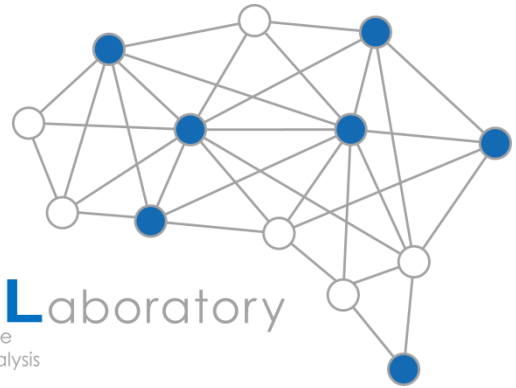
Feature Fusion for Online Mutual Knowledge Distillation

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

Seoul National University
Machine Intelligence Pattern Analysis Lab (MIPAL)
http://mipal.snu.ac.kr



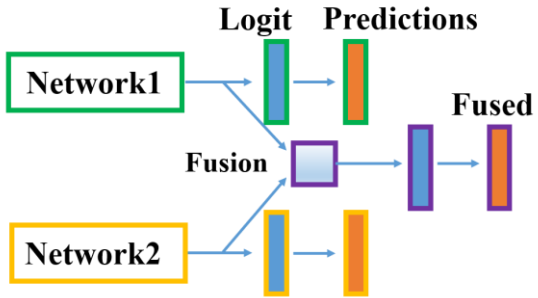
MIPAL aboratory
Machine Intelligence
& Pattern Analysis

Introduction

- Many researches on network architecture that extracts discriminative features.
  - ResNet
  - Wide-ResNet
- 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.

Introduction

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



Overall process of DualNet

Introduction

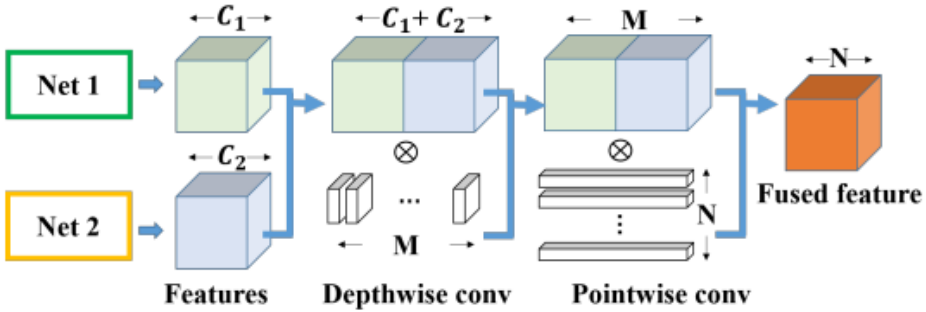
- DualNet coordinated two parallel sub-networks and fused the two-stream features, directly.
- Directly combining feature maps incurs several challenges:
  - The performance of the sub-networks is significantly lower than the performance of the network that is independently trained with the same architecture. This can also influence the performance of fused classifier
  - Because it directly combines feature maps, it is applicable only when the sub-networks have the same architecture.

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

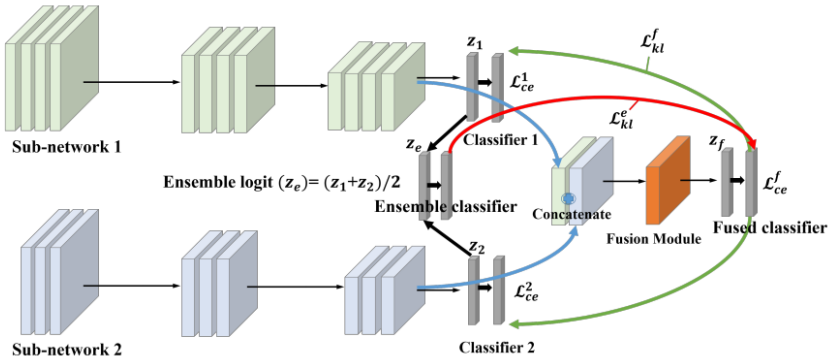
- Fusion module
  - Combining feature maps from the last layer of each sub-network with convolution operation
  - To reduce computational cost, FFL use Depth-wise and Point-wise convolution
  - Combined feature maps is named as fused feature



Fusion module

Method

- Ensemble knowledge distillation (EKD)
  - Using ensemble logits of sub-networks and knowledge distillation, fusion module can generate meaning feature map with this loss
- Fusion knowledge distillation (FKD)
  - Using fused logits and knowledge distillation, sub-networks can be learned with this loss



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

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

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

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.

case	FM	EKD	FKD	CIFAR-100 Fused	CIFAR-100 Sub-network
A	✓	✓	✓	25.56±0.32	27.06±0.34
B	✗	✓	✓	26.1±0.36	27.46±0.31
C	✓	✗	✓	27.03±0.31	28.36±0.44
D	✓	✗	✗	27.29±0.24	31.04±0.31

(Ablation study)

Experiments

- Comparison with Knowledge Distillation

	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)

	Method	Top-1	Top-5
ResNet-34	vanila	26.69	8.58
	ONE	25.61±0.02	7.96±0.02
	FFL-S	25.58±0.06	7.95±0.06
	ONE-E	24.48	7.31
	FFL	23.91	7.17

(ONE; same arch)

Net Types		DML		FFL	
Net 1	Net 2	Net 1	Net 2	Net 1	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

(DML; different arch)

Thank you

- The code is available at the “https://github.com/Jangho-Kim/FFL-pytorch”