Promoting Gradients Alignment in Hint Learning

PaperID 2345

Abstract

Hint learning (HL) is an efficient way to boost the performance of miniaturized neural networks in classification tasks, by jointly supervising the small model with original labels and hints (intermediate feature representation) generated by a heavy model. Whereas, it rarely gives such much improvement while extended to other tasks such as single-stage face detection task. In this work, we observe that the essential problem of HL is the misalignment between the gradients generated by mid-level feature hint and detection labels. To promote the alignment of these gradient flows, we propose a simple mechanism called incremental hint learning (IHL) that disentangles the feature hint learning and detection supervision in a gradual learning manner. Ablation study shows that compared with original HL, IHL can accelerate the process of gradients alignment via faster convergence training, and thus can optimize various miniaturized architectures like pruned ResNet, Inception and MobileNetV2 with stable improvement. Furthermore, we extend IHL to general object classification and face recognition tasks. Experiments on CIFAR10, CIFAR100, ImageNet and Megaface effectively verify its versatility for general tasks. For extremely miniaturized models around 30K parameters, IHL outperforms state-of-the-art methods by $3 \sim 5\%$ on FDDB, MALF, and AFW benchmarks. For general object classification and face recognition tasks, IHL can achieve a stable 1.0∼3.0% improvement over base models.

Introduction

Recent years CNN has shown its prominent capacity for improving the accuracy of face detection (Tang et al. 2018; Zhang et al. 2017b; 2017c; Liu et al. 2017; Zhang et al. 2017a; Shi et al. 2018; Song et al. 2018). This has made visual face detection an attractive application for domains ranging from surveillance to mobile terminal. However, both the accuracy and speed are the key requirements in many applications. The state-of-the-art face detection algorithms (Zhang et al. 2017a; Tang et al. 2018; Yang et al. 2017) have relied on deeper architectures or sufficient parameter, and they are associated with an enormous computational burden at runtime. We can obtain excellent performance with the support of powerful CNN architectures such as ResNet (He et al. 2016), GoogleNet (Szegedy et al. 2015) or SENet (Hu, Shen, and Sun 2017). But the huge parameters and computational cost of these popular backbones make them difficult to perform on applications with the real-time request. Thus, to achieve faster speed and comparable accuracy, some prior works explore the miniaturized CNN structures such as MobileNet (Howard et al. 2017; Sandler et al. 2018) and ShuffleNet (Zhang et al. ; Ma et al. 2018) or models with fewer channels and small filters (Iandola et al. 2016; Kim et al. 2016). In that direction, model compression (Han, Mao, and Dally 2015; Molchanov et al. 2016; Luo, Wu, and Lin 2017) is used to obtain miniaturized models in different computer vision tasks. While these methods obtain impressive speedup, the accuracy gap between original heavy models and miniaturized models is still large, especially on more complex tasks such as face detection. Given a pruned miniaturized CNN, how to maximize its performance has stimulated a lot of in-

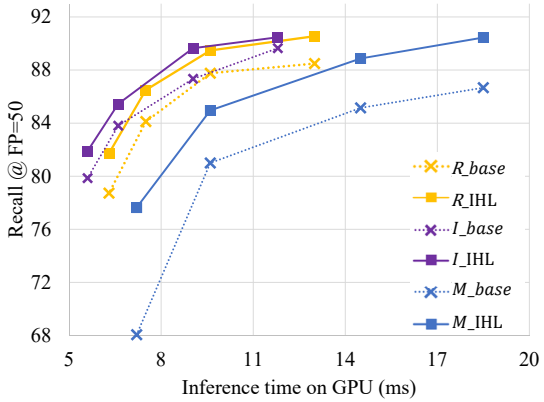


Figure 1: Speed (ms) versus recall on FDDB benchmark. Enabled by the IHL, the miniaturized detector outperforms the detector trained independently. The points from left to right in each curve represent the pruned models with different computational cost $0.01\times$, $0.04\times$, $0.16\times$ and $0.36\times$ compared to original detectors.

sightful works such as knowledge distillation (KD) (Hinton, Vinyals, and Dean 2015; Yim et al. 2017) and hint learning (HL) (Romero et al. 2014). The miniaturized model is used to mimic the behavior of a deeper or more complex model. This mimic learning approach can drive the small model to achieve better performance than training it independently. (Chen et al. 2017; Li, Jin, and Yan 2017; Wei et al. 2018) make attempts on the detection task by using the mimic learning algorithm and achieve good results. They all adopt the two-stage detectors such as powerful Faster-RCNN (Ren et al. 2015) and the student network learns both the behavior of the teacher network and the distribution of ground truth. However, in our experiments, the success of mimic learning on two-stage detectors is difficult to replicate on the single-stage detectors as shown in Tab. 2. We find it very difficult for miniaturized single-stage face detector to learn the behavior of the teacher network and the ground truth at the same time. We observe that the essential problem of these algorithms is the misalignment between the gradients generated by learning the behavior of the teacher and optimization by the detection labels. The smaller the parameter number of the single-stage detector, the more severe the impact of this phenomenon.

Based on this observation, we perform extensive exploratory experiments and find that proper decoupling of hint learning and detection supervision can effectively alleviate the gradient misalignment between them. Encouraged by this, we propose a simple but stable incremental hint learning (IHL) algorithm. The basic idea is that we expect to promote the alignment of gradients flows between learning the behavior of the teacher and optimization by detection labels. IHL can effectually disentangle the feature hint learning and detection supervision in a gradual learning manner.

