Share Tasks among Experts: MTL-MoE Model for Fine-grained Image Classification

Wang Kai
E0509826@u.nus.edu

Hu Weilin
E0385089@u.nus.edu

Lao Guoyin
E0454518@u.nus.edu

Ma Haozhe
E0509813@u.nus.edu

1 Introduction

In traditional computer vision research, the categories of the target objects in image analysis are usually coarse-grained categories, such as "dog", "car" and "bird". However, in many real-world applications, the target objects often belong to fine-grained categories which are from one common /specific coarse-grained category. Fine-grained image classification [1, 2, 3, 4, 5] consists of discriminating between classes in a sub-category of objects, for instance the particular species of bird or dog.

Fine-grained image classification is a challenging computer vision problem due to subtle differences in the overall appearance between various classes (low inter-class variation) and large pose and appearance variations in the same class (large intra-class variation). See Figure 1 for examples. Mixture of experts (MoE) is one of the approaches proposed to solve this challenge. In [6] and [7], by grouping similar images and assigning them to corresponding expert neural networks, the input space is being partitioned so that an expert network can better learn the subtle differences between similar samples.

Figure 1: Example images from CUB 200-2011 dataset which exhibits large intra-class variations and low inter-class variations. Each column represents a unique class.

In this work, for each input image, our objective is to assign it to experts automatically by applying a sparsed gate layer proposed by [8], so that each expert will focus on different specific poses or appearances across all sub-categories, which will decrease intra-class variation and increase inter-class variation for each expert. Our method avoids both manual partitioning of input images and overfitting problems for each expert. Empirical evaluations show that this approach outperforms single expert model approaches in terms of classification accuracy.

Preprint. Under review.

Multi-Task Learning is another approach we introduced to our work to improve the model performance. Neural-based multi-task learning has been successfully used in many real-world large-scale applications. [9] proposed a novel multi-task learning approach, Multi-gate Mixture-of-Experts (MMoE), which explicitly learns to model task relationships from data.

Based on the assumptions that the region detection can facilitate fine-grained feature learning, we aimed to introduce discriminative region localization tasks to help us achieve better fine-grained classification accuracy. More specifically, we adopted MMoE model structure with two tasks: fine-grained classification task and object bounding box coordinate regression task. Our MTL-MoE model, which combines the advantages of Mixture-of-Experts and Multi-task-learning (MTL), can both decrease intra-class variation while enjoying the benefit of additional discriminative region localization tasks. We conducted comprehensive experiments and show that our MTL-MoE model achieves better classification accuracy and converges faster during training than both single-expert model and single-task model. Over the dataset CUB 200-2011, our best result (80.44%) is better than most of the models published before 2018.

2 Related Work

Fine-grained classification Deep learning based methods has made significant progress in the field of fine-grained classification in recent years [10, 11, 12]. One line of work[13, 14, 15] has concentrated on feature encoding. Lin et al. propose a bilinear pooling method[13] that computes local pairwise feature interactions from two CNN branches (shared or not shared).

Another line of work has focus on extracting discriminative part features in a weakly supervised way. To avoid using extensive annotations, Xiao et al. [16] apply a part-level top-down attention and combine candidates proposal attention, object-level attention to train domain-specific deep nets. Zhang et al. [17] propose to elaborately pick deep filters as part detectors before encoding them to final representation. Spatial transformer networks [18] perform transformation on entire feature map to allows networks to select the most relevant (attention) region. Such part-based methods have become dominant in the field of fine-grained classification.

Mixture of Experts (MoE) is one of the most popular combining machine learning techniques, which has great potential to improve performance. MoE is established based on the divide-and-conquer principle in which multiple experts are used to divide the problem space into homogeneous regions. In most of recently proposed frameworks, MoE sets a gating network to decide which expert to use for each input region, learning thus consists of two parts: learning the parameters of individual expert and the parameters of the gating network[19]. MoE can hold models with a larger number of parameters, deal with a variety of data domains and solve more complex problems[20], which is usually embedded into various models as a layer[8]. Many recent researches based on it have achieved good results and have been successfully applied to many fields, such as classification tasks of multiple scenes[21, 22], translation tasks of multiple languages[23], etc.

