Simultaneous Estimation of Facial Landmark and Attributes with Separation Multi-task Networks

Ryo Matsui, Takayoshi Yamashita and Hironobu Fujiyoshi
Chubu University, Kasugai, Aichi, Japan
ryo@mprg.cs.chubu.ac.jp, {takayoshi, fujiyoshi}@isc.chubu.ac.jp

Keywords: Deep Convolutional Neural Network, Multi-task Learning, Channel-wise Convolution, Facial Landmark Detection, Facial Attribute Estimation.

Abstract: Multi-task learning is a machine learning approach in which multiple tasks are solved simultaneously. This approach can improve learning efficiency and prediction accuracy for the task-specific models. Furthermore, it has been used successfully across various applications such as natural language processing and computer vision. Multi-task learning consists of shared layers and task-specific layers. The shared layers extract common low-level features for all tasks, the task-specific layers diverge from the shared layers and extract specific high-level features for each task. Hence, conventional multi-task learning architecture cannot extract the low-level task-specific feature. In this work, we propose Separation Multi-task Networks, a novel multi-task learning architecture that extracts shared features and task-specific features in various layers. Our proposed method extracts low- to high-level task-specific features by feeding task-specific layers in parallel to each shared layer. Moreover, we employ channel-wise convolution when concatenating feature maps of shared layers and task-specific layers. This convolution allows concatenation even if layers have a different number of channels of feature maps. In experiments on CelebA dataset, our proposed method outperformed conventional methods at facial landmark detection and facial attribute estimation.

1 INTRODUCTION

From a person's face, many attributes can be captured many attributes including age, gender, facial expression and the positions of facial landmarks. These facial attributes will benefit many applications such as face recognition, face swapping, and virtual makeup. facial attribute estimation is thus an attractive research field in computer vision(Zhao et al., 2018).

Since the Deep Convolutional Neural Network (DCNN) achieves high recognition accuracy in object recognition (Krizhevsky et al., 2012)(Simonyan and Zisserman, 2015)(He et al., 2016), it is used in many tasks such as object detection(Liu et al., 2016) and human pose estimation(Wei et al., 2016). Similarly, the DCNN can achieve high-accuracy facial landmark detection and facial attribute estimation (Lv et al., 2017)(Liu et al., 2015). However, training and inference time increases in proportion to the number of tasks due to a DCNN needing to be built for each single task. Therefore, facial image analysis will need vast cost because it requires information for many tasks such as facial landmark detection, age estimation, and gender recognition.

By using multi-task learning, multiple recognition tasks can be simultaneously trained and with a single DCNN(Caruana, 1998)(Zhang et al., 2014b). Thus, training and inference time can be greatly reduced. This method is efficient in facial image analysis consisting of multiple tasks. Multi-task learning may share feature representation among all tasks and optimize multiple tasks simultaneously. By training multiple tasks simultaneously, common features can be efficiently extracted in earlier layers. Zhang et al. (Zhang et al., 2014b) proposed a Tasks-Constrained Deep Convolutional Network that detects facial landmarks and estimates attributes simultaneously. Misra et al. (Misra et al., 2016) proposed Cross-stitch Networks that perform surface normal estimation and semantic segmentation. Feichtenhofer et al. (Feichtenhofer et al., 2017) proposed a method that detects object and tracks object. Moreover, multi-task learning has been successfully in not only in the computer vision but also in various fields such as natural language processing(Liu et al., 2017). The methods based on multi-task learning consists of shared layers to extract shared features among all tasks and task-specific layers to extract task-specific features as shown in Fig-

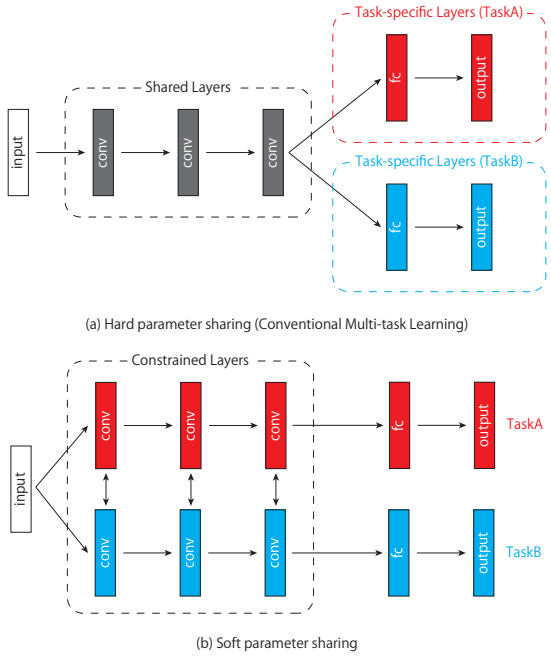

Figure 1: Network architectures of existing multi-task learning: (a) hard parameter sharing which conventional multi-task learning and (b) soft parameter sharing.

