Model Distillation with Knowledge Transfer from Face Classification to Alignment and Verification

Bhavana Jain

Week 10: Paper Summary

1 Introduction

Most previous studies on Knowledge Distillation focus on model distillation in the classification task, where they propose different architectures and initializations for the student network. However, only the classification task is not enough and the authors believe that other related tasks should also be considered. To solve this problem, this paper proposes model distillation with knowledge transfer from face classification to alignment and verification. The objective of face alignment is to locate the key-point locations in each image; while in face verification, we have to determine if two images belong to the same identity. By selecting appropriate initializations and targets in the knowledge transfer, the distillation can be achieved in the non-classification tasks.

2 Distillation Transfer

The knowledge transfer consists of two steps: transfer initialization and target selection.

2.1 Transfer Initialization

Since, face classification, alignment, and verification share the similar domain, the authors transfer the distilled knowledge of classification by taking its teacher and student networks to initialize corresponding networks in alignment and verification. The parameters of teacher and student networks in face classification are denoted as W_T^{cls} and W_S^{cls}. Analogically, they are W_T^{ali} and W_S^{ali} in alignment, while W_T^{ver} and W_S^{ver} in verification.

2.2 Target Selection

Based on the initialization, the second step is to select appropriate targets in the teacher network for distillation. The authors propose the general distillation

for non-classification tasks as follows:

$$L(W_S|W_S^{cls}) = \Phi(W_S, y) + \alpha H(P_S^\tau, P_T^\tau) + \beta\Psi(K_S, K_T)$$

where W_S and y denote the task-specific network parameter and label respectively. $\Phi(W_S, y)$ is the task-specific loss function. $H(P_S^\tau, P_T^\tau)$ denotes the cross-entropy loss between the temperature-softened softmax predictions of the student and teacher P_S^τ and P_T^τ. $\Psi(K_S, K_T)$ is the task-specific distillation term with the targets (hidden layers) selected as K_T and K_S in teacher and student networks. Besides, α and β are the balancing terms between classification and non-classification tasks.

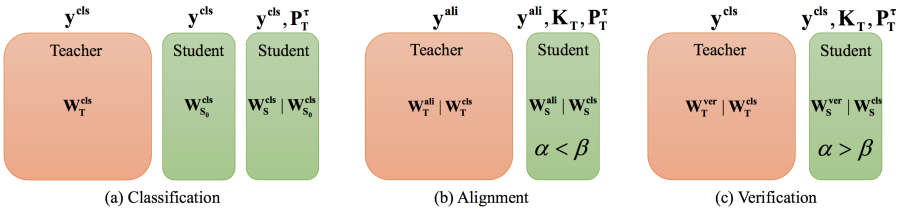

Figure 1: The proposed pipeline

2.2.1 Alignment

Face alignment is usually considered as a regression problem, thus the teacher network is trained by optimizing the Euclidean loss:

$$L(W_T^{ali}|W_T^{cls}) = \|R_T - y_{ali}\|^2$$

wherein R_T is the regression prediction of the teacher network and y_{ali} is the regression label. In distillation, apart from the soft predictions P_T^τ (classification target), another task-specific target is the teacher hidden layer K_T. The difference of key-point locations for different face identities is tiny. As a result, face identity is not the main influencing factor for these locations. Instead, pose and viewpoint variations have a much larger influence. Therefore, in face alignment, the hidden layer is preferred for distillation which gives the following loss function by setting $\alpha < \beta$,

$$L(W_S|W_S^{cls}) = \|R_T - y_{ali}\|^2 + \alpha H(P_S^\tau, P_T^\tau) + \beta \|K_S - K_T\|^2$$

2.2.2 Verification

For verification, the triplet loss is a widely used metric learning method, and the authors take it for model distillation. The teacher network can be trained as,

$$L(W_T^{ver}|W_T^{cls}) = \left[\|K_T^a - K_T^p\|^2 - \|K_T^a - K_T^n\|^2 + \lambda\right]_+$$

where $K_T^a, K_T^p and K_T^n$ are the hidden layers for the anchor, positive and negative samples respectively, i.e. a and p have the same identity, while a and n come from different identities. Besides, λ controls the margin between the positive and negative pairs.

Similar to face alignment, there are two possible targets in distillation – the hidden layer K_T and soft prediction P_T^τ. Since classification focuses on the difference of identities, i.e. the inter-class relation, and this relation can help a lot in telling if two images have the same identity. As a result, classification can be beneficial to boost the performance of verification. Therefore, in face verification, the soft prediction is preferred for distillation, which gives the following loss function by setting $\alpha > \beta$:

$$L(W_T^{ver}|W_T^{cls}) = \left[\|K_T^a - K_T^p\|^2 - \|K_T^a - K_T^n\|^2 + \lambda \right]_+ + \alpha H(P_S^\tau, P_T^\tau) + \beta \|K_S - K_T\|^2$$