A Compact Distilled Network by Accelerated Ensemble Learning for Mobile Visual Food Recognition

Heng Zhao, Kim-Hui Yap, and Alex C. Kot
School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
ICAI’19 Poster Paper

Abstract—Visual food recognition using mobile devices has attracted much attention in recent years, due to its portability and convenience in diet monitoring and health management. Common strategy used in prior works on food recognition involves deployment of a large server-based network, which is used to perform food recognition and provide feedback. However, such network is too large to be deployed on mobile devices directly. Even though some compact networks have been proposed, they are not able to achieve comparable performance similar to that of full-size networks. In view of this, this paper proposes a framework that enhances performance of a compact network by knowledge distillation from an ensemble of large networks, while the training process of the ensemble is greatly reduced by using Function Preserving Transformations.

Keywords: Food Recognition, Compact Network, Knowledge Distillation, Function Preserving Transformations

1. Introduction
Food recognition using mobile devices has raised much interest due to its convenience on diet monitoring and health management. A widely-used deep learning based mobile visual food recognition solution implements a client-server strategy, where a large server-based network is required. This may not be practical since it assumes the user has Internet access and they are willing to upload their private photos. Therefore, it is more beneficial to have a small network that can be deployed on mobile directly. However, the performance of a compact network is often inferior to that of a full-size network.

In view of this, this paper proposes a new framework that aims to enhance the performance of a compact network by knowledge distillation from an ensemble of teacher networks. Further, to address the issue of slow training of the large ensemble, our proposed network applies Function Preserving Transformations (FPT) to greatly reduce each individual training time of the teacher network, by transferring knowledge from a base network to each teacher network in the ensemble.

2. Methodology
Fig. 1 provides an overview of the proposed framework. The base network is based on a state-of-the-art large full network architecture with good performance, which generates better image representation but costs a large storage and longer time to train. Each teacher network in the ensemble shares similar structure to the base network but different in width of layer ($W_{T_n}^i$). The student network has a light structure and less storage requirement. To avoid training each teacher from scratch, we apply FPT which can instantaneously transfer the knowledge from the base network to each teacher network in the ensemble. Further, in order to boost the performance of the student network, we distill the knowledge from the teacher ensemble in the form of logits to the student.

2.1 Ensemble Learning with Function Preserving Transformations
A common strategy to enhance the compact student is to leverage on external knowledge provided by other large network or ensemble of networks. Although ensemble of networks provides a better image representation and generalization ability than a single network [4], it takes much longer to train each network from scratch. To reduce the
training time of each teacher network in the ensemble, we implement a similar idea as Net2Net [2], which allows a quick knowledge transfer from a network to a wider network in terms of layer width.

For a simplified case, suppose a network with layer \( z \) and layer \( z+1 \) that are consecutive fully-connected layers. Widening layer \( z \) requires to expand the weights of the layer. Assume the original layer \( z \) has \( a \) input neurons and \( b \) output neurons and layer \( z+1 \) has \( b \) input neurons and \( c \) output neurons, we can transform the layer \( z \) to have \( d \) outputs, with \( d > b \). For a simplified case, we randomly copy weights of existing neurons to form new neurons:

\[
g(j) = \begin{cases} 
  j, & \text{if } j \leq b, \\
  \text{randomweight from}\{1, 2, ..., b\}, & \text{if } j > b. 
\end{cases} \tag{1}
\]

The new weight matrix of \( g(j) \) for layer \( z+1 \) consists of its original neuron weights and some duplicated neuron weights copied from layer \( z \). The duplicated neuron weights input to layer \( z+1 \), value at each replication is divided by the number of repeating time, so all neurons in layer \( z+1 \) have the same value as in the original. Therefore, each wider teacher network in the ensemble can use the weights copy from the base network directly to speedup convergence.