For evaluating the stability and effectiveness of IHL, we firstly generate several miniaturized models with different computational cost by a common practice where the channels in each CNN layer are preserved with a fixed ratio. The base backbone networks ResNet50 (He et al. 2016), Inception (Szegedy et al. 2015) with Batch Normalization (Ioffe and Szegedy 2015) and MobileNetV2 (Sandler et al. 2018) are used and we intercept the network from *Input to Res3b3* (R), *Input to Inception_3c* (I) and *Input to Block_5_2 with stride 16* (M), respectively. Then, we train these models by the base method and IHL using the same configurations as shown in Fig. 1. For each CNN structure, we conduct several experiments on detectors with different computational cost. The inference time is evaluated on GTX TITAN X with 720P input. We can observe that given the arbitrarily generated miniaturized detectors, IHL can always improve the performance with a considerable gap compared with training it independently, even though the models only have $\sim 20K$ parameters.

To summarize, the contributions of this paper are as follows:

1) We propose a simple but robust method named incremental hint learning (IHL) which disentangles the feature hint learning and detection supervision in a gradual learning manner. IHL can effectively promote the gradients alignment and lead to the stable improvement.

2) To our best knowledge, we are the first to observe the gradient misalignment between traditional hint learning and face detection. We also demonstrate the detailed proof of IHL via theory and experiments.

3) Extensive experiments on standard face detection benchmarks FDDB, AFW and MALF strongly validate the stability and effectiveness of the IHL. Furthermore, IHL can be easily extended to other tasks such as general object classification and face recognition.

Related Work

Fast face detector. Face detection has achieved excellent performance with the assistance of powerful CNN (He et al. 2016; Szegedy et al. 2015). Many object detection methods have been applied to face detection task and the detection pipeline can be divided into single-stage and two-stage detectors. For two-stage detectors, the seminal detection algorithms are RCNN (Girshick et al. 2016) and a series of variants (Girshick 2015; Ren et al. 2015; Dai et al. 2016). Inspired by these insights, (Jiang and Learned-Miller 2017; Wang et al. 2017c; 2017b) have obtained the state-of-the-art performance on FDDB (Jain and Learned-Miller 2010) using the two-stage detection pipeline. Encouraged by the success of single-stage detectors such as SSD (Liu et al. 2016), YOLO (Redmon et al. 2016) and RetinaNet (Lin et al. 2018) in general object detection, many algorithms are inspired by them to perform on face detection task. (Zhang et al. 2017c) proposes a real-time face detector S^3FD which performs superiorly on large scale-variance. Other works (Tang et al. 2018; Liu et al. 2017; Song et al. 2018; Hao et al. 2017; Shi et al. 2018; Najibi et al. 2017; Zhang et al. 2017a) also propose novel methods to design single-stage detector with comparable performance. Most of these algorithms focus on improving the performance or accelerating the detection algorithms and little attention has been paid to the training of miniaturized face detectors with pruned backbones.

Hint learning. With the growing demand for face detection, how to train an efficient miniaturized face detector has become the bottleneck of the practical application. The seminal works in this field are knowledge distillation (Hinton, Vinyals, and Dean 2015; Ba and Caruana 2014) and hint learning (Romero et al. 2014). The former method trains a single neural network by mimicking the soft targets generated by a teacher model and the later algorithm introduces the hint-based training method where students not only mimic the soft targets of the teacher network but also learn the feature maps generated by the guided layers. Other researchers (Zhang et al. 2018; Yim et al. 2017) propose some variants based on mimic learning. However, these approaches mainly focus on classification tasks which are not as complex as detection tasks. To train more efficient miniaturized detectors, (Li, Jin, and Yan 2017; Wei et al. 2018; Chen et al. 2017) make attempts on detection tasks with the assistance of hint-based learning and knowledge distillation where the student detectors are optimized by both the detection labels and the hints generated by a heavy model. Whereas, the success of mimic learning on two-stage detectors is difficult to replicate on the single-stage detectors. We observe that there exists the gradients misalignment between

the HL and detection supervision, severely restricting the performance of the student detector. In order to alleviate this limitation, we propose a simple mechanism called incremental hint learning (IHL) for training single-stage miniaturized face detector. Without bells and whistles, IHL demonstrates its' stability and effectiveness by extensive experiments on variant CNN structures. Furthermore, it's easy to extend IHL to other tasks such as general object detection and large-scale face recognition.

Method

Preliminaries

Before describing our exploration, we provide a mathematical description of gradients misalignment in hint learning and a toy example to illustrate the influence of this problem.

Given the detection task τ with hint learning that induces two losses L_{det} and L_{hint} (**introduced in supplementary material**), we propose to parameterize the solution for τ by a neural network $f(\cdot, w)$ where w is the parameters set. Generally, the full pipeline is to minimize $L = L_{det}(w) + \lambda L_{hint}(w)$ where the λ is to control the balance of face detection and hint learning. Inspired by (Du et al. 2018), define the w_{ij} as the parameter in the j-th kernel of layer i, the gradient misalignment is formulated as

$$\mathcal{G} = \frac{1}{M} \sum_{i=1}^{I} \sum_{j=1}^{J_i} \mathcal{C}(\nabla_{w_{ij}} L_{det}(w), \nabla_{w_{ij}} L_{hint}(w)) \qquad (1)$$

where I and J_i mean the total layer number in CNN and the kernel number in i-th layer. M is the total kernel number in the whole CNN. The $\mathcal{C}(\cdot, \cdot)$ is the cosine similarity. Based on this definition, we conduct several experiments on FDDB with 0.01x_R as shown in Fig. 2. By disturbing the λ in Eq. 1, we obtain different curves to represent the relation between angle of gradients and performance. It's interesting that along with the training, the speed of gradients alignment greatly influences the model performance. For this observation, we first introduce the proposed IHL and then give the detailed proof that it can effectively promote the gradients alignment and lead to the stable improvement.

Incremental hint learning

Based on the former conclusion, IHL is proposed by properly decoupling the feature hint learning and detection supervision in a gradual learning manner. The whole IHL contains initial hint learning, hint-based detection learning and improved refine learning. Each learning stage plays different roles in the training of the student model and the details of them are as follows.

