_n\}$ is predicted by the embedding of CondiNet from conditional inputs Y to perform the linear combination on the basis filters, which produces the convolution filters applied on feature maps X

ing optimization, the filter bank is shared between all conditionals thus only a low-dimension weight vector needs to be leaned for each condition, which significantly compensates the limited information in few-shot setting and facilitates the sample-efficient parameter learning across conditions. We term this model as **D**ynamic **C**onditional **C**onvolutional **N**etwork (DCCN). We evaluate DCCN on four distinct tasks, all of which can be formulated as conditional model learning, including specific object counting, multi-modal image classification, phrase grounding and identity based face generation. The proposed DCCN outperforms other discriminative and generative conditional models for all the tasks.

Our contributions in this paper are summarized as follows. (1) We present a novel and effective deep architecture, which contains a **Dy**namic **Conv**olutional sub**Net** (DyConvNet) and a **Condi**tional sub**Net** (CondiNet) that jointly perform learning to learn in an end-to-end way. This deep architecture provides a unified framework for efficient conditional few-shot learning. (2) The dynamic convolution is achieved through linearly combining the basis filters of the filter bank in the DyConvNet with a set of adaptive weights predicted by the CondiNet from conditional inputs, which is different from existing conditional learning approaches that combine the two inputs through direct concatenation. (3) Our architecture is general and works well for multiple distinct conditional model learning tasks. The source codes as well as the trained models of our deep architecture will be made available to the community.

2 Related Works

Our work is related to several others in the literature. However, we believe to be the first to look at methods that can learn the parameters of deep conditional models in the few-shot setting.

Since its inception, few-shot learning has been widely studied in the context of generative approaches. The real annotated data covering all variations are expensive to achieve, even impossible, thus synthesizing realistic data is beneficial for more efficiently training deep models for few-shot learning, by augmenting the number of samples with desired variations and avoiding costly annotation work [40, 39]. Successful generation from limited labelled training samples usually requires carefully tuned inductive biases using additional available information due to the high dimensionality of the feature space [14]. Such additional information can be accessed through various ways. For instance, 1) more samples of categories of interest can be obtained from huge amount of unlabelled data as in semi-supervised learning [42, 6]; 2) the available labelled training data can be augmented using simple transformations, such as jittering, noise injection, *etc.*, as commonly used in deep learning [7, 8, 18]; 3) samples from other relevant categories can be utilized through transfer learning to assist parameter learning [21]; 4) new virtual samples can be synthesized, either rendered explicitly with GAN-based techniques [10, 24, 2] or created implicitly through compositional representations [25, 41]. Recently, Mehrotra *et al.* [23] argued that having a learnable and more expressive similarity objective is an essential missing component, and proposed a network design inspired by deep residual networks that allows the efficient computation of this more expressive pairwise similarity objective. These approaches can significantly advance the performance of few-shot learning if a generative model that accounts for the underlying data distribution is known. However, such a model is usually unavilable and the generation of additional real or synthesized samples often requires substantial efforts.

A different trend of approaches to few-shot learning is to learn a discriminative embedding space, which is typically done with a siamese network [5]. Given an exemplar of a novel category, recognition is performed in the embedding space by a simple rule such as nearest-neighbor. Training is usually performed by classifying pairs according to distance [9], or by enforcing a distance ranking with a triplet loss [27]. A variant is to combine embeddings using the outer-product, which yields a bilinear classification rule [22]. Built on the advances made by the siamese architecture, Vinyals *et al.* [33] employed ideas from metric learning based on deep neural features and from recent advances that augment neural networks with external memories. They proposed a framework which learns a network that maps a small labelled support set and an unlabelled example to its label, obviating the need for fine-tuning to adapt to new class types. Ravi and Larochelle [31] proposed an Long Short Term Memory (LSTM) based meta-learner model to learn the exact optimization algorithm used to train another learner neural network classifier in the few-shot regime. The parametrization of their model allows it to learn appropriate parameter updates specifically for the scenario where a set amount of updates will be made, while also learning a general initialization of the learner (classifier) network that allows for quick convergence of training. However, these methods did not consider conditional model learning and are usually computational expensive for effectively and efficiently solving the few-shot learning problems.

Compared with previous attempts, our proposed method is conceptually simple yet powerful for conditional few-shot learning, which allows learning all parameters from scratch, generalizing across different tasks, and can be seen as a network that effectively "learns to learn". Detailed comparisons with gernerative and discriminative counterparts on various tasks are provided in Sec. 4.