ure 1(a). The conventional multi-task learning can not exploit the low-level features of task-specific layers, because task-specific layers come after shared layers. Thus, only the high-level features of shared features and task-specific features are extracted from these layers.

In this work, we propose Separation Multi-task Networks, a novel multi-task learning architecture of DCNN that extracts shared features and each task-specific feature simultaneously. Unlike the architectures based on the conventional multi-task learning, Separation Multi-task Networks conduct parallel processing with shared layers that extract shared features among all tasks, and task-specific layers that extract task-specific features in each task. The feature maps of shared layers are concatenated to feature maps of each task-specific layers. Thus, Separation Multi-task Networks are able to extract shared features and task specific features in each layer. To implement a network, our proposed method introduces a two-stage training. The network with shared layers is trained first, and all task-specific layers are appended to train task specific features. When shared features are concatenated to task-specific features, the number of channels of the feature maps increases. Therefore, 1×1 convolution (Channel-wise Convolution) is performed on the concatenated the feature maps. This allows changes to the number of channels of the feature maps that ate input for task-specific layers. Also, it allows fine-tuning by general purpose networks such as VGGNet(Simonyan and Zisserman, 2015) and ResNet(He et al., 2016), which are trained on an Imagenet dataset. In the experiments, we evaluate the effectiveness of Separation Multi-task Networks by performing facial landmark detection and facial attribute recognition on the CelebA dataset. We will also evaluate the effectiveness of introducing channel-wise convolution.

2 RELATED WORKS

2.1 Multi-task Learning

Multi-task learning is a machine learning approach in which multiple learning tasks are optimized at the same time(Caruana, 1998). A single DCNN with multi-task learning can improve learning efficiency and prediction accuracy for the task-specific models. In the conventional DCNN, training and inference time increases in proportion to the number of tasks due to the need to build a single DCNN for each task. Meanwhile, multi-task learning can greatly reduce training and inference time by simultaneously training multiple tasks with a single DCNN. Multi-task learning has the advantage that it can extract efficient common features. There are two types of multi-task learning in DCNN: hard parameter sharing and soft parameter sharing(Ruder, 2017).

2.2 Hard parameter sharing

Hard parameter sharing is a network architecture that shares parameters from an input layer to intermediate layers with multiple tasks. The network architecture is as shown in Figure 1(a). Shared layers extract features common to all tasks. Task-specific layers are stacked after shared layers and extract all task-specific features. Then, task-specific layers of each task output the recognition results. Shared layers can reduce over-fitting more than a single DCNN because these layers train efficient feature representation to all tasks. Hard parameter sharing is a general method of multi-task learning and has many applications (Zhang et al., 2014b)(Misra et al., 2016)(Dai et al., 2016).

Cross-stitch Networks perform surface normal estimation and semantic segmentation simultaneously(Misra et al., 2016). In this method, DCNN is structured for each training task, and shared features are trained by Cross-stitch units. Cross-stitch units integrate feature maps extracted by each channel of each task and generate shared feature maps that are fed into each DCNN. The accuracy in surface normal

estimation and semantic segmentation is higher than when a single DCNN is used. Specifically, the accuracy is significantly improved for the categories that have few training data.

Multi-task Network Cascades arrange several tasks in a cascade structure to perform instance segmentation(Dai et al., 2016). Multi-task Network Cascades sequentially process object detection, mask estimation, and category estimation after shared features extraction. The output of each task is fed to the next task with shared features. The method achieves highly accuracy in instance segmentation tasks.

Since multi-task learning can extract task-independent features, it is also used to improve the accuracy of specific tasks. A Tasks-Constrained Deep Convolutional Network(TCDCN) trains facial landmark detection as the main task and the sub-tasks such as face orientation detection, gender estimation, smiling recognition (Zhang et al., 2014b), simultaneously. These sub-tasks help to improve the accuracy of the main task. This means that sub-tasks categorize the facial appearance, and the main task considers there categorizations to obtain efficient features. TCDCN introducing task-wise early stopping, which stops the training of a sub-task before it finishes, reduces the negative effects over-fitting of sub-tasks on the main task.