MoE has also been proposed to solve fine-grained classification problems. In [24], each expert learns with prior knowledge from the previous expert, so that experts can extract small and large part features, and the gating network determines the contribution of each expert to the final predictions.

In [6] and [7], the images were first partitioned into K non-overlapping sets and K expert systems were learned. By grouping similar images, the input space is being partitioned so that an expert network can better learn the subtle differences between similar samples. However, this input space partition method has two obvious drawbacks. First, in the real applications, it is difficult and tedious to manually perform input images partitioning. Then, since an expert deep neural network can have millions of parameters, training a neural network requires massive amounts of data, and if we do data partitioning, it will also cause serious overfitting for each expert, leading to poor performance on test data.

Multi-Task Learning (MTL) is an approach in machine learning to solve multiple learning tasks at the same time, while sufficiently exploiting commonalities and differences across tasks. This can result in improved learning efficiency and prediction accuracy for the task-specific models, when compared to training the models separately[25, 26]. One of the widely used multi-task learning models is based on a shared bottom model structure, normally, the shared bottom layers mainly aim to extract features and the split high layers aim to do the specific tasks[27]. This structure

substantially reduces the risk of overfitting and reusing the shared layers can efficiently save training time and memory, but can suffer optimization conflicts caused by task differences[26]. This has led to successes in many applications, from natural language processing and speech recognition to computer vision and drug discovery[28].

In [29], the author proposed that in fine-grained classification task, region detection and fine-grained feature learning are mutually correlated and thus can reinforce each other, base on which they proposed recurrent attention convolutional neural network (RA-CNN) which recursively learns discriminative region attention and region-based feature representation at multiple scales in a mutually reinforced way. Enlightened by this, our project introduced a bounding box detection task to help improve the fine-grained classification task.

3 Methodology

Our model is based on a Multi-gate Mixture-of-Experts[9] structure which is shown in Figure 2.

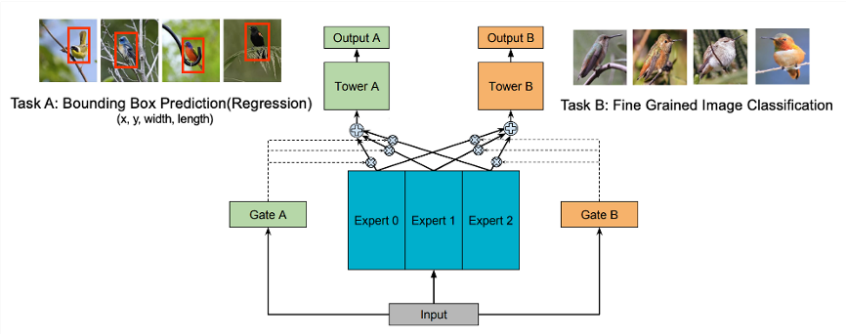

Figure 2: A multi-gate mixture-of-experts (MMoE) model structure

For the shared bottom layers, we are using an MoE model and established a gating network for each task, which is based on a fact that each different task may depend on sum of outputs from experts under different weights, while the objective of the gating networks is to learn the importance of each expert based on a specific task. In our project, we have two tasks, one is classification and another one is finding the bounding box of the target object, so we have two gating networks and for each one, the output of the shared bottom layers can be computed as:

$$y = \sum_{i=1}^{n} g(x)_i f_i(x)$$

where $\sum_{i=1}^{n} g(x)_i = 1$ and $g(x)_i$ is the ith logit of the output of $g(x)$, indicates the probability for expert f_i.

Because different tasks have different outputs, so in the high layers, we build a corresponding tower for each task, that is, a multi-layer perception (MLP) model to further aggregate the extracted features by the bottom layers, and then output the corresponding structure of the specific task, for example, a vector of probabilities for the classification task and four values for the regression task. In transfer learning approach, we usually freeze the MoE part and only train the towers to save training time, which makes it more efficient to be extended to new tasks.