3 Dynamic Conditional Parameter Prediction

Despite the recent success of deep neural networks, it remains challenging to accommodate such models to an extremely large number of categories with limited samples for each, as in the scenario of few-shot learning. Many works to date have mainly focused on learning one-to-one mappings from input to output. However, many interesting problems are more naturally considered as a probabilistic one-to-many mapping. For instance, in the case of image labelling, there may be many different tags that could appropriately be applied to a given image, and different data annotators may use different terms to describe the same image. One way to help address the issue is to leverage additional information from other modalities and to use a conditional model, taking as input small samples and conditional variables, and the one-to-many mapping is instantiated as a conditional predictive distribution.

Since we consider few-shot learning in a conditional modeling task, we start with formulating the standard conditional model learning. It aims to find the parameter W that minimizes the loss $\mathcal{L}$ of a predictor function $h(X|Y;W)$, averaged over N samples x_i and corresponding conditions y_i:

$$\underset{W}{min} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(h(x_i|y_i;W)), \tag{1}$$

where the model can be a discriminative one to learn a classifier or a generative one to learn a conditional distribution over X and Y.

In the case when the dimension of the condition space is too high, the training samples are still scarce for each conditional state even though there are massive training data in total, and the goal is to learn W from small samples with the condition y of interest, called conditional few-shot learning. The main challenge in conditional few-shot learning is to find a mechanism to incorporate domain-specific information into the network. Another challenge, which is of practical importance in applications of few-shot learning, is to enhance efficiency of optimization for Eqn. (1).

We propose to address both challenges by learning the parameter W of the predictor from small samples with the conditions y using a meta-learning process, *i.e.*, a non-iterative feed-forward function φ (meta learner) that maps $(y;W')$ to an optimal W of the predictor (base learner). We parameterize this function using a neural network model and we call it a CondiNet. The CondiNet output depends on the condition y which is a representative of the condition of interest, and contains parameter W' of its own. We train the CondiNet as follows such that it can produce suitable W for different tasks. We optimize the CondiNet using the following objective function. The feed-forward CondiNet evaluation is much faster than solving the optimization problem of Eqn. (1).

$$\min_{\varphi} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(h(x_i; \varphi(y_i; W'))). \tag{2}$$

Importantly, the parameters of the original W of Eqn. (1) now adapt dynamically to each conditional input y. Note that the training scheme is reminiscent of that of siamese networks [5] which also employ dual subnets. However, siamese networks adopt the same network architecture with shared weights, and compute the inner-product of their outputs to produce a similarity score:

$$\min_{W} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(\langle h(x_i; W), h(y_i; W) \rangle). \tag{3}$$

There are two key differences with our model: 1) we treat $h(\cdot)$ and $\varphi(\cdot)$ asymmetrically, which results in a different objective function; 2) more importantly, the output of $\varphi(y; W')$ is used to parametrize convolutional layers that determine the intermediate representations in the network $h(\cdot)$ dynamically. This is significantly different from siamese networks [5] and bilinear networks [22], as well as traditional conditional networks [24] based on the conditional probability $p(X|Y; W)$.

Now, we explain the implementation of the CondiNet $\varphi(\cdot)$ and the main predictor $h(\cdot)$ formally. Given an input tensor $x \in \mathbb{R}^{p \times q \times c}$, weights $W \in \mathbb{R}^{k \times k \times c \times d}$ (where k is the kernel size), and biases $b \in \mathbb{R}^d$, the output $f \in \mathbb{R}^{p' \times q' \times d}$ of a convolutional layer is given by

$$f = W * x + b, \tag{4}$$

where $*$ denotes convolution operation, and the biases b are applied to each of the d channels.

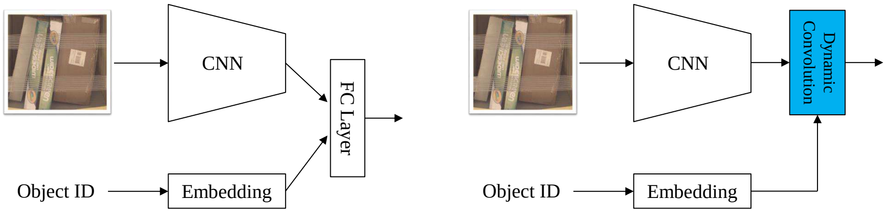
We propose to formulate the weights and biases as functions of y, $W(y)$ and $b(y)$, to represent the dynamic conditional parameters given the conditional input $y \in \mathbb{R}$:

$$f = W(y) * x + b(y). \tag{5}$$

While Eqn. (5) seems to be a straightforward drop-in replacement for convolutional layers, careful analysis reveals that it scales extremely poorly. The main reason is the typically high dimensionality of the output space of the CondiNet $\varphi(\cdot) : \mathbb{R} \rightarrow \mathbb{R}^{k \times k \times c \times d}$. Since k is usually small and so is k^2, for a comparable number of input and output channels in a convolutional layer ($c \simeq d$), the output space of the CondiNet grows quadratically with the number of channels. Overfitting issues, memory and time costs make learning such a regressor difficult in few-shot learning settings.

