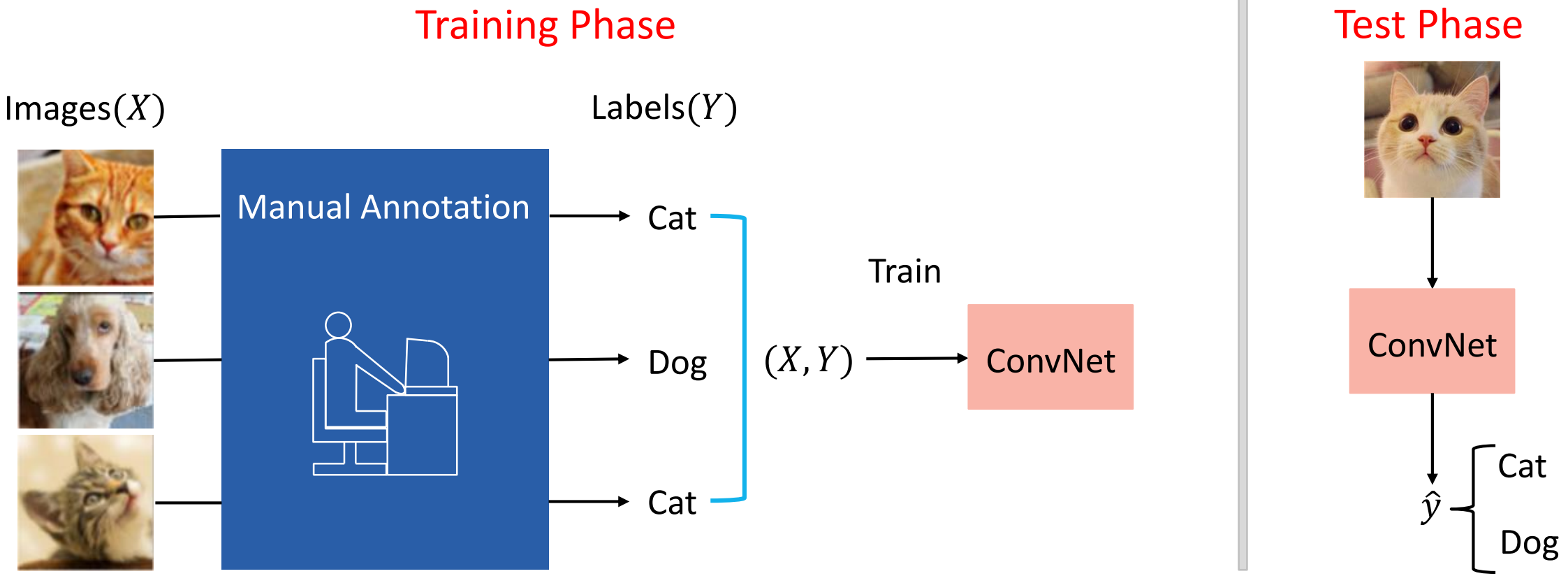
南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Deep Discriminative CNN with Temporal Ensembling for Ambiguously-Labeled Image Classification

Yao Yao[1], Chen Gong[1], Jiehui Deng[1], Xiuhua Chen[1], Jianxin Wu[3] and Jian Yang[1,2*]

AAAI 2020

Supervised Learning Workflow



- The groundtruth of each training image is unique and should be available in training phase.
- Unfortunately, the images may lack clear labels and manually labeling them will incur unaffordable monetary or time cost.

Partial Label Learning

Newsletter

Barcelona news: Lionel Messi and Luis Suarez hold private talks over five player concerns

Crowdsourcing

Annotator 1: Horse Annotator 2: Donkey
Annotator 3: Mule

- Generally, it is a multiple classification problem.
- The images are associated with multiple candidate labels, and only one of them is valid

Formal Definition

- Input space: $\mathcal{X} \in \mathbb{R}^d$ is d-dimensional.
- Output space: $\mathcal{Y} = \{1, 2, \ldots, c\}$ includes c classes.
- Training set: $\mathcal{D} = \{(\mathbf{x}_i, S_i) | 1 \leq i \leq n\}$ has n ambiguously labeled examples.