4 Experiments

4.1 Datasets

We demonstrated the empirical performance of our proposed models and methods on Caltech-UCSD Birds-200-2011 (CUB-200-2011) dataset. CUB-200-2011 dataset has become the de facto standard for fine-grained bird classification task, which contains 11,788 images of 200 bird species. Apart from class labels, CUB-200-2011 dataset also has bounding box annotations around the object of interest for each image. We use this information to extract just the object of interest from the image

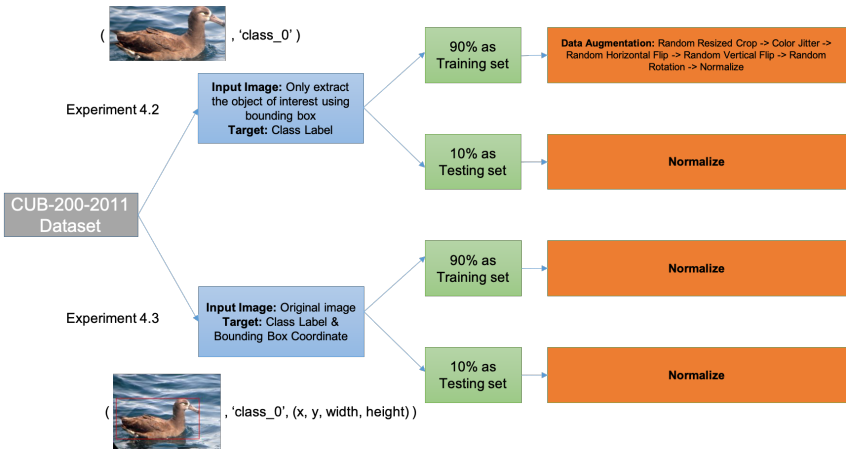

Figure 3: Experiment data preprocessing pipeline

to train our MoE models in experiment 4.2 and use it as the training target in the object bounding box coordinate regression sub-task during training our MTL-MoE models in experiment 4.3.

In both experiment 4.2 and 4.3, we split 90% of the images into training dataset and the rest 10% into testing dataset for each class. In experiment 4.2, for each image, random data augmentation has been performed before the image was involved in model training or evaluation process. The detailed data preprocessing pipeline for two experiments is shown in Figure 4.

4.2 MoE for Fine-Grained Classification

4.2.1 Comparison between Mixture-of-Experts and Single-Expert Model

The structure of the MoE model we used are shown in Figure 4. In this architecture, we use a simple CNN network as the gating network and use EfficientNet-b1 as expert neural networks. Each expert will extract the input's features and output a feature map with shape $\mathbb{R}^{B\times O}$, where B is the batch size and O is the output shape of the experts. The features maps will be put into a feature condense layer and be converted into an output with shape $\mathbb{R}^{B\times f}$, where f is a hyper-parameter and we set it to 1024 in our experiment. After performing a weighted sum, the tower layer will take the new features map as input and output result with shape $\mathbb{R}^{B\times c}$, where c is the number of classes.

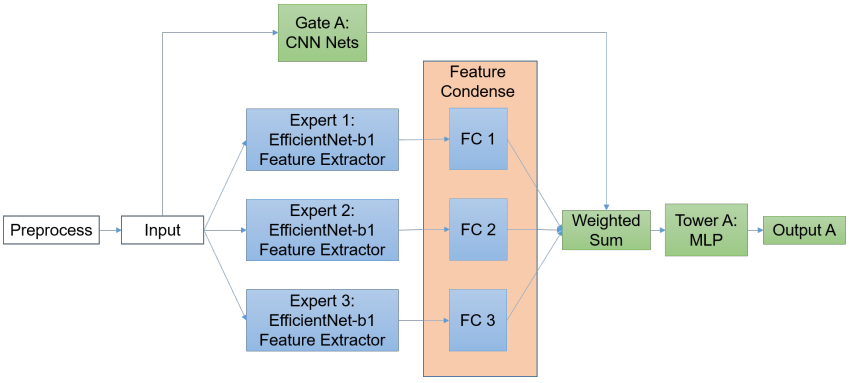

Figure 4: MoE model

We first performed experiments to check whether MoE works better than a single model in the fine-grained classification task. We trained one EfficientNet-b1 and a MoE model with 6 EfficientNet-b1s as experts on 10 classes of the dataset. The results are shown in Figure 5.