In order to address the above-mentioned issue when learning a conditional model in few shot, we herein propose a simple yet effective method to reduce the output space by considering a decomposition as below (we drop the bias term b for simplification),

$$f = \sum_{i=1}^{n} ({w'}_i(y) \cdot w_i) * x, \tag{6}$$



(a) Plain Conditional Network

(b) Dynamic Convolution Based Conditional Network

Fig. 2. Flow charts of object counting based on specific object ID

where $w \in \mathbb{R}^{k\times k\times c\times d}$, $w'(y) \in \mathbb{R}$, the product $({w'}_i(y) \cdot w_i)$ can be seen as a decomposed representation of convolutional kernels, to linearly combine the adaptive weights ${w'}_i$ predicted by the CondiNet $\varphi(\cdot)$ and the basis filters w_i of a filter bank, and n denotes the number of basis filters.

The filter bank is shared between all conditional states and only the weights of basis filters are specific for each conditional state. Both w and w' contain trainable parameters, but they are modest in size compared to the case discussed in Eqn. (5). Importantly, the CondiNet $\varphi(\cdot)$ now only needs to predict a set of adaptive weights, so its output space grows linearly with the number of basis filters in the filter bank (*i.e.*, $\varphi(\cdot) : \mathbb{R} \rightarrow \mathbb{R}^n$). Since the resulted convolutional kernel $({w'}_i \cdot w_i)$ in Eqn. (6) is dynamically changed, depending on the prediction of the CondiNet $\varphi(\cdot)$ and the filter bank of the main predictor $h(\cdot)$, we construct $h(\cdot)$ as another subnet — DyConvNet. The dual subnets operate cooperatively for jointly learning parameters of a deep conditional model with conventional chain rules and **B**ack **P**ropagation (BP) algorithms in few shot.

4 Experiments

We evaluate our model on four conditional few-shot learning problems to verify the effectiveness of dynamically combining the basis filters, including specific object counting, multi-modal image classification, phrase grounding and identity based face generation.

4.1 Specific Object Counting

The specific object counting task is from Amazon Bin Image Dataset (ABID) Challenge, which is to predict the quantity of the object in a bin, given an image and the target category. When the maximal quantity of an object in a bin is set to a constant (here is 5), we formulate this task as a conditional classification model by viewing the object category as a conditional input. As shown in Fig. 2 (b), one network is used to extract image features, and the other network is used to embed the object ID. Finally, our dynamic conditional layer is used

Table 1. Accuracies of object quantity verification and identification on the Amazon Bin Image dataset

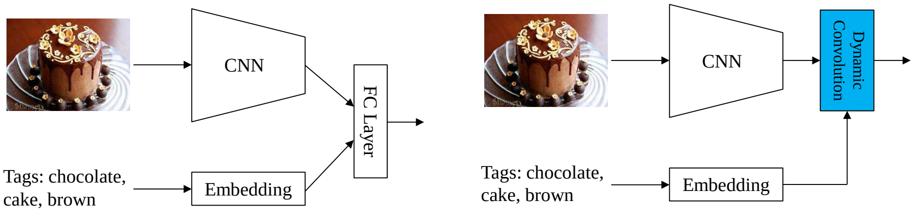
Methods	Dynamic			Plain
	4-D	8-D	16-D	
Identification	76.60%	**76.81**%	75.66%	74.48%
Verification	85.39%	**85.48%**	84.81%	84.87%

to combine the last layers of the two networks to output the quantity of the object. Here we use a plain conditional network (Fig. 2 (a)) as a baseline, i.e., directly concatenating the last layers and substituting our dynamic conditional layer with a fully-connected layer.

Dataset and evaluation metric. We evaluate our model on two subtasks, i.e., object quantity verification and identification. The former is to verify whether the given object quantity is correct for a bin image. The latter is to directly count the objects in a bin image. The dataset contains 535,234 bin images and is divided into two subsets, 481,711 images for training and the remaining images for test. For the object quantity verification, we test on triplets of image, object ID and quantity. The accuracy is used to measure the performance of both the tasks.

Architecture and training. Similar with the model architecture settings provided by the dataset website, we use the ResNet-34 network to extract image features. The embedding dimension of the object ID in the plain conditional network is 512. The dimension of the dynamic conditional layer is set to 4, 8 and 16 respectively to investigate effects of using different numbers of basis filters. All images are resized into 224x224 for convenient training and comparison. Because it is actually a classification task, the Softmax loss is adopted to optimize the entire network. We train for 30 epochs. The initial learning rate is 0.1 and it is dropped by a factor of 10 every 10 epochs.