False positive labels

- Candidate label set: $S_i = A_i \cup \{\mathrm{y}_i\}$

Groundtruth label

- Target: train a classifier $f : \mathcal{X} \to \mathcal{Y}$ from the training set $\mathcal{D}$ so that the correct predictions can be made on test examples.

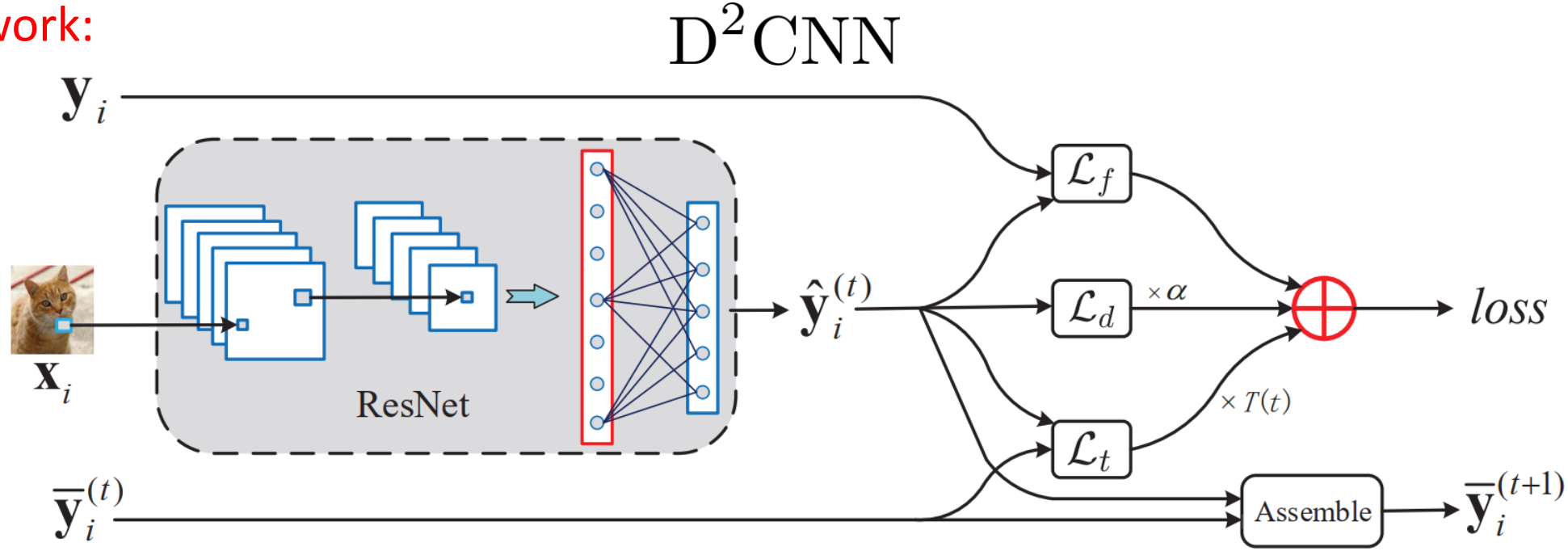
Motivations

- To solve this problem, the common strategy is to disambiguate the set of candidates of each example.
- However, existing methods are usually short of representation ability and discrimination ability.

