Auxiliary Structure for Convolutional Neural Network Training

Cheng-Yeh Chen* and Chen-Kuo Chiang†
* National Chung Cheng University, Taiwan
E-mail: ccy103m@cs.ccu.edu.tw
† National Chung Cheng University, Taiwan
E-mail: ckchiang@cs.ccu.edu.tw Tel/Fax: +886-5-2729111

***Abstract*—Despite recent breakthroughs in the applications of deep neural networks, one challenge remains for deep model training when only limited training data available. New structure may be different from the previous models in different tasks. In addition, there may be insufficient labeled or unlabeled data to train or adapt a deep architecture to new tasks. In view of this, we propose the auxiliary structure for deep model learning with insufficient data via additional alignment layers to migrate the weight of the auxiliary model to the new model. The experimental results demonstrate that the auxiliary structure reduces the overfitting problem and can improve the accuracy using only a few training samples.**

I. INTRODUCTION

The recent success of deep learning depends on the ability to apply gradient-based optimization to high-capacity models. It has achieved state-of-the-art performance in various research fields, such as image classification [1], video classification [2], and speech recognition [3]. These models perform extremely well in domains with large amounts of training data, e.g., (Deng et al. [4], 2009). They have recently outperformed all known methods on a large scale recognition challenge [5].

However, to obtain a large dataset is not easy especially on specific task. This is the critical challenge for using deep learning. With limited training data, the representational capacity of deep architecture dramatically overfits the training data.

In this paper, we proposed an auxiliary structure to solve the task of deep model learning by using transfer learning technique. We take a pre-trained model, which was trained on a large dataset as our auxiliary model to train a new model. Between two models, alignment layer is exploited to help learn the new model without initial weight.

The contributions of our paper are that we propose the solution for deep model training. It can use any pre-trained networks to train the new deep model given the insufficient training samples or lacks of the initial weight. Each path in dual-path deep neural network is flexible, we can choose the different deep models according to different tasks requirement.

II. RELATED WORK

In recent years, convolutional neural network (CNN) is a classical feed-forward network in deep learning, and widely used in computer vision [10, 11, 17, 18]. Deep learning is part of a broader family of machine learning methods based on learning data representations, as opposed to task specific algorithms. In image representation learning approaches, traditional methods use hand-crafted image features. However, these features may not be applied for user-centric tasks, such as image recommendations. Therefore, Lei et al. [8] designed a dual-net deep network to learn user-centric image representations, and proposed a comparative deep learning (CDL) method to solve image recommendations problem. Dual-source deep neural network (DS-CNN) [16] that shares common features and retains flexibility through a dual path architecture to learning new features. DS-CNN differs from traditional deep neural networks. It contains two convolutional paths. Each path has its own model weight. Two outputs of dual paths can be integrated to obtain more discriminative results. We can also analyze and compare their difference.

Knowledge distillation (KD) [20, 21] transfers knowledge from a large highly regularized teacher network into a smaller student network. It provides soft-target information by using an arithmetic or geometric mean of individual predictive distributions computed by the teacher network, so the student network can be trained on much less data than the teacher model. Generalized distillation (GD) extends KD methods by training a teacher network with separate clean and noisy training sets [22, 23, 24]. A student network can be guided by the soft-labels from a teacher network where soft labels are derived using enhanced features generated by a beamformer then processed through a network trained with conventional multi-style training.

In this paper, we hope to construct our auxiliary by DS-CNN and domain adaptation technique. Domain adaptation is one of transfer learning issue. Transfer learning is an important research topic of deep model learning problem. Transfer learning can migrate models that are suitable for large data to small data for model migration [6][7]. In cross domain adaptation problem, Chen et al. [9] proposed a novel double-path deep domain adaptation network (DDAN) to model the data from the two domains jointly. The idea of this architecture is defining the path of shop domain images as source domain because the posture and clothing distribution are consistent. Another path of street domain images as the target domain is used for testing. Then, they proposed an alignment layer between these two paths to force the parameters of two paths

to be similar. It means that the approach can effectively make the features of two different domains to be similar. Accordingly, we wish to construct our model by alignment layer and DS-CNN technique.