Results and analysis. Table 1 reports accuracies of our method under various dimensions of the dynamic convolutional layer and the plain conditional network. One can see that our network using 8-D dynamic layer achieves the best accuracy. The plain conditional network performs not well because the set of object IDs is too large and each ID is only associated with few training examples (one example for most IDs). It is hard to learn a conditional network for each ID when the embedding dimension is too high, and the network coditional output would not be discriminative enough if the embedding dimension is low. In contrast, the proposed DCCN makes different IDs share a filter bank. Only a low-dimension vector is needed to learn to combine the set of filters as a convolutional kernel. Through applying this kernel on the top layer of the feature network, the spatially local correlation of image and object ID can be learned to make the conditional output more discriminative for different IDs. Note that as the dimension of dynamic layer continuously increases, such as 16, the performance decreases instead due to overfitting.



(a) Plain Multi-modal Classification

(b) Dynamic Convolution Based Multi-modal Classification

Fig. 3. Flow charts of multi-modal image classification

4.2 Multi-modal Image Classification

Multi-modal classification can be formulated as a typical conditional model consisting of two networks. Fig. 3 (a) shows a general framework of multi-modal classification. The inputs of the two network are an image and text describing the image, respectively. The texts are usually transformed into bag-of-words vectors at first. The outputs of the networks are then combined into a feature vector through a fully connected layer. We use this plain conditional network as a baseline. Fig. 3 (b) illustrates the proposed dynamic convolution used for multi-modal classification, where we substitute the fully connected layer with the dynamic conditional layer.

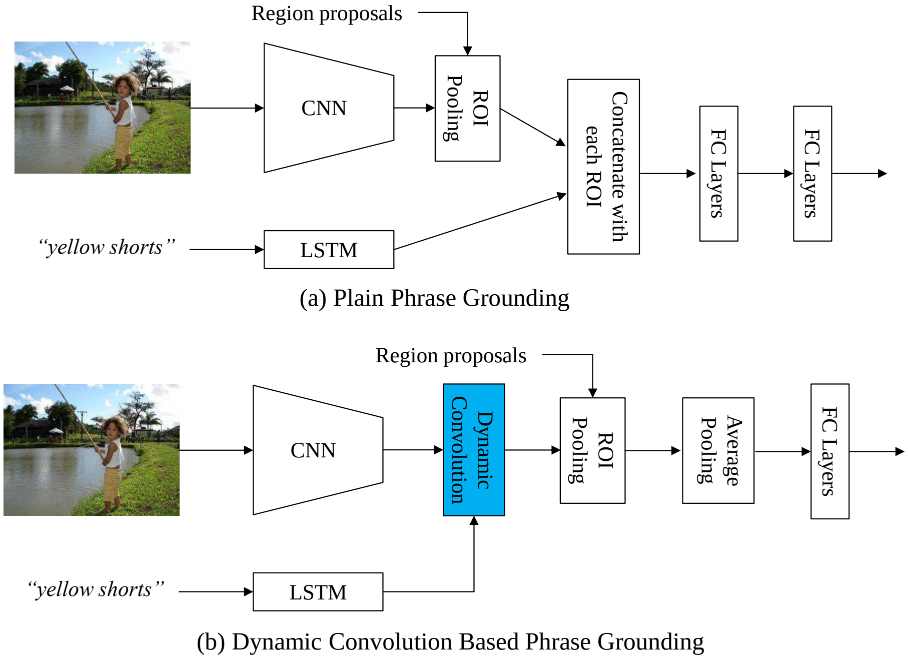
Dataset and evaluation metric. We evaluate our model on the MIRFlickr-25K dataset [16] which consists of 25,000 images downloaded from the social website Flickr. Each image associates with some of 20,000 tags. 38 class labels including various scenes and objects, such as sunset, car and bird, are used to annotate these images and an image may belong to multiple class labels. We randomly sample 20,000 images for training and the rest for testing. The multi-label classification performance is measured by the Intersection over Union (IoU) in the multi-label setting, which is defined as the number of correctly predicted labels divided by the union of predicted and ground-truth labels.

Architecture and training. The base network for both the baseline and our method is the ResNet-34 network. The embedding dimension of the text in the baseline is 512. We set the dimensions of the dynamic conditional layer to 32, 64 and 128 respectively. To deal with the multiple labels in one image, here we adopt the cross entropy loss to learn the conditional networks. The training epoch is 90. Beginning with 0.1, the learning rate is dropped by a factor of 10 every 30 epochs.