According to Figure 5, we found that a single EfficientNet-b1 converges faster than the MoE model. The single model reached an accuracy of 95% in around 2 hours and then got stuck. The MoE model

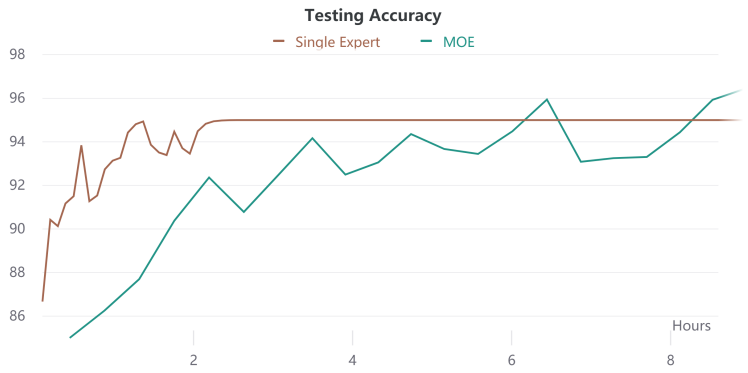

Figure 5: Testing accuracy of MoE and single-expert model

took more time to train. However, it finally got an accuracy of 96.6%, which is higher than the single model. This experiment shows that MoE can be used to get higher results in a fine-grained classification task.

4.2.2 Experiment on Different Structures of MoE Model

Different expert structures would influence the model performance a lot. Normally, a condense layer is added after convolution operations to map the feature matrix to a hidden vector. The weighted summation part will add all hidden vectors together and feed them into towers. However, the problem here is that some features may lose before the weighted summation due to the compression from the condense layer in the expert. Applying weighted summation on outputs of convolution layers directly and doing feature compression in the tower may improve performance. Therefore, in this part, we performed experiments comparing two architectures mentioned above under direct learning and transfer learning respectively.

Our first experiment focused on the accuracy of these two architectures under **direct learning**. For both architectures, we trained them with the first 190 classes of our dataset and checked the highest testing accuracy that it can achieve. The testing accuracy along the training process of these two architectures are shown in Figure 6.

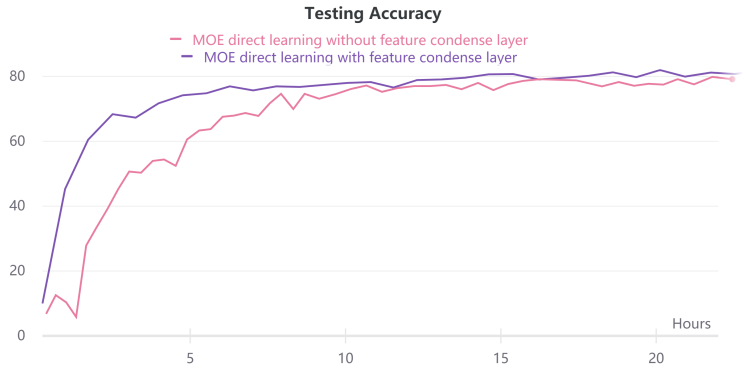

Figure 6: Accuracy of direct learning of MoE with and without feature condense layer

According to Figure 6, MoE model with feature condense layer got an accuracy of 82%, while model without feature condense layer got 80%. In addition, model with feature condense layer converged faster than the other one. This result shows that, In direct learning, feature condense layer is useful for the model to converge faster and get a higher accuracy.

In our second experiment, we tried two architectures under **transfer learning**. While performing transfer learning, in general, we will freeze the parameters in the body of the model and leave the parameters in the last several layers to update. In our implementation, we froze the parameters of all the experts and the feature condense layer and only updated the parameters in the gating network and

tower. We used the two well-trained models in the previous direct learning experiment and partially froze them. Then we performed transfer learning with the models to train with the last 10 classes of the dataset. The comparison of these two models' performance on transfer learning are shown in Figure 7.