III. Auxiliary Structure Learning For Deep Model Training

In this section, we introduce and describe our model. Motivated by background knowledge, we propose a new deep model learning method, Auxiliary Structure Learning (ASL), as shown in Fig. 1. ASL method is used in asymmetric dual-path deep neural network.

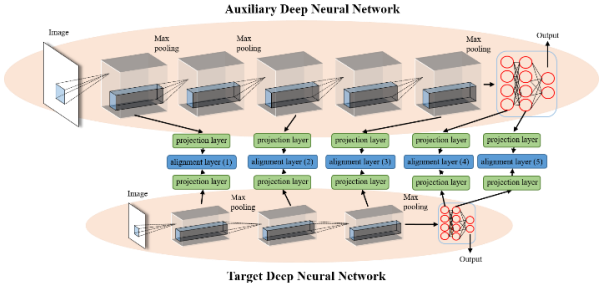

Fig. 1: Dual-path network architecture for ASL.

3.1. Model Architecture for Auxiliary Structure Learning

In the top path, we take a large generative deep learning model (e.g., AlexNet [10], GoogleNet [17]) as auxiliary architecture. Due to these well-known deep learning models use huge amount of data for model training, they have better model weight. In the bottom path, we take a small deep learning model as a target architecture without pre-trained weight. Each path in dual-path deep neural network is flexible, we can choose the different deep models according to different tasks requirement. In this paper, we focus on human action identification task.

We use auxiliary model to train target model. The auxiliary model provides the pre-trained weight, and transfer the model weight to target model. Auxiliary model can help target model convergence easier and improve accuracy. In training stage, auxiliary model and target model need to be trained simultaneously.

Before we link auxiliary model layer and target model layer to alignment layer, we need to add a projection layer. Projection layer projects the input feature representation form auxiliary model layer and target model layer to the same dimension. Between the two models, we use the alignment layer to build dependencies. Alignment layer is actually a cost function in following form:

$$L_k(a,t) = \lambda \times \left\| X_a^{(i)} - X_t^{(j)} \right\| \quad (1)$$

where a and t individually represent auxiliary model and target model. $X_a^{(i)}$ and $X_t^{(j)}$ are the representations from the connection layer (e.g., convolutional layer or fully-connected layer). λ is a constant value to adjust layer importance. i and j represent the connection between alignment layer and layer in dual-path. k is the number of alignment layers.

3.2. Considerations for Auxiliary Structure Learning

In order to make the target architecture converge easier and improve accuracy, we consider three main factors in Auxiliary Structure Learning (ASL) method: (1) The dimensions of projection layer for alignment layer. (2) The number of alignment layer. (3) Weight fixing in auxiliary structure.

3.2.1. The dimensions of projection layer for alignment layer

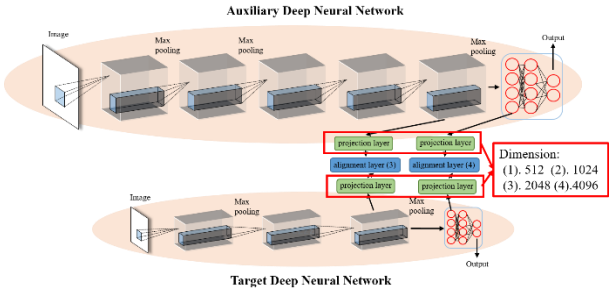

Fig. 2: Projection layer dimension selection.

of alignment layer, we choose Alignment2-3-4. (3) We fix the weight of convolutional layers weight in auxiliary structure. In this design consideration, we consider the dimensions of projection layer for alignment layer show as Fig. 2. We use four settings for projection layer: 512, 1024, 2048 and 4096.

In Table 1, experimental results indicate that when the projection layer dimension is 2048, we can obtain the highest accuracy 40.8 % in target model. We select 2048 for our projection layer dimension.

Table 1: The accuracy for projection layer dimension.