Initial hint learning. We first introduce the initial hint learning to help miniaturized detectors obtain a good initialization. Given a teacher network T with parameter w_T, let the head of the face detector be the hint-guided layer with parameter w_{ht}. The corresponding parameters in the student network S and hint layer are w_S and w_{hs}. The loss of initial hint learning can be formulated as:

$$L_{hint} = \frac{1}{2} ||u(S(x; w_S); w_{hs}) - v(T(x; w_T); w_{ht})||^2 \qquad (2)$$

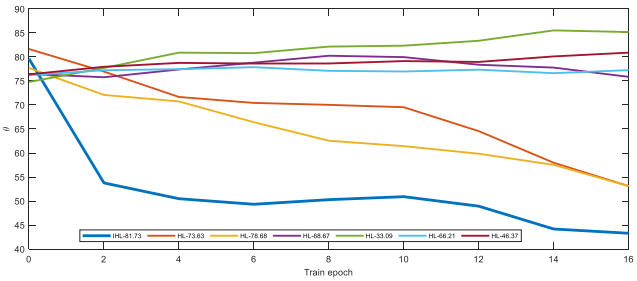


Figure 2: Change of $\theta =< G_*^1, G_*^2 >$ along with the training epoch. The number in legend is the recall at FP50 on FDDB. The λ in IHL is set to value 1 and in HL is set to 1, 10, 20, 50, 200, 500 along with the decrease of performance.

where x means the input image. *u* and *v* represent the hint layer in the student network and the hint-guided layer in teacher network, respectively. In this learning stage, to avoid the gradient misalignment between L_{det} and L_{hint}, L_{det} is omitted.

Hint-based detection learning. The teacher network T can achieve excellent performance due to the sufficient parameters and complex backbone. The *hint* generated by the teacher network contains sufficient semantic information and after completion of initial hint learning, the detection task can easily verify faces from the mimicked *hint*. Empirically, based on this, detection task can converge faster. Thus, in this learning stage, the loss function in this stage is:

$$L = \lambda L_{hint} + L_{det} \qquad (3)$$

This eventually leads to more consistent gradients between the convergent L_{hint} and detection supervision L_{det}.

Improved refine learning. It's an important problem that the small capacity of the miniaturized detectors is insufficient to support it for learning the precise semantic information from the teacher's hint. Affected by the biased hint, the optimization of the detector may suffer from the saddle point problem (Dauphin et al. 2014). In order to alleviate this phenomenon, we design the improved refine learning to further improve the performance of the detector. We gradually lift the restriction of L_{hint} by linearly decreasing and drive the student to further mining the foreground information from the background according to the supervision of ground-truth label. Thus the final loss function in this stage is:

$$L = max(0, 1 - \mathcal{E}) L_{hint} + L_{det} \qquad (4)$$

where $\mathcal{E}$ means the training epoch in the third learning stage. During this stage, the weight of L_{hint} will gradually decrease to 0 in the first epoch. Based on the information learned by the former learning stages, the supervision of ground-truth will guide the detector to obtain further improvement without the limitation of *hint*.

Optimization.

The IHL is performed in each mini-batch based model update step and throughout the whole training process. The optimization of student detector is conducted iteratively until

Algorithm 1 Incremental Hint Learning

Input: Training set $\mathcal{X}$, label set $\mathcal{Y}$, learning rate γ, training epoch state $\mathcal{E}$. The epoch boundaries between different learning stages $\mathcal{E}_1$, $\mathcal{E}_2$ and $\mathcal{E}_3$.

Initialize: Teacher model T with strong performance and student model θ with 'xavier' initialization (Glorot and Bengio 2010).

$\mathcal{E} \leftarrow 0, \gamma \leftarrow 0.01$
while $\mathcal{E} \geq 0$ and $\mathcal{E} < \mathcal{E}_1$ **do**
  Randomly sample data x from $\mathcal{X}$.
  Update the L_{hint} according to Eq. 2.
  Update θ: $\theta \leftarrow \theta + \gamma \frac{\partial L_{hint}}{\partial \theta}$
  Update $\mathcal{E}$ and γ.
end while
$\gamma \leftarrow 0.001$
while $\mathcal{E} \geq \mathcal{E}_1$ and $\mathcal{E} < \mathcal{E}_2$ **do**
  Randomly sample data x from $\mathcal{X}$.
  Update the L according to Eq. 3.
  Update θ: $\theta \leftarrow \theta + \gamma \frac{\partial L}{\partial \theta}$
  Update $\mathcal{E}$ and γ.
end while
$\gamma \leftarrow 0.001$
while $\mathcal{E} \geq \mathcal{E}_2$ and $\mathcal{E} \leq \mathcal{E}_3$ **do**
  Randomly sample data x from $\mathcal{X}$.
  Update the L according to Eq. 4.
  Update θ: $\theta \leftarrow \theta + \gamma \frac{\partial L}{\partial \theta}$
  Update $\mathcal{E}$ and γ.
end while

convergence. We define the $[0, \mathcal{E}_1)$, $[\mathcal{E}_1, \mathcal{E}_2)$ and $[\mathcal{E}_2, \mathcal{E}_3]$ to represent the different learning stages in IHL. Note that in IHL, there are no hyperparameters that need fine-tuning and the $\mathcal{E}_i$ can be easily set to ensure the convergence at each learning stage. The optimization details are summarized in Alg. 1.

Proof of incremental hint learning

For simplicity, define the (G_I^1, G_I^2) and (G_H^1, G_H^2) as the gradients of hint generated by L_{det} and L_{hint} in IHL and HL, respectively.