	Representation ability	Discrimination ability
Reasons	Shallow learning frameworks	Imperfect disambiguation techniques.
Contributions	Employ deep convolutional neural network	Employ entropy minimization regularizer and temporal ensembling technique

Methods

Network:



Loss function:

Discrimination term

$$Loss = \mathcal{L}_f(\mathbf{Y}, \hat{\mathbf{Y}}) + \alpha\mathcal{L}_d(\hat{\mathbf{Y}}) + T(t) \cdot \mathcal{L}_t(\bar{\mathbf{Y}}, \hat{\mathbf{Y}})$$

Fidelity term

Temporal Ensembling Term

Fidelity Term

The predicted probability of the labels in non-candidate label set should be zero.

$$\mathcal{L}_f(\mathbf{Y}, \hat{\mathbf{Y}}) = -\frac{1}{n}\sum_{i=1}^{n}(\mathbf{1} - \mathbf{y}_i)^\top \log(\mathbf{1} - \hat{\mathbf{y}}_i)$$

- Label matrix: $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n] \in \{0, 1\}^{n \times c}$
  - $\mathrm{y}_{ij} = 1$: if the *j*-th label is a candidate label of the image $\mathbf{x}_i$
  - $\mathrm{y}_{ij} = 0$: otherwise
- Predicted matrix: $\hat{\mathbf{Y}} = [\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_n] \in [0, 1]^{n \times c}$
  - $\hat{\mathrm{y}}_{ij}$: the probability of the image $\mathbf{x}_i$ belongs to the *j*-th class
  - $\hat{\mathrm{y}}_{ij} \geq 0$ and $\sum_{j=1}^{c} \hat{\mathrm{y}}_{ij} = 1$

Discrimination Term

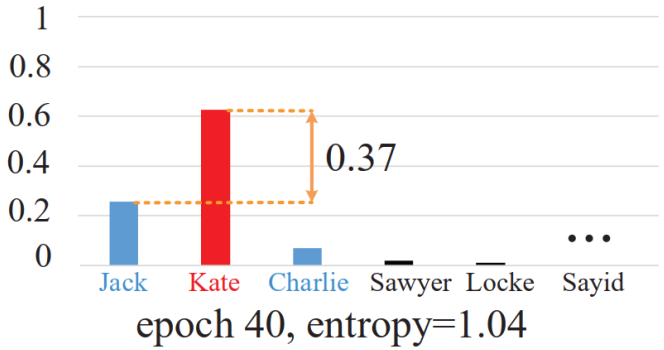
Minimizing entropy makes the potential groundtruth label become prominent among all labels.

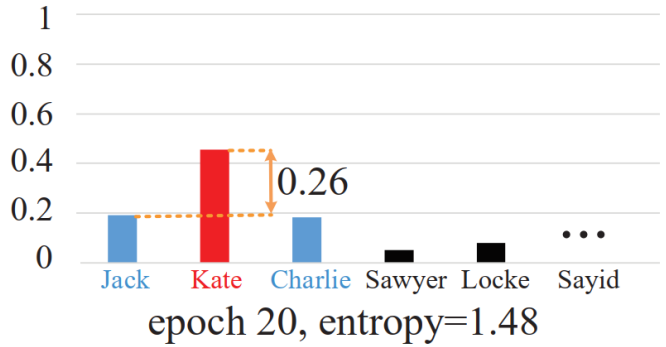
$$\mathcal{L}_d(\hat{\mathbf{Y}}) = -\frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{y}}_i^{\top}\log\hat{\mathbf{y}}_i$$