Results and analysis. Table 2 reports the results of different methods on the MIR-Flickr25K dataset. It can be observed that when the dimension of the dynamic layer is 64, our dynamic conditional network outperforms the baseline under the condition of various numbers of tags. We argue that although the total training images are sufficient, there are only a few ones for each tag. That is to say, it is still a few-shot learning for each tag. Thus the dynamic layer which

Table 2. IoU of multi-modal multi-label classification on the MIR-Flickr25K dataset

Methods		Dynamic			Plain
		32-D	64-D	128-D	
Tags	5k	0.6517	**0.6553**	0.6520	0.6489
	10k	0.6513	**0.6606**	0.6560	0.6516
	20k	0.6549	**0.6577**	0.6543	0.6490



(a) Plain Phrase Grounding

(b) Dynamic Convolution Based Phrase Grounding

Fig. 4. Flow charts of phrase grounding

reduces the parameters of the conditional network is able to address the problem of overfitting effectively for this kind of few-shot conditional learning, and the filter bank shared by tags can be learned easily by using all training images.

4.3 Phrase Grounding

The task of phrase grounding is to localize objects or scenes described by text phrases in images [32, 29]. This task can also be modeled as a conditional model. A typical framework of phrase grounding is illustrated in Fig. 4 (a). One convolutional neural network is used to produce a spatial feature map of an input image, and Long Short-Term Memory network (LSTM) [11] is used to embed an input phrase into a vector with fixed length. Then the features of a set of region proposals (i.e.,Edge Boxes [43]) are extracted by applying the ROI pooling on

Table 3. Accuracy (IoU $>$ 0.5) of phrase grounding on the Flickr30k Entities dataset

Methods	Dynamic			SMPL	NonlinearSP	GroundeR
	8-D	16-D	32-D			
Accuracy	50.18	**50.65**	50.52	42.08	43.89	47.81

Table 4. Accuracy (IoU $>$ 0.5) of phrase grounding for various phrase types on the Flickr30k Entities dataset

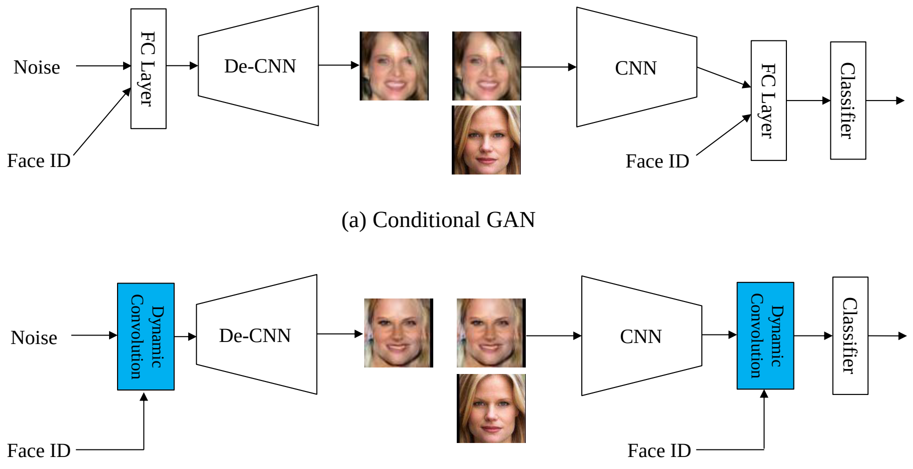
Methods	People	Clothing	Body parts	Animals	Vehicles	Instruments	Scene	Other
SMPL	57.89	34.61	15.87	55.98	52.25	23.46	34.22	26.23
GroundeR	61.00	38.12	10.33	62.55	**68.75**	36.42	**58.18**	29.08
Dynamic	**67.37**	**38.12**	**18.22**	**69.93**	56.04	**37.57**	54.05	**32.59**

the spatial feature map. Finally, the proposal features are concatenated with the phrase vector respectively to compute correlation scores by two fully connected layers. Fig. 4 (b) shows the proposed dynamic convolution based phrase grounding. We firstly use the dynamic conditional layer to combine the phrase vector with the image feature map to obtain a correlative feature map. Then the ROI pooling is applied to obtain the correlative feature map for each region proposal, which is fed into the average pooling and a fully connected layer sequentially to compute the correlation score.

Dataset and evaluation metric. The Flickr30k Entities dataset [30] is used to evaluate our model for phrase grounding, which is an extension of the Flickr30K dataset [38]. It consists of 31,000 images and their captions which are associated with 276,000 manually annotated bounding boxes. We use 2,000 images for testing and the remaining images for training. Following [30], if a single phrase (*e.g.*, rainbow flags) has multiple ground truth bounding boxes, the union of the boxes is used to represent the phrase. If the IoU of an image region predicted for a phrase and the ground truth bounding box is larger than 0.5, the predicted region is deemed correct for the phrase.