Lemma 1. *Define the $F_S \in \mathbb{R}^D$, whether hint learning or IHL, with the initialization parameter w in the classification layer, there are*

$$\lim_{D \to +\infty} < G_*^1, G_*^2 > = 0.$$

Consider the two gradients G_*^1 and G_*^2 for the feature F_S w.r.t. the two loss functions L_{hint} and L_{det} that are symmetrically and randomly distributed in the D-dim space (Wan et al. 2018), the volume of D-dim sphere with radius R is:

$$\begin{aligned} V_D(R) =& (\int_0^R r^{D-1} dr)(\int_0^\pi sin^{D-2}(\varphi_1) d\varphi_1) \cdots (\int_0^{2\pi} d\varphi_{D-1}) \\ =& \frac{R^D}{D} \frac{\Gamma(\frac{D-1}{2})\Gamma(\frac{1}{2})}{\Gamma(\frac{D}{2})} \cdots \frac{\Gamma(1)\Gamma(\frac{1}{2})}{\Gamma(\frac{3}{2})} \cdot 2 \frac{\Gamma(\frac{1}{2})\Gamma(\frac{1}{2})}{\Gamma(1)} \\ =& \frac{\pi^{\frac{D}{2}}}{\Gamma(\frac{D}{2}+1)} R^D, \end{aligned} \tag{5}$$

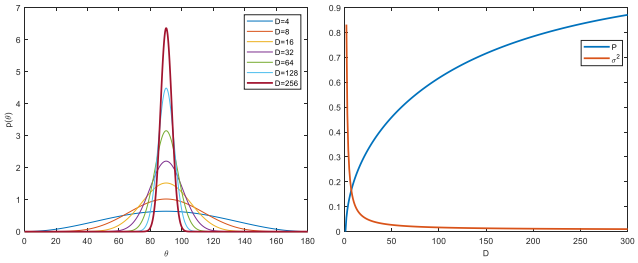


Figure 3: Probability density function p(θ) in D-dim space. θ indicates angle between G_*^1 and G_*^2. σ^2 and P in right fig indicate the variance and the probability that θ is in [85,95].

where $\Gamma(\cdot)$ indicates the gamma function. The surface area of it can be formulated as:

$$S_D = \frac{dV_D}{dR} = \frac{2\pi^{\frac{D}{2}}}{\Gamma(\frac{D}{2})} R^{D-1} \tag{6}$$

Assume that $\nabla_{F_S} L_{hint}(w)$ is on a coordinate axis and $\nabla_{F_S} L_{det}(w)$ has an angle of θ with it. The probability density can be computed as:

$$p(\theta)d\theta = \frac{\Delta S}{S_D} = \frac{S_{D-1}(Rsin\theta)Rd\theta}{S_D(R)} \tag{7}$$

According to Eq.(5)(6)(7), the probability density function of θ can be formulated as:

$$p(\theta) = \frac{\Gamma(\frac{D}{2})}{\Gamma(\frac{D-1}{2})} \frac{sin^{D-2}(\theta)}{\sqrt{\pi}} \tag{8}$$

As shown in Fig. 3, G_*^1 and G_*^2 tend to be orthogonal as the hint dimensions increase.

Lemma 2. *The detection task in hint-based detection learning convergences faster than the traditional HL.*

Following the denotation of G_*^1 and G_*^2, The angle α_* between the total gradients and gradient of the detection task is:

$$\alpha_* = arccos(\frac{G_*^1 + \lambda G_*^2}{\|G_*^1 + \lambda G_*^2\|} \cdot \frac{G_*^1}{\|G_*^1\|}) \tag{9}$$

Based on the **Lemma 1** that $G_*^1 \cdot G_*^2 \approx 0$, there are:

$$\alpha_* \approx arccos(\frac{\|G_*^1\|}{\|G_*^1 + \lambda G_*^2\|}) \tag{10}$$

After the initial hint learning stage in IHL, $\|G_I^2\|$ is much less than $\|G_H^2\|$ and also the $\|G_I^1\|$ is approximately equal to $\|G_I^2\|$ due to the random initialization of weights in detection head. So this naturally leads to that $\alpha_I < \alpha_H$ and thus the detection task can convergence faster than it in HL because of the consistency of gradients direction between G_*^1 and $G_*^1 + \lambda G_*^2$.

Lemma 3. *L_{hint} will not be greatly disturbed alone with the faster converging L_{det}.*

For the loss function $L_*(w)$ in IHL, to approximate the change of it, we use the first-degree Taylor polynomial, and the taylor expansion at weight point $w = w^{(0)}$ is

$$L_*(w) \approx L_*(w^{(0)}) + (w - w^{(0)})^T J_* + \frac{1}{2}(w - w^{(0)})^T H_* (w - w^{(0)}) \tag{11}$$

where J_* and H_* are the jacobian matrix and hessian matrix at weight point $w^{(0)}$ and $G = G_I^1 + \lambda G_I^2$. Define the ε as the learning rate and the new weight point after one step will be $w^{(1)} = w^{(0)} - \varepsilon G$. The Eq. 11 can be updated as:

$$L_*(w^{(1)} \approx L_*(w^{(0)}) - \varepsilon G^T J_* + \frac{1}{2}\varepsilon^2 G^T H_* G. \quad (12)$$

As the seminal works done (Tian et al. 2019; Sagun, Bottou, and LeCun 2016; Sagun et al. 2017; Lipton 2016; Baity-Jesi et al. 2018), deep networks often converge to 'flat minima' whose hessian matrix has a lot of small eigenvalues. Thus, the J_{hint} and H_{hint} has the smaller numerical distribution and there are:

$$\Delta L_{hint} \approx -\varepsilon G^T J_{hint} + \frac{1}{2}\varepsilon^2 G^T H_{hint} G \approx 0 \quad (13)$$

Note that the IHL can converge faster along with the converging detection task.

Lemma 4. *Faster convergence of IHL promotes the gradients alignment and leads to better performance.*

As illustrated in Fig. 2, along with the convergence of $L_{det} + L_{hint}$, the gradients between them will gradually become harmonious. The faster converged detection task based on the converged hint learning can accelerate the process of harmonious gradient and lead to the higher performance.