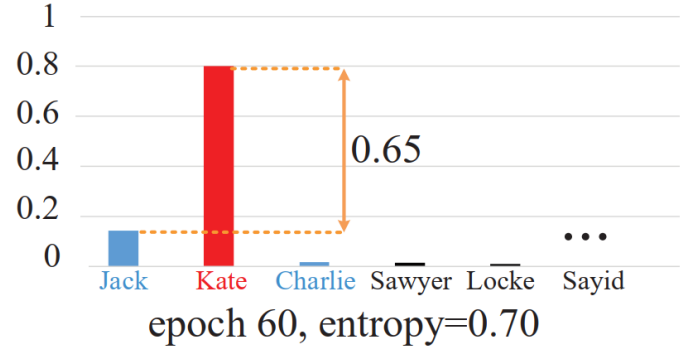
Entropy

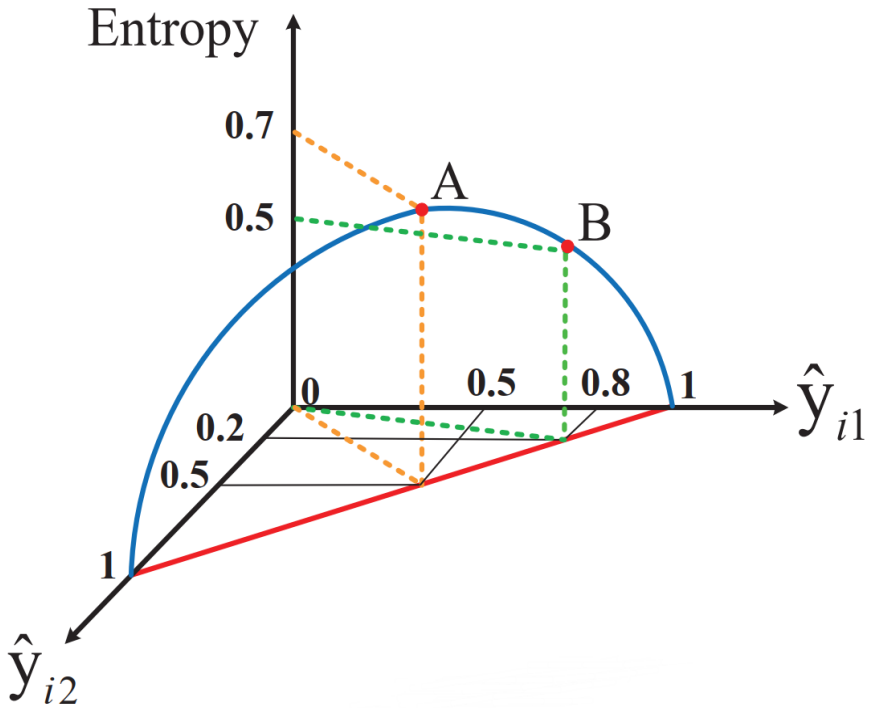
Input $\mathbf{x}_i$

S_i : {Jack, Kate, Charlie}









Temporal Ensembling Term

Assemble the model predictions of different epochs and regard them as the auxiliary supervision information for the next epoch.

$$\mathcal{L}_t(\bar{\mathbf{Y}}, \hat{\mathbf{Y}}) = -\frac{1}{n}\sum_{i=1}^{n} w_i \cdot \bar{\mathbf{y}}_i^{\top} \log \hat{\mathbf{y}}_i$$

$$\mathbf{S}^{(t)} = \gamma \mathbf{S}^{(t-1)} + (1-\gamma)\hat{\mathbf{Y}}^{(t)}$$

Assembled predictions at t-th epoch

$$\bar{\mathbf{Y}}^{(t+1)} = \mathbf{S}^{(t)} / (1-\gamma^t)$$

Training target of (t+1)-th epoch

$$w_i = \begin{cases} 0 & m_i \leq 0 \\ {m_i}^2 & otherwise \end{cases}$$

$$m_i = \max_{y_j \in S_i} \overline{y}_{ij} - \max_{y_k \notin S_i} \overline{y}_{ik}$$

Formation of $T(t)$

- $T(t)$ is a time-dependent weighting function.
- Roughly speaking, $T(t)$ increases with the number of epochs.

$$T(t) = T_{max} \exp\left[-5(1-t)^2\right]$$

- T_{max} denotes the maximum value that $T(t)$ could reach.
- t increases linearly from zero to one during the rising phase.
- At the initial training phase, the network mainly learns from the original ambiguous labels.
- Then it gradually learns from the assembled predictions when then training process proceeds

Datasets

- Synthesized datasets:
  - Fashion MNIST
  - SVHN
- Real-world datasets:
  - Yahoo!News
  - Lost

Datasets	# Images	# Classes	Avg # labels
FM-v1	70,000	10	2
FM-v3	70,000	10	4
SVHN-v1	99,289	10	2
SVHN-v3	99,289	10	4
Lost	1,122	16	2.23
Yahoo!News	14,322	38	1.44

Characteristics of the adopted datasets.