### 2.2 Knowledge Distillation using Ensemble Logits

For an input feature image \( x_i \), the pre-softmax logits vector generated by one teacher \( \phi_t \) in the ensemble can be denoted as \( \mathbf{v}(1)_i \). Similarly, student \( \phi_s \) logits can be denoted as \( \mathbf{w}_i \).

The dimension of the vector \( \mathbf{v}(1)_i \) is the number of classes \( C \) in the dataset. A generalized softmax layer converts the logits \( \mathbf{v}(1)_i \) to the probability distribution \( q_i \) as follows:

\[
M_T(\mathbf{v}(1)_i) = q_i, \quad \text{where} \quad q_i^j = \frac{\exp(v(1)_i^j)}{\sum_k \exp(v(1)_i^k)}, \tag{2}
\]

For normal softmax classification, \( T = 1 \). By increasing its value, a "softer" probability distribution from the teacher \( \mathbf{v}(1)_i \) could be obtained. To calculate ensemble logits \( \mathbf{v}_i \), we average the logits from each teacher network in the ensemble by its importance, given by \( h_m \). Therefore, the ensemble logits can be formulated as:

\[
\mathbf{v}_i = \sum_{m=0}^{M} h_m \mathbf{v}(m)_i, \tag{3}
\]

where \( M \) represents the number of teacher models in the ensemble. We propose to minimize the Kullback-Leibler (KL) divergence between the soft probability distribution of the teacher ensemble and the normal probability distribution of the student, as a result, the knowledge distillation loss can be formulated as:

\[
L_{KLD}(\phi_t, \phi_s) = \frac{1}{N} \sum_{i=1}^{N} KL(M_T(\mathbf{v}_i)| | M_T(\mathbf{w}_i)), \tag{4}
\]

where \( N \) is the total number of training images. The loss function to train the student is given by a weighted combination of the cross entropy loss and the KD loss:

\[
L(\phi_s) = \eta L_{\text{softmax}}(\phi_s) + (1 - \eta)L_{KLD}(\phi_t, \phi_s), \tag{5}
\]

### 3. Results and Conclusion

To evaluate the performance of the proposed framework, we use the food benchmark Food-101 [1], which consists of 101 popular Western dishes with 1000 images per category. We use the official split of training and testing. We choose VGG-16 as the base network and the basic structure for \( M = 4 \) teacher networks. A MobileNet [3] is used as the student network.

Table 1 summarizes the Top-1 and Top-5 recognition accuracy of the student network. Results on a standalone student and distillation using only one teacher are included for comparison.

<table>
<thead>
<tr>
<th>Method</th>
<th>Top-1 (%)</th>
<th>Top-5 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standalone Student</td>
<td>82.5</td>
<td>93.8</td>
</tr>
<tr>
<td>One Teacher Distillation</td>
<td>84.0</td>
<td>94.5</td>
</tr>
<tr>
<td>Proposed method</td>
<td>84.7</td>
<td>95.2</td>
</tr>
</tbody>
</table>

Table 2 summarizes the iterations required to train the network, which including training the ensemble to reach optimal performance using FPT and training the student network to final performance using knowledge distillation, compared to the case without FPT.

<table>
<thead>
<tr>
<th>Method</th>
<th>Iterations (k)</th>
<th>Ensemble Training</th>
<th>Knowledge Distillation</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>142</td>
<td>54</td>
<td>190</td>
</tr>
<tr>
<td>Proposed method</td>
<td>101</td>
<td>54</td>
<td></td>
<td>155</td>
</tr>
</tbody>
</table>

In conclusion, this paper proposes a new framework for mobile visual food recognition. It leverages on knowledge distillation to enhance the performance of the compact network by using an ensemble of large networks. The training time of the ensemble is greatly reduced via knowledge transfer from a base network using Function Preserving Transformations.

### Acknowledgement

This research is supported by the National Research Foundation, Prime Minister’s Office, Singapore, under the NRF-NSFC grant NRF2016NRF-NSFC001-098. The research work was done at the Rapid-Rich Object Search (ROSE) Lab, Nanyang Technological University, Singapore.

### References