2.3 Soft parameter sharing

Soft parameter sharing is a network architecture consisting of constrained layers that make the distance between the parameters of each task uniform. Unlike hard parameter sharing, soft parameter sharing has DCNNs for each task. The network architecture of soft parameter sharing is shown in Figure 1(b). To make the distance between the parameters of each task uniform, soft parameter sharing utilizes regularization by the L2 norm(Duong et al., 2015) and trace norm(Yang and Hospedales, 2016). Because the architecture of soft parameter sharing constructs a DCNN for a task, training and inference time increases in proportion to the number of tasks.

3 PROPOSED SEPARATION MULTI-TASK NETWORKS

In this section, we describe the network architecture of Separation Multi-task Networks in 3.1, the two-stage training in 3.2, and channel-wise convolution in 3.3.

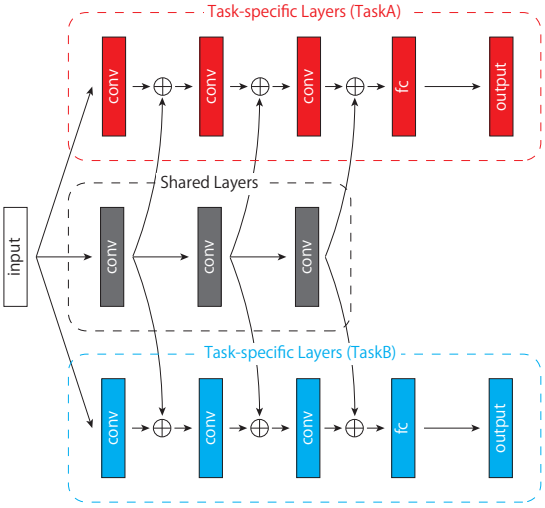

Figure 2: Our proposed Separation Multi-task Networks.

3.1 Network architecture

As shown in Figure 2, in our proposed Separation Multi-task Networks, shared layers that extract shared features among all tasks and task-specific layers that extract task-specific features in each task are constructed in parallel. The feature maps of shared layers are concatenated to the feature maps of each task-specific layer. The concatenated feature maps are fed into each task-specific layer. This allows training of both low- and high-level features in shared layers and corresponding task-specific layers.

3.2 Training procedure

Separation Multi-task Networks extract features common to all tasks and task-specific features in each task, simultaneously. First, the method needs to train only shared layers to extract common features, and then, it focuses on training task-specific features. Therefore, we use a two-stage training procedure.

Stage 1. First, multiple tasks are trained simultaneously using a conventional multi-task learning approach based on hard parameter sharing such as in Figure 1(a). The parameters of shared layers that extract common features among all tasks are trained in advance. These parameters are updated by using the training loss of each task. For example, training tasks are Task A and Task B when training losses are E_{TaskA} and E_{TaskB}, respectively. The training loss E_{all} of the whole network is defined in Equation (1). At this time, the training loss is obtained by the mean square error function when the training task is the regression. On the other hand, the training loss is obtained by softmax cross entropy loss function when the training

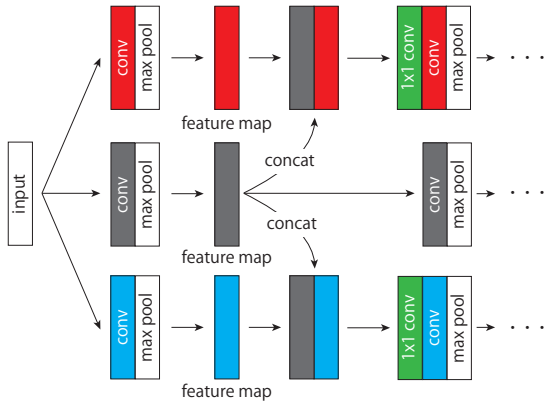

Figure 3: Channel-wise Convolution.

task is the recognition.

$$E_{all} = E_{TaskA} + E_{TaskB} \quad (1)$$

Stage 2. In second stage, as shown in Figure 2, the proposed method newly constructs task-specific layers in parallel with the shared layers. At this time, the parameters of shared layers obtained in stage 1 are fixed in the training in stage 2. The method only trains the parameters of each task-specific layer. As a result, only task specific features are extracted from task-specific layers. The parameters of the network in stage 2 are optimized by the training loss function of each task as in stage 1.