Experiments on Synthesized Datasets

	FM-v1	FM-v3	SVHN-v1	SVHN-v3
RegISL	-	-	-	-
SURE	-	-	-	-
WMCar-ICE	-	-	-	-
MCar	0.913 ± 0.003 •	0.825 ± 0.002 •	0.796 ± 0.004 •	0.545 ± 0.004 •
PLKNN	0.897 ± 0.004 •	0.848 ± 0.004 •	0.736 ± 0.003 •	0.636 ± 0.002 •
M3PL	0.884 ± 0.002 •	0.874 ± 0.004 •	0.827 ± 0.002 •	0.788 ± 0.003 •
IPAL	0.912 ± 0.005 •	0.905 ± 0.003 •	0.798 ± 0.003 •	0.777 ± 0.001 •
DCNN	0.902 ± 0.007 •	0.890 ± 0.008 •	0.922 ± 0.003 •	0.908 ± 0.007 •
D^2CNN	$\mathbf{0.936 \pm 0.002}$	$\mathbf{0.927 \pm 0.003}$	$\mathbf{0.937 \pm 0.003}$	$\mathbf{0.929 \pm 0.001}$

- D^2CNN achieves superior performance against other baselines on these synthesized datasets.
- The performances of all baselines decrease, when the number of the candidate labels increase.

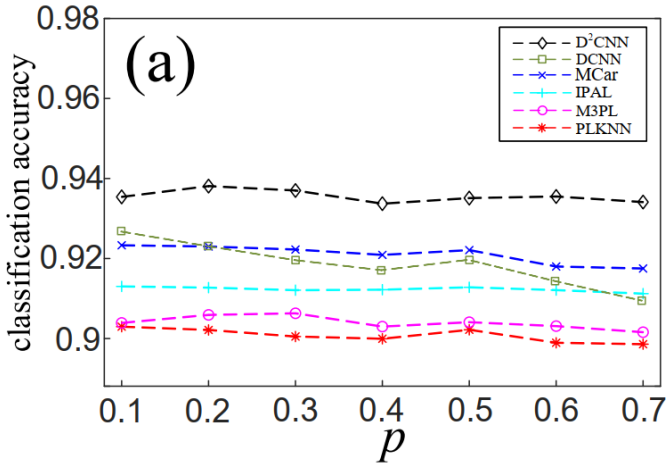
Experiments on Real-world Datasets

	Lost	Yahoo!News
RegISL	0.761 ± 0.037 •	0.598 ± 0.016 •
SURE	0.794 ± 0.037 •	0.729 ± 0.010 •
WMCar-ICE	0.795 ± 0.020 •	0.705 ± 0.010 •
MCar	0.743 ± 0.011 •	0.671 ± 0.010 •
PLKNN	0.651 ± 0.012 •	0.562 ± 0.017 •
M3PL	0.678 ± 0.032 •	0.613 ± 0.001 •
IPAL	0.790 ± 0.034 •	0.647 ± 0.017 •
DCNN	0.580 ± 0.031 •	0.740 ± 0.006 •
$\mathrm{D}^2\mathrm{CNN}$	$\mathbf{0.838 \pm 0.014}$	$\mathbf{0.833 \pm 0.009}$

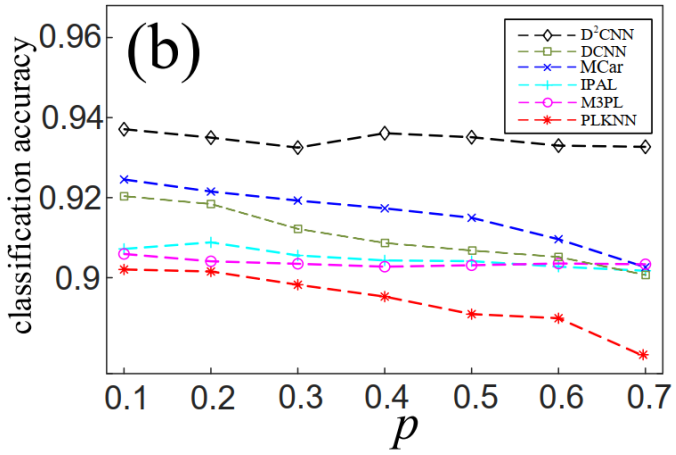
- The accuracy of $\mathrm{D}^2\mathrm{CNN}$ is higher than the second best method on *Lost* dataset by approximately 4%.
- On *Yahho!News* dataset, $\mathrm{D}^2\mathrm{CNN}$ significantly outperforms other baselines and leads the second best method with the margin of 9.3%.

