Feature Fusion for Online Mutual Knowledge Distillation

Jangho Kim, Minsung Hyun, Inseop Chung, Nojun Kwak  
 już91, minsung.hyun, ji3613, nojunk)@snu.ac.kr  
Seoul National University

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.

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

\[ L_{total} = \sum_{k=1}^{K} L^{k}_{ECE} + L^{f}_{CE} + T^2 \times (L^{k}_{FKD} + L^{f}_{FKD}) \]

Experiments

Comparison with Feature Fusion Method

\[
\begin{array}{cccc}
& \text{CIFAR-10} & \text{CIFAR-100} \\
& \text{DualNet} & \text{FFL} & \text{DualNet} & \text{FFL} \\
\text{ResNet-32} & 6.21 \pm 0.20 & 5.78 \pm 0.13 & 27.49 \pm 0.31 & 25.36 \pm 0.32 \\
\text{ResNet-56} & 5.97 \pm 0.12 & 5.25 \pm 0.17 & 25.87 \pm 0.29 & 23.33 \pm 0.25 \\
\text{WRN-16-2} & 5.92 \pm 0.16 & 5.97 \pm 0.13 & 25.71 \pm 0.20 & 24.74 \pm 0.31 \\
\text{WRN-40-2} & 4.94 \pm 0.10 & 4.46 \pm 0.13 & 23.22 \pm 0.25 & 25.88 \pm 0.25 \\
\end{array}
\]

Comparison with Knowledge Distillation Methods

\[
\begin{array}{cccc}
& \text{ResNet-32} & \text{ResNet-56} \\
& \text{ONE} & \text{FFL-S} & \text{ONEE} & \text{FFL} \\
26.64 (26.94 \pm 0.21) & 26.61 & 24.63 & 23.30 \\
26.1 & 26.66 & 24.51 & 24.06 \\
24.75 (24.93 \pm 0.20) & 24.67 & 23.27 & 23.24 \\
24.31 (24.82 \pm 0.33) & 23.20 (23.43 \pm 0.19) & & \\
\end{array}
\]

References


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.

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.

Comparison with Feature Fusion Method

\[
\begin{array}{cccc}
& \text{CIFAR-10} & \text{CIFAR-100} \\
& \text{DualNet} & \text{FFL} & \text{DualNet} & \text{FFL} \\
\text{ResNet-32} & 6.21 \pm 0.20 & 5.78 \pm 0.13 & 27.49 \pm 0.31 & 25.36 \pm 0.32 \\
\text{ResNet-56} & 5.97 \pm 0.12 & 5.25 \pm 0.17 & 25.87 \pm 0.29 & 23.33 \pm 0.25 \\
\text{WRN-16-2} & 5.92 \pm 0.16 & 5.97 \pm 0.13 & 25.71 \pm 0.20 & 24.74 \pm 0.31 \\
\text{WRN-40-2} & 4.94 \pm 0.10 & 4.46 \pm 0.13 & 23.22 \pm 0.25 & 25.88 \pm 0.25 \\
\end{array}
\]

Comparison with Knowledge Distillation Methods

\[
\begin{array}{cccc}
& \text{ResNet-32} & \text{ResNet-56} \\
& \text{ONE} & \text{FFL-S} & \text{ONEE} & \text{FFL} \\
26.64 (26.94 \pm 0.21) & 26.61 & 24.63 & 23.30 \\
26.1 & 26.66 & 24.51 & 24.06 \\
24.75 (24.93 \pm 0.20) & 24.67 & 23.27 & 23.24 \\
24.31 (24.82 \pm 0.33) & 23.20 (23.43 \pm 0.19) & & \\
\end{array}
\]

<Same architecture>

\[
\begin{array}{cccc}
& \text{Net 1} & \text{Net 2} & \text{Net 1} & \text{Net 2} \\
& \text{ResNet-32} & \text{ResNet-56} & \text{ResNet-32} & \text{ResNet-56} \\
\text{WNR-16-2} & 28.33 \pm 0.28 & 26.45 \pm 0.30 & 27.06 \pm 0.36 & 25.93 \pm 0.30 \\
\text{WRN-40-2} & 28.77 \pm 0.21 & 23.83 \pm 0.27 & 26.23 \pm 0.30 & 23.06 \pm 0.47 \\
\end{array}
\]

<Different architecture>