3.3 Channel-wise Convolution

The number of channels of the feature maps fed into each task-specific layer increases, because the proposed method concatenates the feature maps of shared layers and each task-specific layers. Thus, each task-specific layers can not fine-tuning a pretrained model such as VGGNet(Simonyan and Zisserman, 2015) and ResNet(He et al., 2016). We introduce 1×1 convolution (Channel-wise Convolution) after concatenating the feature maps as shown in Figure 3. This allows adjustment of the number of channels of the feature maps fed into each task-specific layers and fine-tuning of each task-specific layer.

4 EXPERIMENTS

We evaluate the effectiveness of our proposed method and compare it with related methods. For multiple tasks, we perform facial landmark detection and facial attribute estimation on the CelebA dataset. For the comparison method in facial landmark detection, multi-task learning based on hard parameter sharing is used (baseline). This network architecture is same as that in Stage 1. of Sec. 3.2. In facial attribute estimation, we compare our proposed Separation Multi-task Networks with FaceTracer(Kumar et al., 2008), PANDA-w(Zhang et al., 2014a), PANDA-l(Zhang et al., 2014a), and LNets+ANet(Liu et al., 2015), which is described by Liu et al. (Liu et al., 2015).

We implement three training models, 1) the two-stage training model, 2) the two-stage training model with channel-wise convolution, and 3)the two-stage training model based on fine-tuning from shared layers. Furthermore, we evaluate the variation of the network architectures with different numbers of task-specific layers.

Table 1: Details of the network architecture of Separation Multi-task Networks used in this experiment.

	TaskA	Shared	TaskB
Input	size $= 128 \times 128$, channels $= 3$		
Conv1-1	ksize $= 3 \times 3$, channels $= 64$, pad $= 1$		
Conv1-2			
max pooling	ksize $= 2 \times 2$		
Conv2-1	ksize $= 3 \times 3$, channels $= 128$, pad $= 1$		
Conv2-2			
max pooling	ksize $= 2 \times 2$		
Conv3-1	ksize $= 3 \times 3$, channels $= 256$, pad $= 1$		
Conv3-2			
Conv3-3			
max pooling	ksize $= 2 \times 2$		
Conv4-1	ksize $= 3 \times 3$, channels $= 512$, pad $= 1$		
Conv4-2			
Conv4-3			
max pooling	ksize $= 2 \times 2$		
Fc	2048		2048
Output	10		80

4.1 Experiments details

The CelebA dataset used in this experiment consists of about 200,000 facial images. Annotations are 5 facial landmarks (eyes, nose, and mouth corners) and 40 facial attributes such as "hat," "black hair," and "smiling." In training and evaluation, 162,770 samples are used as training data, and 19,962 samples as evaluation data.

In these expriments, the baseline and our proposed Separation Multi-task Networks use a network model that improved VGG16(Simonyan and Zisserman, 2015). Table 1 lists the details of the network architecture of our Separation Multi-task Networks used in these experiments, where Task A is facial landmark detection and Task B is facial attributes estimation. In addition, the activation function of the network uses ReLU. To optimize both the baseline and our proposed Separation Multi-task Networks, MomentumSGD was used. The learning rate is 0.0001

Table 2: Accuracy of facial landmark detection with baseline and our proposed Separation Multi-task Networks [%].

	Channel-wise Convolutioin	Fine-tuning	Left eye	Right eye	Nose	Left mouth corner	Right mouth corner	Average
Ours			**97.7**	**98.0**	**87.5**	**94.9**	**94.7**	**94.6**
	✓		96.9	97.0	82.2	93.1	93.0	92.4
	✓	✓	96.1	96.0	50.6	85.4	85.1	82.6
Baseline			96.3	96.5	54.0	92.0	91.9	86.1