Architecture and training. The same with [32], we adopt the VGG-16 network to extract the image feature map, which is pretrained on the PascalVOC dataset for object detection and then is fixed when training the entire conditional model. Both the numbers of the hidden and input units of LSTM are 512. The dimension of the dynamic layer is set to 8, 16 and 32, respectively. 100 region proposals generated by Edge Boxes are used as candidate bounding boxes. We employ the Softmax loss to learn the model to maximize the correlation score of the input phrase with the correct region proposal. We train for 90 epochs. The initial learning rate is 0.01 and every 30 epochs it is dropped by a factor of 10.

Results and analysis. Table 3 reports the accuracy of phrase grounding for different methods under the condition of IoU $>$ 0.5 on the Flickr30k Entities dataset. One can see that our dynamic conditional network achieves the best accuracy compared with the state-of-the-art methods when the dimension of the dynamic layer is 16. NonlinearSP [34] and GroundR [32] have similar frameworks with Fig. 4 (a), i.e., using fully connected layers to combine the features of the



(a) Conditional GAN

(b) Dynamic Convolution Based Conditional GAN

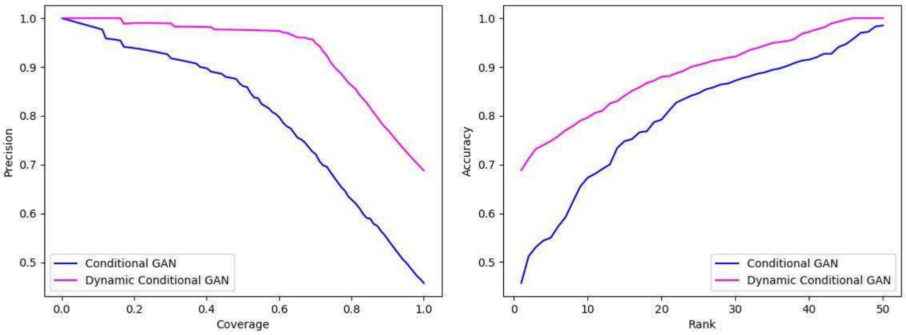
Fig. 5. Flow charts of identity based face generation

image region and phrase. SMPL [35] utilizes a bipartite matching to compute their correlation score. However, all of these methods do not consider that only a few training images are available for each phrase although there are a large number of training images in this dataset. In this sense, this task can be viewed as a conditional few-shot learning problem which can be solved better by our dynamic conditional layer. Table 4 reports the accuracy of phrase grounding for different types of phrases. Our method has better performance than other methods for most phrase types.

Although there are some phrase grounding methods which have better performance than our method, e.g., RtP [28] and SPC+PPC [29], we argue that these methods employ additional cues to improve correlation learning of image region and phrase, such as region-phrase compatibility, candidate position and size. Actually, our model is mostly like a proof-of-concept and applied on the task of phrase grounding to verify its effectiveness on the conditional few-shot learning. It is orthogonal to many technical improvements found in the phrase grounding literature.

4.4 Identity Based Face Generation

The proposed DCCN can also be used to improve conditional generative models. Here we test DCCN on the task of identity based face generation. Fig. 5 (a) shows a general framework based on conditional generative adversarial nets (GAN) [24], which consists of a generative model G and a discriminative model D. In G, the prior input noise and the face ID are combined through a fully connected layer to obtain a joint hidden representation. Then the representation is fed into a deconvolutional neural network to generate a face image of the input ID. In D,



(a) Precision-Coverage curve

(b) Cumulative match characteristic curve

Fig. 6. Precision-Coverage and cumulative match characteristic curves of face identification for identity based face generation.

a convolutional neural network is employed to extract the features of the faces generated by G and the real faces. Then the feature and the embedding vector of the face ID are concatenated and fed into a classifier, which judges whether the face is real or not for this ID. The proposed dynamic convolution based conditional GAN is illustrated in Fig. 5 (b). We use the dynamic conditional layer to integrate the face ID with the noise in D and the image feature in G, respectively.

Dataset and evaluation metric. We evaluate our model on the MS-Celeb-1M dataset [12] which contains about 10M face images for 100K subjects. For the training set, we randomly sample 100 subjects and 10 face images for each subject to simulate the conditional few-shot setting. In testing, given a generated face image, a pretrained face recognition model is used to predict which one of the 100 subjects it belongs to. 50 images are generated for each subject. The precision-coverage (PC) curve and the cumulative match characteristic (CMC) curve are used to measure the performance of face identification.

