Feature Fusion for Online Mutual Knowledge Distillation

Jangho Kim, Minsung Hyun, Inseop Chung, Nojun Kwak
{kjh91, minsung.hyun, jis3613, nojunk}@snu.ac.kr
Seoul National University

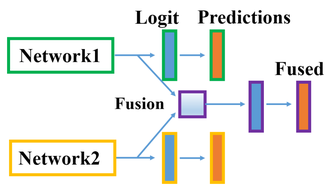
MIPAL laboratory
Machine Intelligence & Pattern Analysis

ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

Introduction

- Many researches on network architecture that extracts discriminative features.
- New approach : the feature fusion method that can combine different feature maps gained from multiple sub-networks.
- Feature fusion methods have been used in many previous deep learning studies.

DualNet – Example of feature fusion method



- DualNet which is feature fusion method trains independent two sub-networks with iterative training.
- This framework combine complementary two feature maps with fused classifier.

Motivation & Contribution

- Motivation

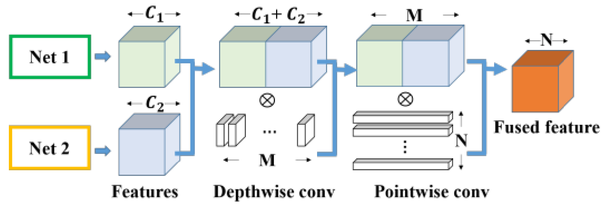
- Sub-networks can not help fused classifier with positive synergy.
- Only same architecture type can be used.

- Contribution

- Our method, Feature fusion learning (FFL) can improve the accuracy of sub-networks where gives positive synergy to a fused classifier.
- FFL can handle various architecture type.
- FFL can create meaningful feature maps used at computer vision tasks.

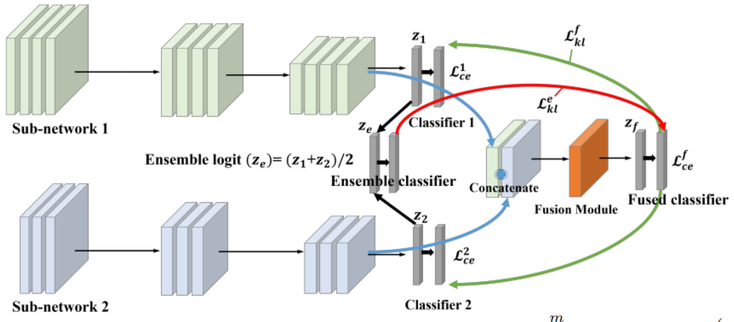
Method

Fusion Module



- Combining feature maps from the last layer of each sub-network with convolution operation.
- To reduce computational cost, FFL use Depth-wise and Point-wise convolution.
- Combined feature maps is named as fused feature.

Online Mutual Knowledge Distillation



- Ensemble knowledge distillation (EKD)

$$\mathcal{L}^e_{kl} = \sum_{i=1}^{m} \sigma_i(\mathbf{z}_e;T)\log\left(\frac{\sigma_i(\mathbf{z}_e;T)}{\sigma_i(\mathbf{z}_f;T)}\right)$$

- Using ensemble logits of sub-networks and knowledge distillation, fusion module can generate meaning feature map with this loss

- Fusion knowledge distillation (FKD)

$$\mathcal{L}^f_{kl} = \sum_{k=1}^{n}\sum_{i=1}^{m} \sigma_i(\mathbf{z}_f;T)\log\left(\frac{\sigma_i(\mathbf{z}_f;T)}{\sigma_i(\mathbf{z}_k;T)}\right)$$

- Using fused logits and knowledge distillation, sub-networks can be learned with this loss

- Feature fusion learning (FFL)

- With Cross entropy loss, EKD and FKD (Total loss), FFL framework trains sub-networks and fusion module

$$\mathcal{L}_{total} = \sum_{k=1}^{n}\mathcal{L}^k_{ce} + \mathcal{L}^f_{ce} + T^2 \times (\mathcal{L}^e_{kl} + \mathcal{L}^f_{kl})$$

Experiments

Comparison with Feature Fusion Method

(%)	CIFAR-10 DualNet	CIFAR-10 FFL	CIFAR-100 DualNet	CIFAR-100 FFL
ResNet-32	6.21±0.20	5.78±0.13	27.49±0.31	25.56±0.32
ResNet-56	5.67±0.12	5.26±0.17	25.87±0.29	23.53±0.25
WRN-16-2	5.92±0.16	5.97±0.13	25.71±0.20	24.74±0.31
WRN-40-2	4.94±0.10	4.6±0.13	23.22±0.25	21.05±0.25

(a) Top-1 classification error rate of fused classifiers. DualNet outputs results from the average of classifiers and FFL uses fusion module for classification.

(%)	CIFAR-10 DualNet	CIFAR-10 FFL	CIFAR-100 DualNet	CIFAR-100 FFL
ResNet-32	8.23±0.31	6.06±0.15	34.91±1.23	27.06±0.34
ResNet-56	7.34±0.25	5.58±0.13	32.67±1.14	24.85±0.30
WRN-16-2	7.53±0.20	6.09±0.09	31.7±1.00	25.72±0.28
WRN-40-2	6.25±0.14	4.75±0.16	28.4±0.61	22.06±0.20

(b) Top-1 classification error rate of sub-network classifiers.

Comparison with Knowledge Distillation Methods

	ResNet-32	ResNet-56
ONE	26.64 (26.94±0.21) {26.61*}	24.63 (25.10±0.29)
FFL-S	26.3 (26.66±0.21)	24.51 (24.85±0.31)
ONE-E	24.75 (25.19±0.20) {24.63*}	23.27 (23.59±0.24)
FFL	24.31 (24.82±0.33)	23.20 (23.43±0.19)

<Same architecture>

Net Types Net 1	Net 2	DML Net 1	DML Net 2	FFL Net 1	FFL Net 2
ResNet-32	WRN-16-2	28.31±0.28	26.45±0.30	27.06±0.26	25.93±0.30
ResNet-56	WRN-40-2	26.75±0.21	23.33±0.27	26.23±0.30	23.06±0.43

<Different architecture>

References

[1] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[2] Zagoruyko, Sergey, and Nikos Komodakis. "Wide residual networks." arXiv preprint arXiv:1605.07146 (2016).

[3] Zhang, Ying, et al. "Deep mutual learning." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[4] Hou, Saihui, Xu Liu, and Zilei Wang. "Dualnet: Learn complementary features for image recognition." Proceedings of the IEEE International Conference on Computer Vision. 2017.

[5] Zhu, Xiatian, and Shaogang Gong. "Knowledge distillation by on-the-fly native ensemble." Advances in neural information processing systems 31 (2018): 7517-7527.