Discussion

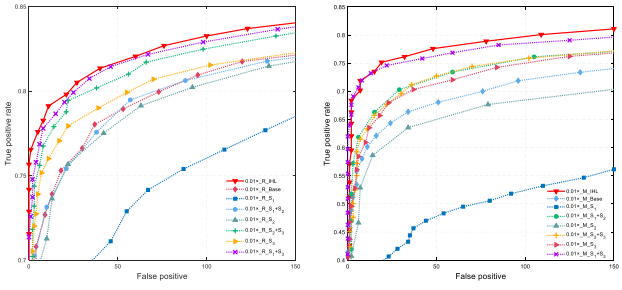
It's natural to ask that **can HL promote the gradients alignment via adjusting the λ?** To better evaluate this, we conduct several experiments with variant λ as shown in Fig 2. A suitable λ will moderately promote the the gradient alignment, which is consistent with the fact that the hint learning needs to carefully adjust the loss weight to get the performance improvement. However, it's still difficult to achieve the same promotion effect as IHL.

Experiments

In this section, we first introduce our detailed setup of experiments and then ablation studies are conducted to validate the effectiveness of each learning stage in IHL. Exhaustively experiments with different CNN structures are performed on standard face detection benchmarks. And also, we compare our methods with other traditional mimicking methods on training single-stage miniaturized face detectors. We also quantitatively and qualitatively analyze the effectiveness and stability of the IHL (refer to supplementary materials). Finally, we extend IHL to other tasks such as general object classification and face recognition to verify its versatility.

Setup and implementation details

In order to evaluate the stability and effectiveness of IHL, we conduct **16 detectors** corresponding to Sec. including a teacher network T with stride 16 (*from input to res3b3 in ResNet50 (He et al. 2016)*) , three large detectors (*the original ResNet50 (R_L) (He et al. 2016), Inception (I_L) (Szegedy et al. 2015) and MobileNetV2 (M_L) (Sandler et al. 2018)*) and other 12 student models with stride 16. We intercept the network R_L, I_L and M_L from *Input to Res3b3* (R), *Input to Inception_3c* (I) and *Input to Block_5_2* (M), respectively. We conduct four pruned detectors with different computational cost $0.01\times$, $0.04\times$, $0.16\times$ and $0.36\times$ by reducing the channel number in each layer for each CNN structure. The details of the students' speed and parameter number are shown in Tab. 1.



(a) $0.01\times\_R$ (b) $0.01\times\_M$

Figure 4: Ablation study on FDDB. S_i means the index of the learning stage in IHL.

The FDDB (Jain and Learned-Miller 2010), AFW (Zhu and Ramanan 2012) and MALF (Yang et al. 2015) are used for testsets and the protocols are the same as (Liu et al. 2017). Our training set is same to (Liu et al. 2017). The base face detection pipeline is similar to (Song et al. 2018). Encouraged by the conclusion of SNIP (Singh and Davis 2018) and S^3FD (Zhang et al. 2017c), our detector mainly focuses on the specific face scale range corresponding to the receptive field of the detector. The detecting face scale range is set to $[64, 128]$. Faces will be treated as positive samples if the scales of them fall into the detecting face scale range. To accommodate the configuration of the detector, we only adopt an anchor with size $[64\sqrt{2}, 64\sqrt{2}]$. For training this face detector, we assign a binary class label to the anchor. We assign a positive label to two kinds of anchors: (i) the center of the anchor and the ground-truth box are coincidence which can guarantee the recall of the detector, or (ii) an anchor that has an IOU overlap higher than 0.5 with any ground-truth boxes, otherwise it will be assigned as the negative sample. The loss function of detection is L_{det}, the sum of binary-class cross entropy loss and smooth L1 loss (Jiang and Learned-Miller 2017). In the training process, we resize the image to make sure at least one face falls into the scale of [64, 128]. All of the detectors are trained from scratch.**More details are introduced in supplementary materials**.

Ablation study on IHL

In this section, extensive experiments are conducted to evaluate this. Using different combinations of these learning stages, where only one stage, two stages, or the whole IHL are performed. $0.01\times\_R$ and $0.01\times\_M$ are evaluated on FDDB benchmark. Fig. 4 shows the performance of the baseline and the detectors trained by different learning stages. During the inference stage of the detectors trained with the only stage S_1, we use the detection head of the teacher model to detect faces from the mimicked hint because, in the initial hint learning stage, the detection head of

Model	#Parameter	Inference time(ms)	Recall@FDDB FP=50	Recall@AFW FP=1	Recall@MALF 1%FPPI
R_L	∼24.04M	46.4/131.8	91.03	80.38	82.81
I_L	∼ 10.44M	34.8/102.2	91.71	92.11	84.49
M_L	∼ 2.56M	72.1/259.1	90.3	93.82	83.44
T	∼1.58M	19.7/97.7	91.73	95.95	83.73
0.36×_R	∼611k	13/38.7	88.47/+2.10	94.03/+0.85	79.92/+1.81
0.36×_I	∼443k	11.8/35.4	89.65/+0.8	92.96/+1.38	79.77/+1.94
0.36×_M	∼109k	18.5/51.6	86.67/+3.78	92.32/+3.2	78.33/+4.75
0.16×_R	∼296k	9.6/30.4	87.72/+1.74	92.75/+2.13	79.13/+1.92
0.16×_I	∼218k	9.04/28.6	87.27/+2.31	92.32/+1.71	77.39/+3.02
0.16×_M	∼222k	14.5/46	85.15/+3.72	91.47/+3.0	76.38/+5.06
0.04×_R	∼92k	7.5/26.8	83.81/+1.62	87.63/+5.33	73.72/+2.89
0.04×_I	∼71k	6.6/23.6	83.81/+1.62	78.68/+9.38	71.75/+2.95
0.04×_M	∼36k	9.6/34.4	81.0/+3.96	76.55/+3.43	66.96/+8.80
0.01×_R	∼34k	6.3/22.6	78.73/+3.0	71.0/+8.32	61.74/+5.98
0.01×_I	∼27k	5.6/20.6	79.76/+2.14	65.67+4.69	62.56/+5.34
0.01×_M	∼15k	7.2/28.5	68.06/+9.64	62.47+7.89	53.6/+8.04