Architecture and training. We use five-layer fully convolutional and deconvolutional network in the generative and discriminative models, respectively. To learn the conditional GAN, we optimize the generative model G and the discriminative model D alternatively. D is trained to minimize the classification loss under the condition of the input ID, and G is trained to maximize the loss under the same condition, i.e., G trying to generate face images which can confuse G.

Results and analysis. Fig. 6 illustrates the PR and CMC curves of the face identification for generated face images. Table 5 reports the precision when Coverage=0.99 and 0.95 and the accuracies of rank 1 and 5. It can be observed that dynamic conditional GAN achieves better performance than the plain conditional GAN in terms of all the metrics. The dynamic layer can effectively incorporate the information of conditional input through sharing filter bank across conditions when limited training data are available for each condition. Some examples of generated faces are shown in Fig. 7. Each row of faces corresponds to

Table 5. Accuracy and Precision@Coverage of face identification for generated faces

Methods	Accuracy		Pricison@Coverage	
	Rank 1	Rank 5	P@C = 0.99	P@C = 0.95
Plain	0.457	0.550	0.05	0.18
Dynamic	**0.688**	**0.748**	**0.21**	**0.71**

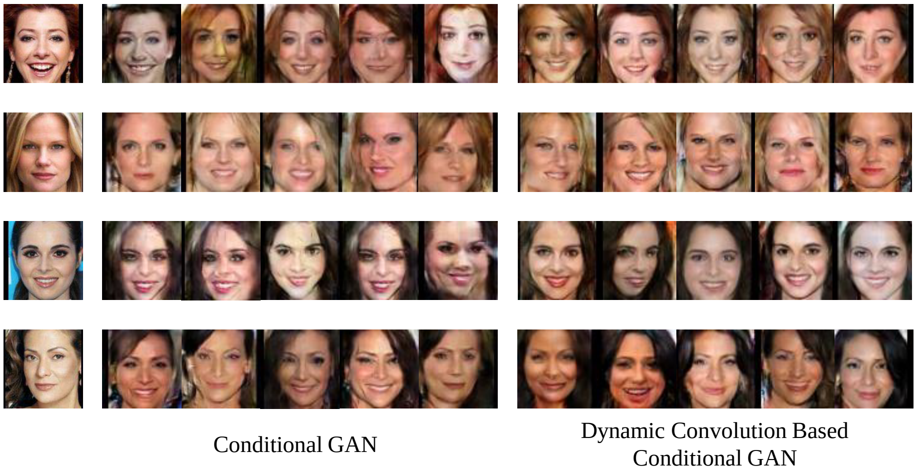


Fig. 7. Examples of identity based face generation

one subject. The faces generated by the dynamic conditional GAN are obviously more similar with the real face of the subject.

5 Conclusion

This paper addressed the problem of conditional few-shot learning. A Dynamic Conditional Convolutional Network is presented to incorporate conditional input in a deep model when only a few training samples are available for each condition. In this model, a set of adaptive weights from conditional inputs is predicted to linearly combine the basis filters of a filter bank shared by all conditions. Then a dynamic convolutional kernel can be obtained according to different conditional inputs. Finally the dynamic kernel is applied on the top layer of the other network to provide conditional output. Qualitative and quantitative experiments on four tasks demonstrate that the proposed model achieves better performance compared with other conditional learning models.

Acknowledgments Jian Zhao was partially supported by China Scholarship Council (CSC) grant 201503170248. Jiashi Feng was partially supported by NUS IDS R-263-000-C67-646, ECRA R-263-000-C87-133 and MOE Tier-II R-263-000-D17-112.