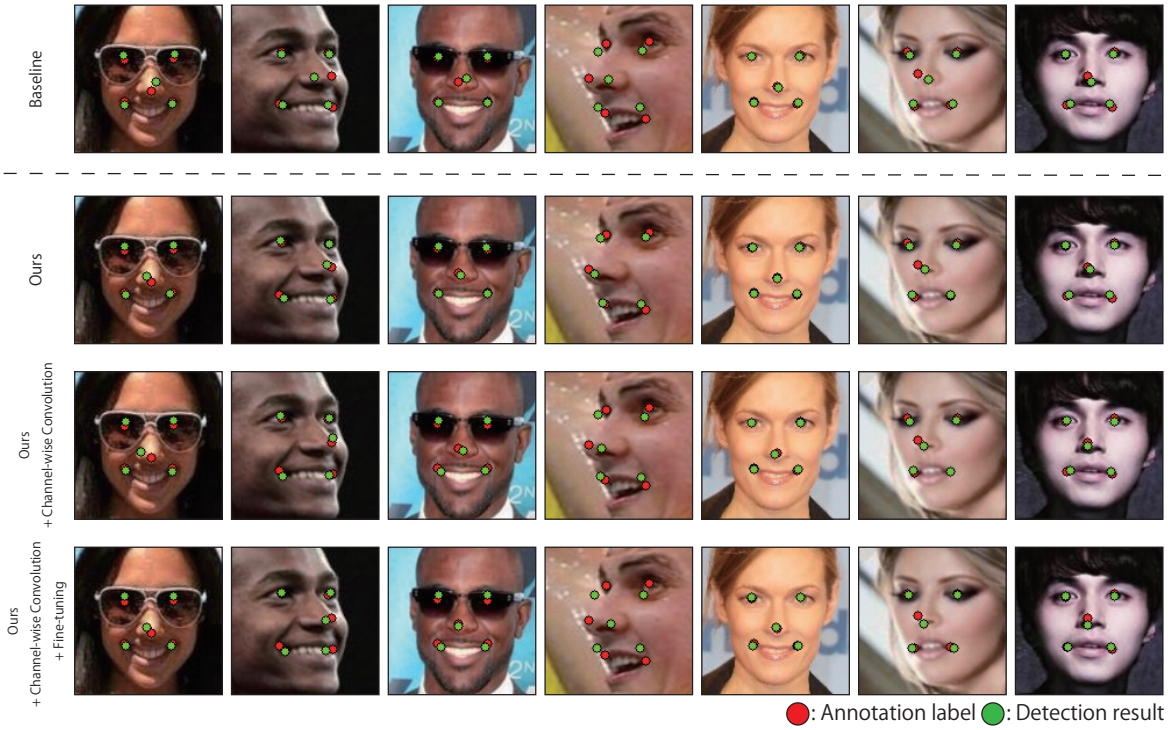

Figure 4: Example of facial landmark detection by baseline and our Separation Multi-task Networks.

and momentum is 0.9. The iterations of training is 100 epochs and mini-batch size is 32. Input image size is 128×128.

Evaluation metrics of Facial Landmark Detection. The evaluation metrics of facial landmark detection in this experiments determines that detection is successful when Equation (2) is satisfied in each facial landmark. In Equation (2), x_i and y_i are the annotation labels, x'_i and y'_i are the detection results, L is the distance between both eyes, and α is the threshold. In these experiments, threshold α is set to 0.1, and if the error between the annotation label and the detection result is within 10% of the distance between both eyes, the detection is successful.

$$\frac{\sqrt{(x_i - x'_i)^2 + (y_i - y'_i)^2}}{L} \leq \alpha \qquad (2)$$

4.2 Results for proposed Separation Multi-task Networks

Facial Landmark Detection. Table 2 shows accuracy of facial landmark detection for the baseline and our Separation Multi-task Networks. From Table 2, the average detection accuracy of our Separation Multi-task Networks is 94.6%, while baseline method is 86.1%. It is higher than the accuracy of baseline. Also, our Separation Multi-task Networks achieves high detection accuracy than the baseline for all facial landmarks. Specifically, our Separation Multi-task Networks accurately detected about 87.5% of nosed whereas the baseline is about 54.0%, a 33.5% difference. This shows the effectiveness in facial landmark detection of simultaneously extracting features shared among all tasks and task-specific features for each task. The two-stage training model introducing channel-wise convolution into our Separation Multi-task Networks outperformed the baseline but it was

Table 3: Accuracy of each method on 40 facial attributes [%].