Projection dimension	*Accuracy (Auxiliary)*	*Accuracy (Target)*
512	51.0	34.7
1024	37.0	38.2
2048	32.8	**40.8**
4096	28.6	39.1

3.2.2. The number of alignment layer

In this design consideration, we consider the number of alignment layers. We use five settings for alignment layers: (1) Alignment1-2-3-4-5. (2) Alignment2-3-4-5. (3) Alignment2-3-4. (4) Alignment3-4. (5) Alignment3. Details as shown in Fig. 3.

In Table 2, the experimental results indicate that when alignment layer setting is alignment2-3-4, we can obtain the highest accuracy 42.7 % in target model.

Table 2: The accuracy for the number of alignment layer.

	Accuracy (Auxiliary)	*Accuracy (Target)*
Alignment 1-2-3-4-5	35.3	40.1
Alignment 2-3-4-5	40.7	38.5
Alignment 2-3-4	36.8	**42.7**
Alignment 3-4	37.0	38.2
Alignment 3	41.5	39.5

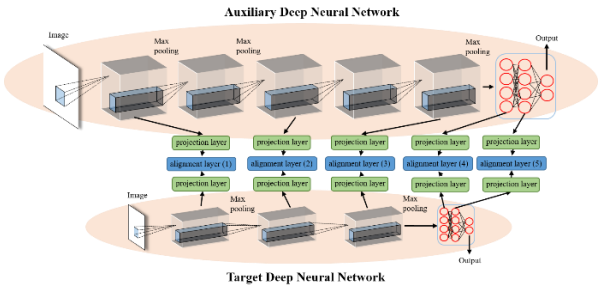

Fig. 3: The number of alignment layer.

3.2.3. Weight fixing in auxiliary structure

After doing the above experiments, most of experimental results demonstrate ASL method can be used to improve the accuracy for target model. At the same time, we found an unexpected situation: the accuracy for auxiliary architecture is decrease. We think the weight of auxiliary architecture is affected by the weight of target model in training stage. By fixing a part of parameters of auxiliary model, we hope it can reduce the problem for auxiliary model accuracy decrease.

In this design consideration, we consider three situations to fix the weight in auxiliary architecture: (1) Weight fixing in convolutional layers and fully-connected layers (Both Fixed). (2) Weight fixing in convolutional layers (Conv Fixed). (3) Without weight fixing in auxiliary architecture (Unfixed).

In Table 3, experimental results indicate that when fixing the convolutional layers weights in auxiliary architecture, we can obtain the highest accuracy for auxiliary architecture is 40.71%, and we also obtain the highest accuracy for target model is 41.13%.

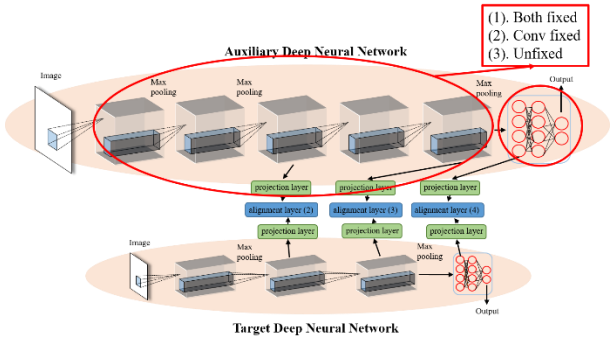

Fig. 4: Weight fixing in auxiliary structure.

Table 3: The accuracy for weight fixing.

	Accuracy (Auxiliary)	*Accuracy (Target)*
Both Fixed	20.83	39.67
Conv Fixed	**40.71**	**41.13**
Unfixed	38.42	39.17

3.3 Loss function for auxiliary structure learning

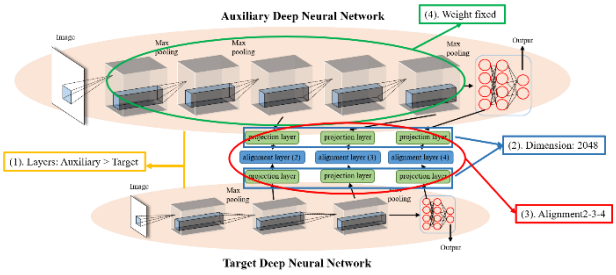

Fig. 5: Final deep architecture selection for ASL.