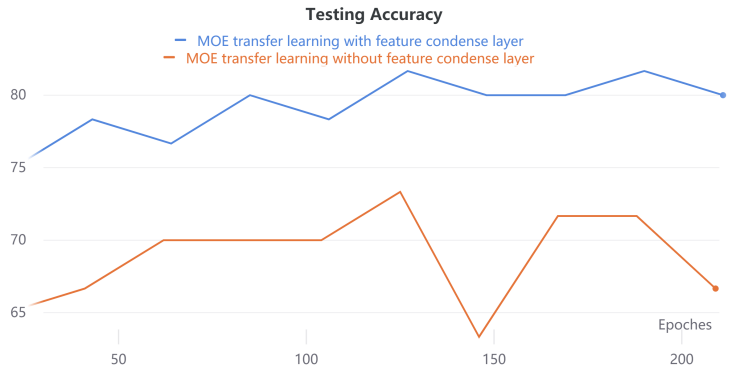

Figure 7: Accuracy of transfer learning of MoE with and without feature condense layer

According to Figure 7, MoE with feature condense layer got much higher accuracy in transfer learning than MoE without feature condense layer. In addition, the number of parameters for MoE without feature condense layer is much more than the other one. We think that the huge amount of parameters in MoE without the feature condense layer make it easy to get overfit, especially in a dataset with small amounts of data, which results in the low testing accuracy. Because model with feature condense layer works better in both direct learning and transfer learning, we will use MoE with the feature condense layer by default in the later experiments.

4.3 Multi-Task Mixture-of-Experts Model

In the following experiment, we will investigate how multi-task learning strategy and the number of experts influence the model performance. In order to leverage Multi-Task Learning advantages, we added a regression task to predict bounding box coordinate of the object as Task A. And the classification task we focus on is Task B. As shown in Figure 8, we choose Efficient-Net as the expert network, a simple convolutional neural network as the gate. In terms of training, we used one single optimizer to reduce the summation of two task losses. We choose the complete-IoU as the loss function for the bounding box regression task A, which considers overlap area, central point distance, and aspect ratio simultaneously and can lead to better performance. The loss function used in classification task B is a Cross-Entropy with label smoothing.

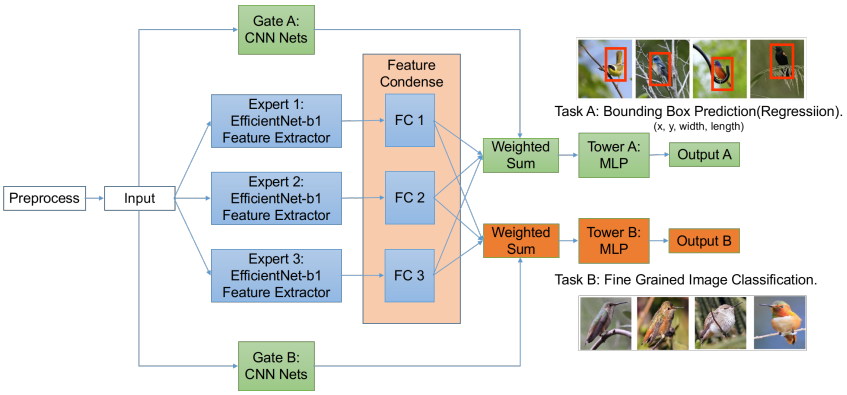

Figure 8: Our MMoE model structure

4.3.1 Comparison of Single-task MoE and Multi-task MoE

In order to test whether MTL is able to give us a better and more robust result, we design the experiment as shown in Table 1. Our objective is to compare classification accuracy between single-task (Task B) model and two-task (Task A&B) model. The empirical evaluation results in Table 1 and Figure 9 have shown that MTL improved the overall classification accuracy by 1.4%, from which we can conclude that introducing additional discriminative region localization tasks indeed facilitate fine-grained classification task.

Table 1: Settings and results of MoE models with different number of towers

	Batch Size	Learning Rate	Classes Num	Epoch	Experts	Best Classification Accuracy
One Tower MoE model	16	1e-4	200	100	3 eff-b1	0.7758
MTL-MoE (**Two Towers**) model	16	1e-4	200	100	3 eff-b1	**0.7893**