	Channel-wise Convolution	Fine-tuning	5 o Clock Shadow	Arched Eyebrows	Attractive	Bags Under Eyes	Bald	Bangs	Big Lips	Big Nose	Black Hair	Blond Hair	Blurry	Brown Hair	Bushy Eyebrows	Chubby	Double Chin	Eyeglasses	Goatee	Gray Har	Heavy Makeup	High Cheekbones	Male
Ours			**93**	**81**	80	**83**	**98**	**95**	69	82	85	94	**95**	84	**91**	**95**	**96**	**99**	**97**	97	89	85	97
	✓		**93**	80	79	82	**98**	**95**	69	81	85	94	**95**	85	**91**	**95**	95	**99**	96	97	89	85	97
	✓	✓	**93**	**81**	80	**83**	**98**	**95**	**70**	**83**	85	94	**95**	**86**	**91**	**95**	**96**	**99**	**97**	**98**	**90**	85	97
Baseline			**93**	**81**	80	**83**	**98**	**95**	69	82	85	94	**95**	85	**91**	**95**	**96**	**99**	**97**	97	89	86	97
FaceTracer			85	76	78	76	89	88	64	74	70	80	81	60	80	86	88	98	93	90	85	84	91
PANDA-w			82	73	77	71	92	89	61	70	74	81	77	69	76	82	85	94	86	88	84	80	93
PANDA-l			88	78	**81**	79	96	92	67	75	85	93	86	77	86	86	88	98	93	94	**90**	86	97
LNets+ANet			91	79	**81**	79	**98**	**95**	68	78	**88**	**95**	84	80	90	91	92	**99**	95	97	**90**	**87**	**98**

	Channel-wise Convolution	Fine-tuning	Mouth Slightly Open	Mustache	Narrow Eyes	No Beard	Oval Face	Pale Skin	Pointy Nose	Receding Hairline	Rosy Cheeks	Sideburns	Smiling	Straight Hair	Wavy Hair	Wearing Earrings	Wearing Hat	Wearing Lipstick	Wearing Necklace	Wearing Necktie	Young		Average
Ours			92	**96**	**86**	**95**	71	**96**	74	**92**	**94**	**97**	91	79	78	87	98	**93**	84	**95**	86		89
	✓		92	**96**	**86**	**95**	70	**96**	74	91	**94**	**97**	91	79	78	87	98	92	84	**95**	85		89
	✓	✓	**93**	**96**	**86**	**95**	**72**	**96**	**75**	**92**	**94**	**97**	91	**80**	79	**88**	98	**93**	**85**	**95**	86		**90**
Baseline			**93**	**96**	**86**	**95**	71	**96**	74	**92**	**94**	**97**	91	**80**	79	**88**	98	**93**	**85**	**95**	86		89
FaceTracer			87	91	82	90	64	83	68	76	84	94	89	63	73	73	89	89	68	86	80		81
PANDA-w			82	83	79	87	62	84	65	82	81	90	89	67	76	72	91	88	67	88	77		79
PANDA-l			**93**	93	84	93	65	91	71	85	87	93	**92**	69	77	78	96	**93**	67	91	84		85
LNets+ANet			92	95	81	**95**	66	91	72	89	90	96	**92**	73	**80**	82	**99**	**93**	71	93	**87**		87

outperformed by the model that did not introduce it. Similarly, the model that fine-tuned each task-specific layers by introducing channel-wise convolution with shared layers has lower accuracy than the baseline.

Figure 4 shows an example of facial landmark detection results with the baseline and our Separation Multi-task Networks. Red points are annotation label and green points are detection results. Similarly to Figure 4, our Separation Multi-task Networks achieves higher detection accuracy than the baseline. Also, for profile face, it deviations happen at all facial landmarks when the baseline is used. However, our Separation Multi-task Networks mostly solve this problem.

Facial Attribute Estimation. Table 3 shows results for state-of-the-art method in addition to the baseline and our Separation Multi-task Networks on 40 facial attributes. From Table 3, our proposed method outperforms the comparison methods in average estimation accuracy. Also, our proposed method achieves the highest estimation accuracy in specific tasks such as "Bald," "Eyeglasses," and "Pale Skin." The model that introduces channel-wise convolution into our Separation Multi-task Networks had the same estimation accuracy as the model without it. However, the model that fine-tuned each task-specific layers by introducing channel-wise convolution with shared layers achieves the highest estimation accuracy with an average estimation accuracy about 90%. Similarly to facial landmark detection, our proposed method is also efficient at facial attribute estimation. Thus, it is considered that the features shared between tasks improves the accuracy for the problems such as facial attribute estimation.