According the above experiments, we summarize our Considerations for ASL method: (1) The dimensions of projection layer for alignment layer is 2048. (2) In the number

Our final model architecture is shown in Fig. 10. In this model, the total loss function definition is in following form:

$$L_{total} = \sum_{k=1}^{n} L_k(a,t) + \sum_{p=1}^{m} \tilde{L}_p(y,z) \qquad (2)$$

$L_k(a,t)$ is loss term for alignment layer. There are three alignment layers in our model, so $n = 3$. Other details for alignment loss is in Section 3.2. $\tilde{L}_k(y,z)$ is softmax loss, and it define in following form:

$$\tilde{L}_p(y,z) = \log\left(\sum_{j=1}^{m} e^{z_j}\right) - z_y \qquad (3)$$

z_j is the j-th linear prediction result, y is the label for input data. p is the number of deep models, in our dual-path deep model, $m = 2$.

IV. Experimental Results

In this section, we conduct some experiments on our collected dataset to verify the performance of our proposed method. We focus on human action identification task.

4.1. Dataset

We evaluate the performance of our models on three datasets, People Playing Musical Instrument (PPMI) dataset, Willow dataset and Uiuc-sports dataset.

First, we use People Playing Musical Instrument (PPMI) dataset [12] for our experiments. It contents 4800 Normalized images. The PPMI dataset contains 24 human interaction categories, with 12 kinds of musical instruments, and each instrument contains 2 different interactions. The images number of each class is 200. In each class, we take 100 samples for training and 100 samples for testing.

The second dataset we use is Willow dataset [13]. It contains 991 still images and 7 kinds of action classes. We take 430 images for training and 481 images for testing.

Last, we use Uiuc-sports dataset [14]. It contents 1579 images and 8 kinds of action classes. In each class, we take half samples for training and the rest for testing.

4.2 Experiment Settings

We implement our model using the Caffe software package [15]. All neural network models use the same parameter settings with learning rate 0.0002, momentum 0.9, and weight decay 0.0002. All input image size in our auxiliary model is resized to 227×227, and in our target model is resized into 57×57.

In the following experiments, we use the model architecture as shown in Fig. 1. We take AlexNet as auxiliary model with the pre-trained weight to train target model without pre-trained weight. Between two models, we use three alignment layers to make the feature representations similar and migrate the weight of the auxiliary model to target model. We train two models simultaneously.

4.3 Experiment on PPMI dataset

In this experiment, we compared our proposed method with other model training methods on PPMI dataset. First, we directly train our small target model without any pre-trained weight. Second, we train auxiliary model, AlexNet. Then, we use ASL method to train small target model. Sample images of this dataset and Classification accuracy showed in Fig. 6 and Table 4, respectively. We can see by adopting ASL, the accuracy of target model can be improved by nearly 15% compared to the target model trained stand along.

Fig. 6: Final deep architecture selection for ASL.

4.4 Experiment on Willow dataset

In this experiment, we compared our proposed method with other model training methods on Willow dataset. Sample images and classification accuracy showed in Fig.7 and Table 5, respectively. The classification accuracy can be 5% higher.

Table 4: Classification accuracy on PPMI dataset.

	Accuracy (Auxiliary)	*Accuracy (Target)*
Target	56.4	-
AlexNet	-	28.1
Target (ASL)	36.8	**42.7**

Fig. 7: Final deep architecture selection for ASL.

Table 5: Classification accuracy on Willow dataset.

	Accuracy (Auxiliary)	*Accuracy (Target)*
Target	50.8	-
AlexNet	-	39.8
Target (ASL)	46.0	**44.8**

4.5 Experiment on Uiuc-sports dataset

In this experiment, we compared our proposed method with other model training methods on Uiuc-sports dataset. Classification accuracy showed in Table 6. The improvement can be also achieved by 5% higher by using the proposed ASL.