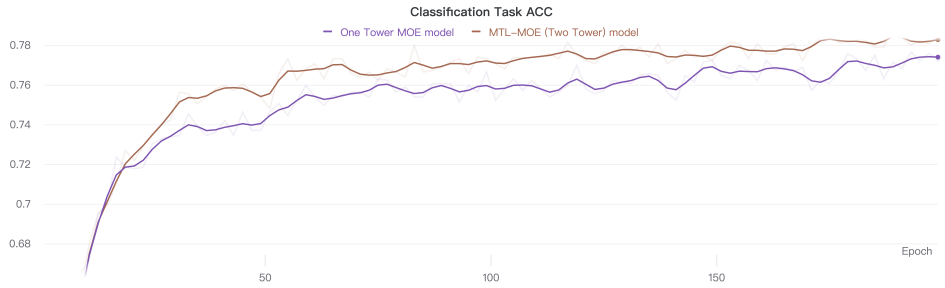

Figure 9: Classification accuracy of MoE models with different number of towers

4.3.2 Comparison of Different Number of Experts

And then we want to research on the influence of the expert number on model performance and learning efficiency. Here we test 1, 3, 6 experts respectively. The result of the experiment illustrated in Table 2 and Figure 10&11 reflects that the model with 3 experts reaches the highest accuracy both in the regression task (94.43%) and the classification task (80.44%). Both the 3-expert model and 6-expert model achieve better performance than single-expert model, which means that more experts are able to bring higher accuracy. However, there should be a trade-off between more representation ability and training difficulty. In another perspective, the model with more experts can converge faster than the model with fewer experts, i.e. the model with more experts can reach the same accuracy using fewer epochs.

Table 2: Settings and results of MTL-MoE models with different number of experts

	Batch Size	Learning Rate	Classes Num	Epoch	Experts	Best Regression Accuracy(Task A)	Best Classification Accuracy(Task B)
One Expert MTL-MoE model	16	1e-4	200	100	**1 eff-b1**	0.9337	0.7918
Three Experts MTL-MoE model	16	1e-4	200	100	**3 eff-b1**	**0.9424**	**0.8044**
Six Experts MTL-MoE model	8	1e-4	200	100	**6 eff-b1**	0.9391	0.8010

5 Conclusion

In this project, we have made efforts on using the mixture-of-experts model to solve fine-grained image classification task. The training strategy of transfer learning and the model design of multi-task learning have been exploited to improve our model performance. Empirical results of our experiments reflected that muti-task learning strategy and mixture-of-experts model structure have made significant contributions to faster training and more robust training. Over the dataset CUB 200-2011, our best result (80.44%) is better than most of the models published before 2018.

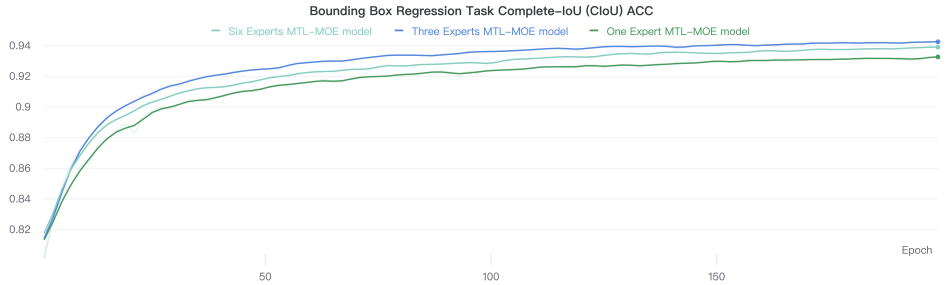

Figure 10: Regression accuracy of MTL-MoE models with different number of experts

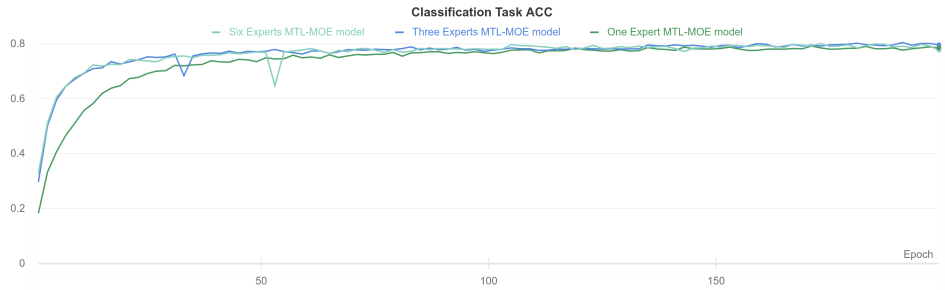

Figure 11: Classification accuracy of MTL-MoE models with different number of experts

In future work, we are interested in designing more specific preprocessing methods for fine-grained datasets and are prepared to combine the transfer learning and the multi-task learning to further improve the training efficiency and model accuracy. We will also research on how multiple tasks influence each other in multi-task learning in the next step.