Table 1: The details of the speed and the performance for all of the base detectors. The inference time is evaluated on the GTX TITAN X. The first value is tested with the 720P input and the second value is tested on FDDB with multi-scale testing. The two numbers in the performance mean baseline performance and the improvement of IHL relative to baseline with red color.

the student is not be optimized. It's obvious that the detector trained only by the initial hint learning (S_1) depicts very poor performance. The hint mimicked by the student has a big gap with the teacher's hint, especially for the miniaturized detectors. The traditional hint learning (only using S_2 in our experiments) obtains almost no improvement or even decline for the miniaturized detectors and the performance of S_1+S_2 is better than S_2 with a clear margin. Although the performance of S_1 is poor, we can find it can also help improve the performance of S_2 and S_3.

Note that S_2 can obtain the better performance based on S_1 and S_3 can further improve the performance based on the former $S_1 + S_2$ which shows the robustness of the incremental hint learning. From the extensive ablation studies, we can conclude that each learning stage plays different roles for training the miniaturized detectors.

Detailed evaluation on IHL

In order to evaluate our IHL method on different popular CNN structures, we conduct extensive experiments on detectors with variant computational cost. The details of these detectors are shown in Tab. 1. All of the experiments are trained with the same number of epochs and BatchNorm (Ioffe and Szegedy 2015) is used for accelerating the convergence. Following the views of (Li, Jin, and Yan 2017) that the performance of the small network directly depends on the large model in mimic learning, we adopt T (*Input to Res3b3 in ResNet50*) with excellent performance as the teacher. Results demonstrate that IHL can effectively improve the performance of arbitrary structures with variant computational cost by a large margin, especially for tiny detectors with $\sim 20K$ parameters.

Considering that we are not to focus on which kind of network architectures perform better on face detection task, instead we want to show the IHL can steadily improve the performance of miniaturized single-stage face detectors. The inference time for different detectors in Tab. 1 is evaluated on GTX TITAN X with the multi-scale testing where the image is scaled to have long edges of $1400\times2^k, k = \{0, -1, \cdots, -4\}$. Note that we only record the inference time of the detectors and other processing time is not included.

Model	Method	FDDB	AFW	MALF
0.01×_R	IHL	**81.73**	**79.32**	**67.72**
0.01×_R	BaseLine	78.73	71.0	61.74
0.01×_R	KD	78.62	75.27	63.29
0.01×_R	HL+KD	78.68	65.03	63.66
0.01×_M	IHL	**77.7**	**70.36**	**61.64**
0.01×_M	BaseLine	68.06	62.47	53.6
0.01×_M	KD	72.24	68.66	59.38
0.01×_M	HL+KD	70.37	64.18	55.74

Table 2: Comparison with KD and HL+KD on different benchmarks. The evaluation metrics in each benchmark are the same as Tab. 1. 0.01×_R and 0.01×_M are used as the student detectors. All of the experiments are trained with the same number of epochs.

Comparison with distillation and other methods

As our method is closely related to model distillation and hint learning, we provide a focused comparison to Distillation (KD) (Hinton, Vinyals, and Dean 2015) and jointly KD and Hint Learning (HL) (Romero et al. 2014; Chen et al. 2017) with the miniaturized CNN structures 0.01×_R and 0.01×_M. Note that the S_2 in IHL shares the same operation with hint-based detection learning (HL) (Li, Jin, and Yan 2017) and in ablation study, the performance between it and IHL has been compared. Tab. 2 compares our IHL with knowledge distillation and hint learning. In knowledge distillation (KD), the teacher network T is pre-trained and pro-

Network	Top1 (%)	Top5 (%)
ResNet50 (Teacher)	76.1	92.9
ResNet18	69.7	89.0
MobileNetV2_half	64.4	85.5
MobileNetV2	71.3	89.9
ResNet18 + IHL	71.3	89.8
MobileNetV2_half + IHL	65.7	86.6
MobileNetV2 + IHL	72.8	91.2

Table 3: Performance on ImageNet, comparison for different networks.

vides fixed classification and regression targets for the student network. Compared with KD, the hint targets are also provided in HL. Note that in KD and KD+HL, the supervision of the ground-truth is also assigned. Tab. 2 shows that IHL significantly and steadily enhances the performance of miniaturized single-stage detectors.

Generality of IHL on general tasks

Implement details

In these experiments, the feature layer (fully connected layer before the classification layer) is assigned as the 'hint' layer. The large-scale datasets ImageNet (Russakovsky et al. 2015) and the cleaned MegaFace (Deng et al. 2018) are used as the test benchmarks. L_{det} is replaced by L_{cls} (the classification loss in classification tasks).

For general object classification, ResNet50 (He et al. 2016) is assigned as the teacher to guide the learning of ResNet18 (He et al. 2016), MobileNetV2 (Sandler et al. 2018) and MobileNetV2_half which is the half channel version of MobileNetV2. For training on ImageNet, the total training epoch is 150 and the base learning rate is set to 0.8 with a decrease of 0.1 at epoch 30, 60 and 90. We use the standard SGD with the momentum 0.9 and weight decay 0.0001. For face recognition, we evaluate the effectiveness of IHL on learning ensemble of multi-models. We use the ensemble of model R100 (Deng et al. 2018), PolyNet (Zhang et al. 2017d) and Attention56 (Wang et al. 2017a) to guide the learning of PolyNet, R18 and MobileNetV2. MS1MV2 (Deng et al. 2018) and IMDb-Face (Wang et al. 2018) are used as the training set. We remove the last global average pooling layer of these models for face recognition. The configuration of loss function and training parameters are the same as (Deng et al. 2018). For CIFAR10 and CIFAR 100, we follow the configuration of CETL (Chen, Zhang, and Dong 2018). ResNet26 and CI-CNN in CETL are used as the teacher and student. The last fully connected layer before the classification layer is assigned as the hint-guided layer and hint layer, respectively.