References

1. Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., et al.: Deep speech 2: End-to-end speech recognition in english and mandarin. In: International Conference on Machine Learning. pp. 173–182 (2016)
2. Berthelot, D., Schumm, T., Metz, L.: Began: Boundary equilibrium generative adversarial networks. arXiv preprint arXiv:1703.10717 (2017)
3. Bertinetto, L., Henriques, J.F., Valmadre, J., Torr, P., Vedaldi, A.: Learning feed-forward one-shot learners. In: Advances in Neural Information Processing Systems. pp. 523–531 (2016)
4. Bloom, P.: How children learn the meanings of words. The MIT Press (2000)
5. Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R.: Signature verification using a" siamese" time delay neural network. In: Advances in Neural Information Processing Systems. pp. 737–744 (1994)
6. Chapelle, O., Scholkopf, B., Zien, A.: Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. IEEE Transactions on Neural Networks **20**(3), 542–542 (2009)
7. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: Delving deep into convolutional nets. arXiv preprint arXiv:1405.3531 (2014)
8. Dosovitskiy, A., Springenberg, J.T., Riedmiller, M., Brox, T.: Discriminative unsupervised feature learning with convolutional neural networks. In: Advances in Neural Information Processing Systems. pp. 766–774 (2014)
9. Fan, H., Cao, Z., Jiang, Y., Yin, Q., Doudou, C.: Learning deep face representation. arXiv preprint arXiv:1403.2802 (2014)
10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014)
11. Guadarrama, S., Rodner, E., Saenko, K., Zhang, N., Farrell, R., Donahue, J., Darrell, T.: Long short-term memory. Neural Computation (1997)
12. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: MS-Celeb-1M: A dataset and benchmark for large scale face recognition. In: European Conference on Computer Vision (2016)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
14. Hertz, T., Hillel, A.B., Weinshall, D.: Learning a kernel function for classification with small training samples. In: Proceedings of the international conference on Machine learning. pp. 401–408. ACM (2006)
15. Huang, G., Liu, Z., Weinberger, K.Q., van der Maaten, L.: Densely connected convolutional networks. arXiv preprint arXiv:1608.06993 (2016)
16. Huiskes, M.J., Lew, M.S.: The mir flickr retrieval evaluation. In: Proceedings of the ACM International Conference on Multimedia Information Retrieval (2008)
17. Krause, E.A., Zillich, M., Williams, T.E., Scheutz, M.: Learning to recognize novel objects in one shot through human-robot interactions in natural language dialogues. (2014)
18. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)

19. Kumar, A., Irsoy, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., Zhong, V., Paulus, R., Socher, R.: Ask me anything: Dynamic memory networks for natural language processing. In: International Conference on Machine Learning. pp. 1378–1387 (2016)
20. Lake, B.M., Salakhutdinov, R.R., Tenenbaum, J.: One-shot learning by inverting a compositional causal process. In: Advances in neural information processing systems. pp. 2526–2534 (2013)
21. Lim, J.J., Salakhutdinov, R.R., Torralba, A.: Transfer learning by borrowing examples for multiclass object detection. In: Advances in neural information processing systems. pp. 118–126 (2011)
22. Lin, T.Y., RoyChowdhury, A., Maji, S.: Bilinear cnn models for fine-grained visual recognition. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1449–1457 (2015)
23. Mehrotra, A., Dukkipati, A.: Generative adversarial residual pairwise networks for one shot learning. arXiv preprint arXiv:1703.08033 (2017)
24. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
25. Movshovitz-Attias, Y., Yu, Q., Stumpe, M.C., Shet, V., Arnoud, S., Yatziv, L.: Ontological supervision for fine grained classification of street view storefronts. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1693–1702 (2015)
26. Oord, A.v.d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K.: Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499 (2016)
27. Parkhi, O.M., Vedaldi, A., Zisserman, A., et al.: Deep face recognition.
28. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer imageto-sentence models. International Journal of Computer Vision **123**(1), 74–93 (2017)
29. Plummer1, B.A., Mallya, A., Cervantes, C.M., Hockenmaier, J., Lazebnik., S.: Phrase localization and visual relationship detection with comprehensive image-language cues. In: Proceedings of the IEEE International Conference on Computer Vision (2017)
30. Plummer1, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE International Conference on Computer Vision (2015)
31. Ravi, S., Larochelle, H.: Optimization as a model for few-shot learning (2016)
32. Rohrbach, A., Rohrbach, M., Hu, R., Darrell, T., Schiele, B.: Grounding of textual phrases in images by reconstruction. In: European Conference on Computer Vision (2016)
33. Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al.: Matching networks for one shot learning. In: Advances in Neural Information Processing Systems. pp. 3630–3638 (2016)
34. Wang, L., Li, Y., Lazebnik, S.: earning deep structure preserving image-text embeddings. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
35. Wang, M., Azab, M., Kojima, N., Mihalcea, R., Deng, J.: Structured matching for phrase localization. In: European Conference on Computer Vision (2016)
36. Wang, Y.X., Hebert, M.: Learning to learn: Model regression networks for easy small sample learning. In: European Conference on Computer Vision. pp. 616–634. Springer (2016)

37. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al.: Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 (2016)
38. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics (2014)
39. Zhang, Z., Song, Y., Qi, H.: Age progression/regression by conditional adversarial autoencoder. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
40. Zhao, J., Xiong, L., Jayashree, K., et al.: Dual-agent gans for photorealistic and identity preserving profile face synthesis. In: Advances in Neural Information Processing Systems. pp. 66–76 (2017)
41. Zhu, X., Vondrick, C., Fowlkes, C.C., Ramanan, D.: Do we need more training data? International Journal of Computer Vision **119**(1), 76–92 (2016)
42. Zhu, X.: Semi-supervised learning literature survey (2005)
43. Zitnick, C.L., Dollar, P.: Edge boxes: Locating object proposals from edges. In: European Conference on Computer Vision (2014)