Table 6: Classification accuracy on Uiuc-sports dataset.

	Accuracy (Auxiliary)	*Accuracy (Target)*
Target	86.6	-
AlexNet	-	74.8
Target (ASL)	84.5	**81.6**

V. CONCLUSION

In this paper, we propose a method for deep model training, use auxiliary structure and alignment layer to train deep convolutional neural network. Our experimental results demonstrate that our method can improve the performance in target model. With the replacement of different auxiliary architecture, we can apply our method in different tasks, such like detection and retrieval. In the future, we will solve the problem for auxiliary model layers less than personalization model layers (Small to Big). We can replace the larger auxiliary model (e.g., GoogleNet, VGGNet [18]) with AlexNet or dividing the target model into several sub-target model to make every sub-model layers less than auxiliary model layers.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1026-1034, Dec 2015.

[2] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Largescale video classification with convolutional neural networks," in 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1725-1732, June 2014.

[3] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," IEEE Signal Processing Magazine, vol. 29, pp. 82-97, Nov 2012.

[4] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "Imagenet: A largescale hierarchical image database," in 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 248-255, June 2009.

[5] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," International Journal of Computer Vision (IJCV), vol. 115, no. 3, pp. 211-252, 2015.

[6] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," CoRR, vol. abs/1411.1792, 2014

[7] J. Donahue, Y. Jia, O. Vinyals, J. Ho_man, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in International Conference in Machine Learning (ICML), 2014.

[8] C. Lei, D. Liu, W. Li, Z. J. Zha, and H. Li, "Comparative deep learning of hybrid representations for image recommendations," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2545-2553, June 2016.

[9] Q. Chen, J. Huang, R. Feris, L. M. Brown, J. Dong, and S. Yan, "Deep domain adaptation for describing people based on fine-grained clothing attributes," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5315-5324, June 2015.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems 25, pp. 1097-1105, Curran Associates, Inc., 2012.

[11] S. Liu, X. Liang, L. Liu, X. Shen, J. Yang, C. Xu, L. Lin, X. Cao, and S. Yan, "Matching-cnn meets knn: Quasi-parametric human parsing," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1419-1427, June 2015.

[12] B. Yao and L. Fei-Fei, "Grouplet: A structured image representation for recognizing human and object interactions," in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 9-16, June 2010.

[13] V. Delaitre, I. Laptev, and J. Sivic, "Recognizing human actions in still images: a study of bag-of-features and part-based representations," 2010. updated version, available at http://www.di.ens.fr/willow/research/stillactions/.

[14] L. J. Li and L. Fei-Fei, "What, where and who? classifying events by scene and object recognition," in 2007 IEEE 11th International Conference on Computer Vision, pp. 1-8, Oct 2007.

[15] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," CoRR, vol. abs/1408.5093, 2014.

[16] Xiaochuan Fan, Kang Zheng, Yuewei Lin, Song Wang, "Combining Local Appearance and Holistic View: Dual-Source Deep Neural Networks for Human Pose Estimation", in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2015.

[17] C. Szegedy, et al., "Going deeper with convolutions," in Proc.

[18] IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 1–9.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv:1409.1556, 2014.

[20] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," arXiv preprint arXiv:1503.02531, 2015.

[21] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio, "Fitnets: Hints for thin deep nets," arXiv preprint arXiv:1412.6550, 2014.

[22] David Lopez-Paz, L´eon Bottou, Bernhard Sch¨olkopf, and Vladimir Vapnik, "Unifying distillation and privileged information," arXiv preprint arXiv:1511.03643,

[23] Konstantin Markov and Tomoko Matsui, "Robust speech recognition using generalized distillation framework," Interspeech 2016, pp. 2364–2368, 2016.

[24] Shinji Watanabe, Takaaki Hori, Jonathan Le Roux, and John Hershey, "Student-teacher network learning with enhanced features," ICASSP'17, pp. 5275–5279, 2017.