Results on classification and face recognition

The results on object classification and face recognition are shown in Tab. 3 and Tab. 4. With the guidance of the teacher network, IHL can steadily improve the performance of the student network by 1$\sim$1.5% whether learning a single teacher network or the ensemble of multiple networks. It is worth noting that when IHL is performed on the large model PolyNet, its performance even exceeds the teacher network. Results of CIFAR10 and CIFAR100 are shown in Tab 5 where IHL has the higher performance than other methods with a large margin and even better than the teacher in CIFAR100.

Network	Id (%)	# Param
PolyNet	98.52	70.87M
R100	98.48	58.72M
Attention56	98.49	55.5M
Ensemble (Teacher)	98.74	-
R18	95.66	17.6M
MobileNetV2	93.42	18.29M
PolyNet+IHL	98.8	70.87M
R18+IHL	96.58	17.6M
MobileNetV2+IHL	94.81	18.29M

Table 4: Face identification of different networks on Megaface challenge1. We use the refined probe set and 1M distractors provided by (Deng et al. 2018). "Id" refers to the rank-1 face identification accuracy with 1M distractors. # Param means the parameter number.

Method	CIFAR-10	CIFAR-100
Teacher	92.34%	65.36%
Student	87.38%	61.25%
HL (Romero et al. 2014)	88.57%	61.28
KD (Hinton, Vinyals, and Dean 2015)	88.45%	61.03
FSP DNN (Yim et al. 2017)	88.70%	63.33
AT (Zagoruyko and Komodakis 2016)	89.14%	62.13%
CETL (Chen, Zhang, and Dong 2018)	89.11%	64.83
IHL	**90.99%**	**68.56%**

Table 5: Comparison on CIFAR-10 and CIFAR-100. The references will be added in the revised version due to the limited space.

Conclusion

In this paper, we focus on training the strong and efficient miniaturized single-stage face detectors and we observe that the traditional HL will bring the essential problems where the gradients generated by mimicking the teacher model and learning the detection labels are misaligned. To alleviate the misalignment of gradient flows, we propose a stable and effective mechanism called incremental hint learning(IHL) that disentangles the feature hint learning and detection supervision in a gradual learning manner. Extensive experiments and analysis demonstrate that IHL is excellent on training the miniaturized face detectors. Furthermore, the hyperparameter-free IHL can be easily extended to other tasks such as general object classification and face recognition. Experiments on ImageNet and Megaface show that IHL can steadily increase the performance by 1.0$\sim$1.5% over base models and it can also lead to improvement even for large models.

References

Ba, J., and Caruana, R. 2014. Do deep nets really need to be deep? In *Advances in neural information processing systems*, 2654–2662.

Baity-Jesi, M.; Sagun, L.; Geiger, M.; Spigler, S.; Arous, G. B.; Cammarota, C.; LeCun, Y.; Wyart, M.; and Biroli, G. 2018. Comparing dynamics: Deep neural networks versus glassy systems. *arXiv preprint arXiv:1803.06969*.

Chen, G.; Choi, W.; Yu, X.; Han, T.; and Chandraker, M. 2017. Learning efficient object detection models with knowledge distillation. In *NIPS*, 742–751.

Chen, S.; Zhang, C.; and Dong, M. 2018. Coupled end-to-end transfer learning with generalized fisher information. In *CVPR*, 4329–4338.

Dai, J.; Li, Y.; He, K.; and Sun, J. 2016. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, 379–387.

Dauphin, Y. N.; Pascanu, R.; Gulcehre, C.; Cho, K.; Ganguli, S.; and Bengio, Y. 2014. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in neural information processing systems*, 2933–2941.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2018. Arcface: Additive angular margin loss for deep face recognition. *arXiv preprint arXiv:1801.07698*.

Du, Y.; Czarnecki, W. M.; Jayakumar, S. M.; Pascanu, R.; and Lakshminarayanan, B. 2018. Adapting auxiliary losses using gradient similarity. *arXiv preprint arXiv:1812.02224*.

Girshick, R.; Donahue, J.; Darrell, T.; and Malik, J. 2016. Region-based convolutional networks for accurate object detection and segmentation. *IEEE transactions on pattern analysis and machine intelligence* 38(1):142–158.

Girshick, R. 2015. Fast r-cnn. In *ICCV*, 1440–1448.

Glorot, X., and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 249–256.

Han, S.; Mao, H.; and Dally, W. J. 2015. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*.

Hao, Z.; Liu, Y.; Qin, H.; Yan, J.; Li, X.; and Hu, X. 2017. Scale-aware face detection. In *CVPR*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Howard, A. G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; and Adam, H. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Hu, J.; Shen, L.; and Sun, G. 2017. Squeeze-and-excitation networks. *arXiv preprint arXiv:1709.01507* 7.

Iandola, F. N.; Han, S.; Moskewicz, M. W.; Ashraf, K.; Dally, W. J.; and Keutzer, K. 2016. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*.

Ioffe, S., and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Jain, V., and Learned-Miller, E. 2010. Fddb: A benchmark for face detection in unconstrained settings. Technical Report UM-CS-2010-009, University of Massachusetts, Amherst.

Jiang, H., and Learned-Miller, E. 2017. Face detection with the faster r-cnn. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, 650–657. IEEE.