Algorithm Analysis

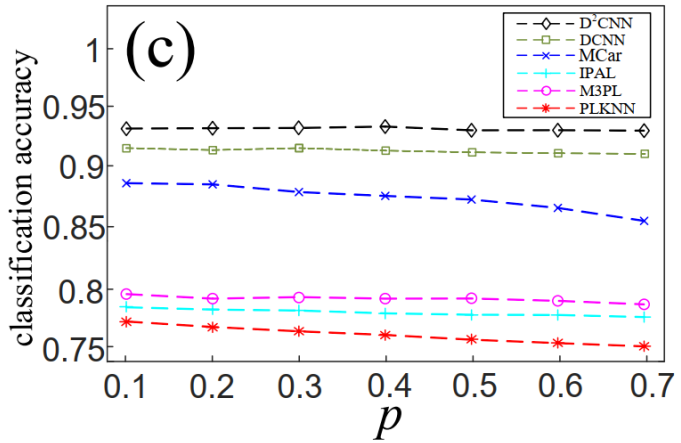
The robustness to proportion of ambiguously-labeled images



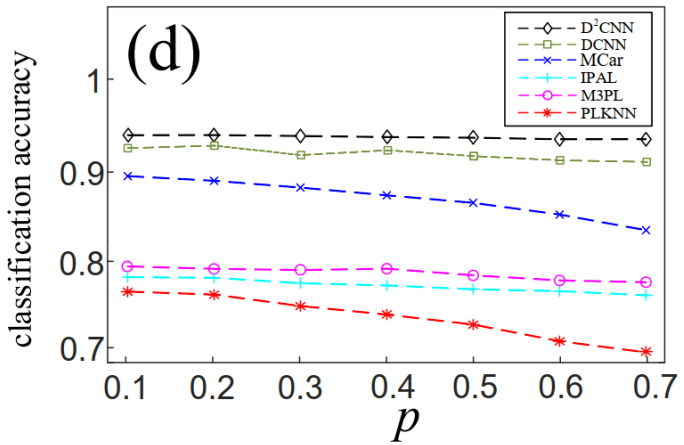
Fashion-Mnist dataset when $r = 1$



Fashion-Mnist dataset when $r = 3$