4.3 Comparison of the network architecture

In our Separation Multi-task Networks, both shared layers and task-specific layers use the same network architecture, and images are simultaneously inputs into each networks. In these experiments, the feature maps of shared layers are fed into each task-specific layers instead of the input image as shown in Figure

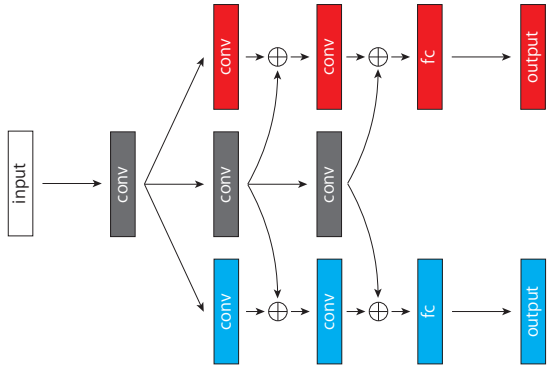

Figure 5: Example of the change of network architecture of Separation Multi-task Networks.

Table 4: The accuracy due to variation of the network architectures [%].

	Branch point		
	Conv1-2	Conv2-2	Conv3-3
Facial Landmark Detection (Baseline)	**92.9**	**95.5**	**95.2**
Facial Landmark Detection (Changed Ours)	80.4	79.5	78.8
Facial Attribute Estimation (Baseline)	**89.9**	**89.8**	**89.9**
Facial Attribute Estimation (Changed Ours)	88.8	88.8	88.8

5. The network uses the same parameters as in Sec. 4.1, where the start position input to each task-specific layers from shared layers is taken as the branch point, and the network model implements three models in which the branch point is "Conv1-2," "Conv2-2," and "Conv3-3" in Table 1. The comparison method is multi-task learning based on hard parameter sharing of the same branch point, in addition to the baseline and our original Separation Multi-task Networks.

The accuracy due to variation of the network architectures are shown in Table 4. From Table 4, our Separation Multi-task Networks shows the accuracy lower than the baseline in both tasks. Also, before changing the network architecture, our Separation Multi-task Networks has the highest estimation accuracy. From this, in multi-task learning, it is more effective to use the features shared between tasks more supplementally for each task like our proposed method.

5 DISCUSSION

We input a image to Separation Multi-task Networks and visualize these feature maps on Conv1-2 layer of shared layer and task-specific layer as shown in Figure 6. In shared layer, facial contour features are extracted. On the other hand, task-specific layer of facial landmark detection only extracts insufficient features from the feature of shared layer such as endpoint of landmark. Similarly, task-specific layer of facial attribute estimation extracts context features such as facial wrinkle. Thus, Separation Multi-task Networks shows it is able to extract common features for all tasks and task-specific features in each task, separately.

6 CONCLUSION

In this work, we proposed Separation Multi-task Networks, a novel multi-task learning method that simultaneously extracts features shared between tasks, and task-specific features in each task. Our proposed method was able to train and inference taking into account features shared between all tasks and task-specific in each task. Moreover, by introducing channel-wise convolution, our proposed method was able to adjust the number of channels of the feature maps input to each task-specific layers and fine-tune each task-specific layers. In experiments, our Separation Multi-task Networks performed facial landmark detection and facial attribute estimation on the CelebA dataset and outperformed the existing methods in both tasks. Then, in multi-task learning, we showed that it is effective to use features shared between tasks more supplementally for each task.

Future tasks include applying Separation Multi-task Networks to images other than facial image. In addition, our proposed method separates shared features shared between tasks and task-specific features in each task by two-stage training. Therefore, it is considered that Separation Multi-task Networks can be improved to train with end-to-end training while separating these two features. By improving the network model of our proposed method, it is considered the number of parameters can be reduced by changing the activation function to CReLU(Shang et al., 2016).

REFERENCES

Caruana, R. (1998). Multitask learning. In *Learning to learn*, volume 1, pages 95–133. Springer.

Dai, J., He, K., and Sun, J. (2016). Instance-aware semantic segmentation via multi-task network cascades. In *Conference on Computer Vision and Pattern Recognition*, pages 3150–3158.