Kim, K.-H.; Hong, S.; Roh, B.; Cheon, Y.; and Park, M. 2016. Pvanet: deep but lightweight neural networks for real-time object detection. *arXiv preprint arXiv:1608.08021*.

Li, Q.; Jin, S.; and Yan, J. 2017. Mimicking very efficient network for object detection. In *CVPR*.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2018. Focal loss for dense object detection. *IEEE transactions on pattern analysis and machine intelligence*.

Lipton, Z. C. 2016. Stuck in a what? adventures in weight space. *arXiv preprint arXiv:1602.07320*.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*, 21–37. Springer.

Liu, Y.; Li, H.; Yan, J.; Wei, F.; Wang, X.; and Tang, X. 2017. Recurrent scale approximation for object detection in cnn. In *ICCV*, volume 5.

Luo, J.-H.; Wu, J.; and Lin, W. 2017. Thinet: A filter level pruning method for deep neural network compression. *arXiv preprint arXiv:1707.06342*.

Ma, N.; Zhang, X.; Zheng, H.-T.; and Sun, J. 2018. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *ECCV*.

Molchanov, P.; Tyree, S.; Karras, T.; Aila, T.; and Kautz, J. 2016. Pruning convolutional neural networks for resource efficient inference. *arXiv preprint arXiv:1611.06440*.

Najibi, M.; Samangouei, P.; Chellappa, R.; and Davis, L. S. 2017. Ssh: Single stage headless face detector. In *ICCV*.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *CVPR*, 779–788.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 91–99.

Romero, A.; Ballas, N.; Kahou, S. E.; Chassang, A.; Gatta, C.; and Bengio, Y. 2014. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *IJCV* 115(3):211–252.

Sagun, L.; Evci, U.; Guney, V. U.; Dauphin, Y.; and Bottou, L. 2017. Empirical analysis of the hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*.

Sagun, L.; Bottou, L.; and LeCun, Y. 2016. Eigenvalues of the hessian in deep learning: Singularity and beyond. *arXiv preprint arXiv:1611.07476*.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 4510–4520.

Shi, X.; Shan, S.; Kan, M.; Wu, S.; and Chen, X. 2018. Real-time rotation-invariant face detection with progressive calibration networks. In *CVPR*.

Singh, B., and Davis, L. S. 2018. An analysis of scale invariance in object detection snip. In *CVPR*.

Song, G.; Liu, Y.; Jiang, M.; Wang, Y.; Yan, J.; and Leng, B. 2018. Beyond trade-off: Accelerate fcn-based face detector with higher accuracy. In *CVPR*.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. In *CVPR*.

Tang, X.; Du, D. K.; He, Z.; and Liu, J. 2018. Pyramidbox: A context-assisted single shot face detector. *arXiv preprint arXiv:1803.07737*.

Tian, Y.; Jiang, T.; Gong, Q.; and Morcos, A. 2019. Luck matters: Understanding training dynamics of deep relu networks. *arXiv preprint arXiv:1905.13405*.

Wan, W.; Zhong, Y.; Li, T.; and Chen, J. 2018. Rethinking feature distribution for loss functions in image classification. In *CVPR*, 9117–9126.

Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; and Tang, X. 2017a. Residual attention network for image classification. In *CVPR*.

Wang, H.; Li, Z.; Ji, X.; and Wang, Y. 2017b. Face r-cnn. *arXiv preprint arXiv:1706.01061*.

Wang, Y.; Ji, X.; Zhou, Z.; Wang, H.; and Li, Z. 2017c. Detecting faces using region-based fully convolutional networks. *arXiv preprint arXiv:1709.05256*.

Wang, F.; Chen, L.; Li, C.; Huang, S.; Chen, Y.; Qian, C.; and Change Loy, C. 2018. The devil of face recognition is in the noise. In *ECCV*.

Wei, Y.; Pan, X.; Qin, H.; Ouyang, W.; and Yan, J. 2018. Quantization mimic: Towards very tiny cnn for object detection. In *ECCV*.

Yang, B.; Yan, J.; Lei, Z.; and Li, S. Z. 2015. Fine-grained evaluation on face detection in the wild. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 1, 1–7. IEEE.

Yang, S.; Xiong, Y.; Loy, C. C.; and Tang, X. 2017. Face detection through scale-friendly deep convolutional networks. *arXiv preprint arXiv:1706.02863*.

Yim, J.; Joo, D.; Bae, J.; and Kim, J. 2017. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *CVPR*, volume 2.

Zagoruyko, S., and Komodakis, N. 2016. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*.

Zhang, X.; Zhou, X.; Lin, M.; and Sun, J. Shufflenet: an extremely efficient convolutional neural network for mobile devices (2017). arxiv preprint. *arXiv preprint arXiv:1707.01083*.

Zhang, K.; Zhang, Z.; Wang, H.; Li, Z.; Qiao, Y.; and Liu, W. 2017a. Detecting faces using inside cascaded contextual cnn. In *ICCV*, 3171–3179.

Zhang, S.; Zhu, X.; Lei, Z.; Shi, H.; Wang, X.; and Li, S. Z. 2017b. Faceboxes: A cpu real-time face detector with high accuracy. In *IJCB)*, 1–9. IEEE.

Zhang, S.; Zhu, X.; Lei, Z.; Shi, H.; Wang, X.; and Li, S. Z. 2017c. Sˆ 3fd: Single shot scale-invariant face detector. In *ICCV*, 192–201. IEEE.

Zhang, X.; Li, Z.; Change Loy, C.; and Lin, D. 2017d. Polynet: A pursuit of structural diversity in very deep networks. In *CVPR*, 718–726.

Zhang, Y.; Xiang, T.; Hospedales, T. M.; and Lu, H. 2018. Deep mutual learning. In *CVPR*.

Zhu, X., and Ramanan, D. 2012. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, 2879–2886. IEEE.