References

[1] Thomas Berg and Peter N Belhumeur. Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 955–962, 2013.

[2] Yuning Chai, Victor Lempitsky, and Andrew Zisserman. Symbiotic segmentation and part localization for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 321–328, 2013.

[3] Ryan Farrell, Om Oza, Ning Zhang, Vlad I Morariu, Trevor Darrell, and Larry S Davis. Birdlets: Subordinate categorization using volumetric primitives and pose-normalized appearance. In *2011 International Conference on Computer Vision*, pages 161–168. IEEE, 2011.

[4] Efstratios Gavves, Basura Fernando, Cees GM Snoek, Arnold WM Smeulders, and Tinne Tuytelaars. Fine-grained categorization by alignments. In *Proceedings of the IEEE international conference on computer vision*, pages 1713–1720, 2013.

[5] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. Part-based r-cnns for fine-grained category detection. In *European conference on computer vision*, pages 834–849. Springer, 2014.

[6] ZongYuan Ge, Christopher McCool, Conrad Sanderson, and Peter Corke. Subset feature learning for fine-grained category classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 46–52, 2015.

[7] ZongYuan Ge, Alex Bewley, Christopher McCool, Peter Corke, Ben Upcroft, and Conrad Sanderson. Fine-grained classification via mixture of deep convolutional neural networks. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–6. IEEE, 2016.

[8] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

[9] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H Chi. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1930–1939, 2018.

[10] Yue Chen, Yalong Bai, Wei Zhang, and Tao Mei. Destruction and construction learning for fine-grained image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5157–5166, 2019.

[11] Zhe Xu, Shaoli Huang, Ya Zhang, and Dacheng Tao. Augmenting strong supervision using web data for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision*, pages 2524–2532, 2015.

[12] Yaming Wang, Vlad I Morariu, and Larry S Davis. Learning a discriminative filter bank within a cnn for fine-grained recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4148–4157, 2018.

[13] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 1449–1457, 2015.

[14] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 317–326, 2016.

[15] Shu Kong and Charless Fowlkes. Low-rank bilinear pooling for fine-grained classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 365–374, 2017.

[16] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 842–850, 2015.

[17] Xiaopeng Zhang, Hongkai Xiong, Wengang Zhou, Weiyao Lin, and Qi Tian. Picking deep filter responses for fine-grained image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1134–1142, 2016.

[18] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. *arXiv preprint arXiv:1506.02025*, 2015.

[19] Seniha Esen Yuksel, Joseph N Wilson, and Paul D Gader. Twenty years of mixture of experts. *IEEE transactions on neural networks and learning systems*, 23(8):1177–1193, 2012.

[20] Saeed Masoudnia and Reza Ebrahimpour. Mixture of experts: a literature survey. *Artificial Intelligence Review*, 42(2):275–293, 2014.

[21] David J Miller and Hasan S Uyar. A mixture of experts classifier with learning based on both labelled and unlabelled data. In *Advances in neural information processing systems*, pages 571–577, 1997.

[22] Markus Enzweiler and Dariu M Gavrila. A multilevel mixture-of-experts framework for pedestrian classification. *IEEE Transactions on Image Processing*, 20(10):2967–2979, 2011.

[23] Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017.

[24] Lianbo Zhang, Shaoli Huang, Wei Liu, and Dacheng Tao. Learning a mixture of granularity-specific experts for fine-grained categorization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8331–8340, 2019.

[25] Thomas G Dietterich. Machine-learning research. *AI magazine*, 18(4):97–97, 1997.

[26] Jonathan Baxter. A model of inductive bias learning. *Journal of artificial intelligence research*, 12:149–198, 2000.

[27] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

[28] Kim-Han Thung and Chong-Yaw Wee. A brief review on multi-task learning. *Multimedia Tools and Applications*, 77(22):29705–29725, 2018.

[29] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4438–4446, 2017.