Duong, L., Cohn, T., Bird, S., and Cook, P. (2015). Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International*

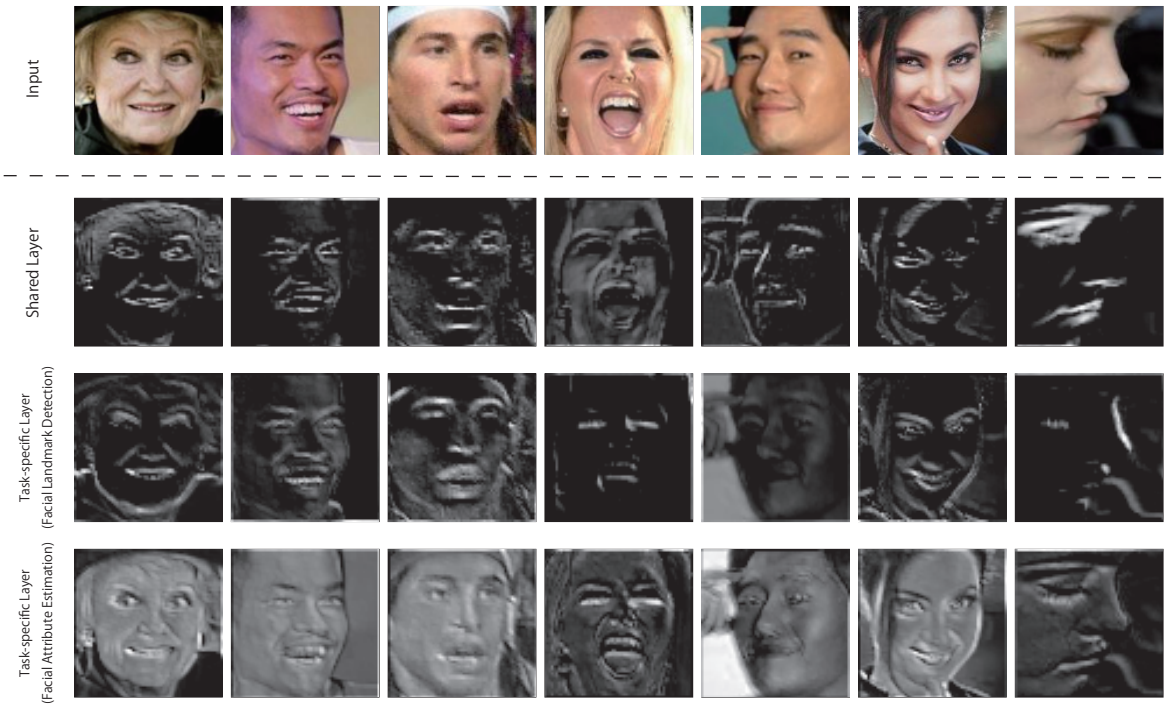

Figure 6: Visualization example of feature maps on Conv1-2 of our Separation Multi-task Networks.

Joint Conference on Natural Language Processing, volume 2, pages 845–850.

Feichtenhofer, C., Pinz, A., and Zisserman, A. (2017). Detect to track and track to detect. In *International Conference on Computer Vision*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition*, pages 770–778.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Neural Information Processing Systems*, pages 1097–1105.

Kumar, N., Belhumeur, P. N., and Nayar, S. K. (2008). Facetracer: A search engine for large collections of images with faces.

Liu, P., Qiu, X., and Huang, X. (2017). Adversarial multi-task learning for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1–10.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. (2016). Ssd: Single shot multibox detector. In *European Conference on Computer Vision*.

Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild. In *International Conference on Computer Vision*.

Lv, J.-J., Shao, X., Xing, J., Cheng, C., Zhou, X., et al. (2017). A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In *Computer Vision and Pattern Recognition*.

Misra, I., Shrivastava, A., Gupta, A., and Hebert, M. (2016). Cross-stitch networks for multi-task learning. In *Conference on Computer Vision and Pattern Recognition*.

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Shang, W., Sohn, K., Almeida, D., and Lee, H. (2016). Understanding and improving convolutional neural networks via concatenated rectified linear units. In *International Conference on Machine Learning*.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.

Wei, S.-E., Ramakrishna, V., Kanade, T., and Sheikh, Y. (2016). Convolutional pose machines. In *Conference on Computer Vision and Pattern Recognition*.

Yang, Y. and Hospedales, T. M. (2016). Trace norm regularised deep multi-task learning. *arXiv preprint arXiv:1606.04038*.

Zhang, N., Paluri, M., Ranzato, M., Darrell, T., and Bourdev, L. (2014a). Panda: Pose aligned networks for deep attribute modeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1637–1644.

Zhang, Z., Luo, P., Loy, C. C., and Tang, X. (2014b). Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision*.

Zhao, Y., Tang, F., Dong, W., Huang, F., and Zhang, X. (2018). Joint face alignment and segmentation via deep multi-task learning. *Multimedia Tools and Applications*, pages 1–18.