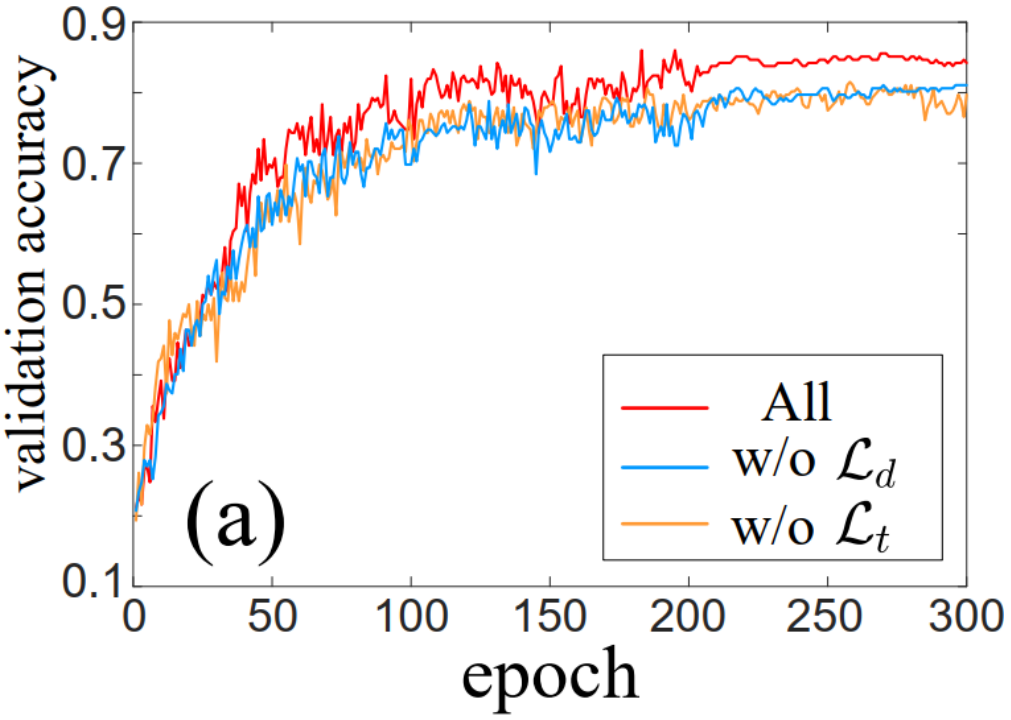


SVHN dataset when $r = 1$

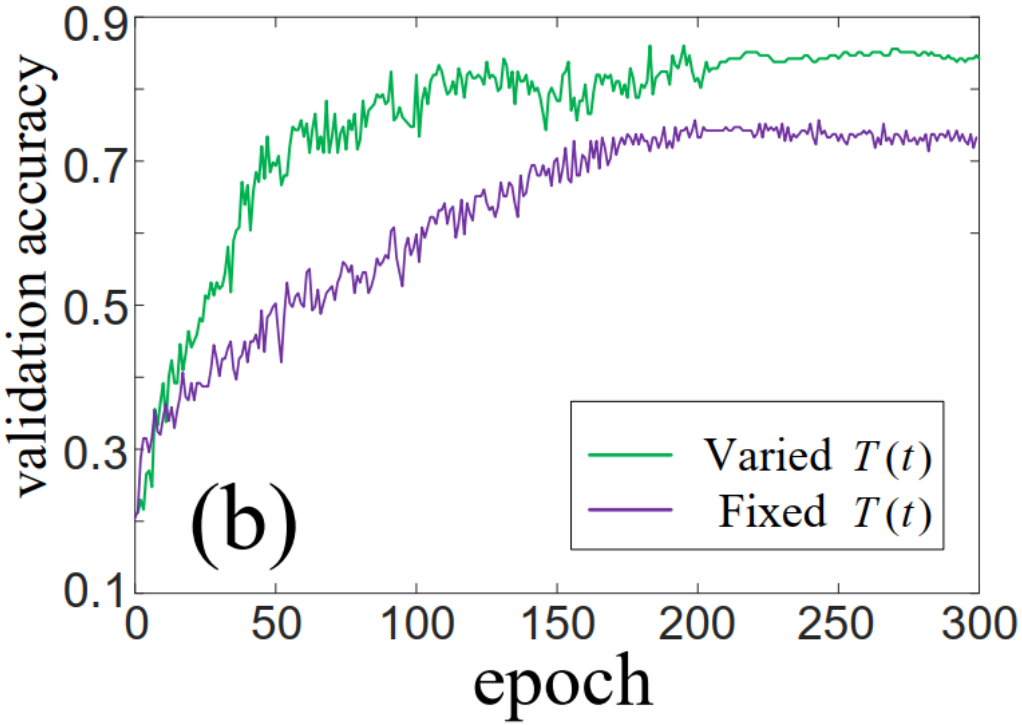


SVHN dataset when $r = 3$.

Algorithm Analysis



The effectiveness of $\mathcal{L}_d$ and $\mathcal{L}_t$



The effectiveness of varied $T(t)$

THANKS