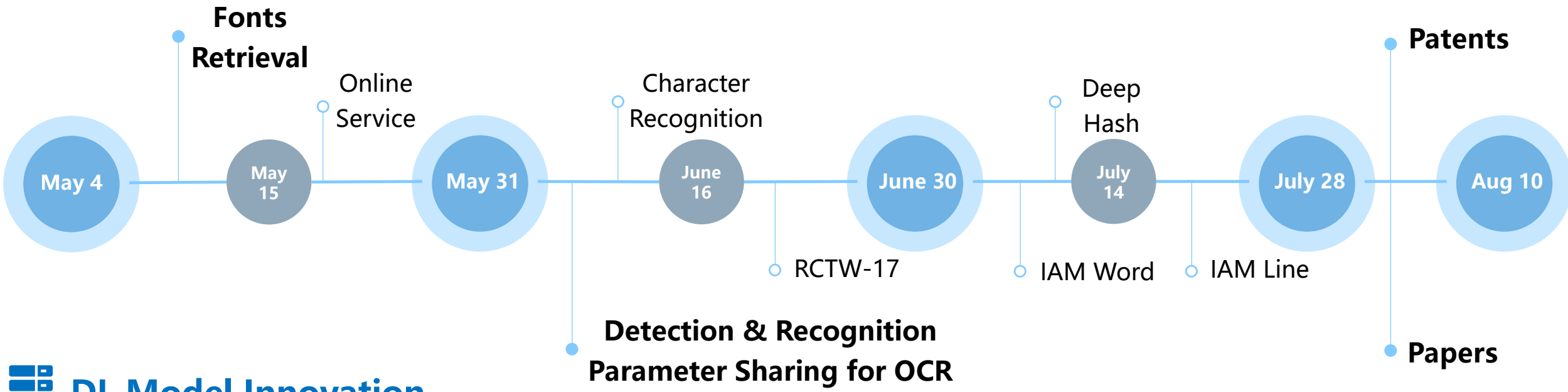
实习总结汇报

Bin Xu-Large Scale Learning-Deep Learning

Timeline

UED Intelligent Design



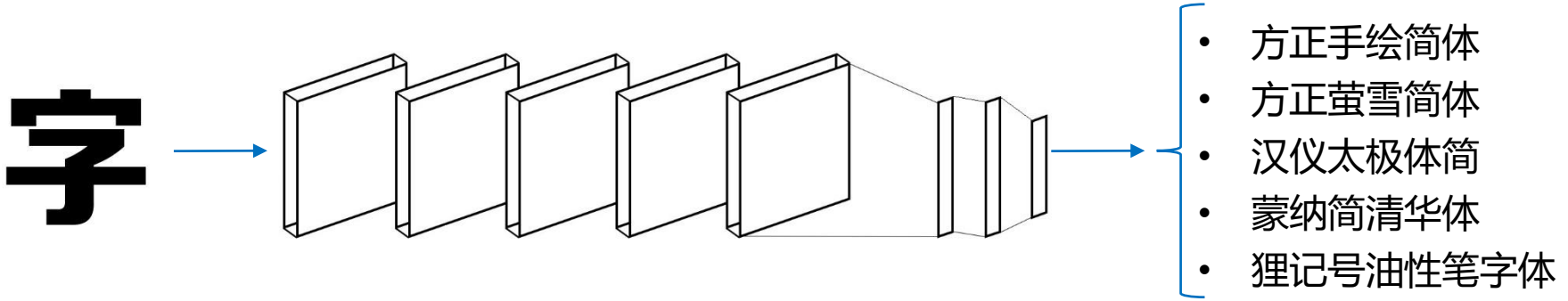
DL Model Innovation


UED

UED Project

Fonts Retrieval

- Model: we use the CNNs for the retrieval task.



- Dataset: we build two datasets according to the font files provided by International UED team.

Font Classes	Training Set	Testing Set	Validation set
506	20000k	20K	20K
1411	50000k	20K	20K

Examples:

UED Project

Training Model — we train and evaluate three classification models: ResNet, GoogleNet and VGG on 506 and 1411 font classes dataset, respectively.

- 506 font classes:

Model	val-top5	val-top1	test-top5	test-top1	Iteration
ResNet-18	0.96985	0.5235	0.96645	0.49785	210000
GoogleNet-4c	0.96015	0.4886	0.9578	0.4735	180000
VGG-16	0.97835	0.5161	0.97645	0.51085	185000

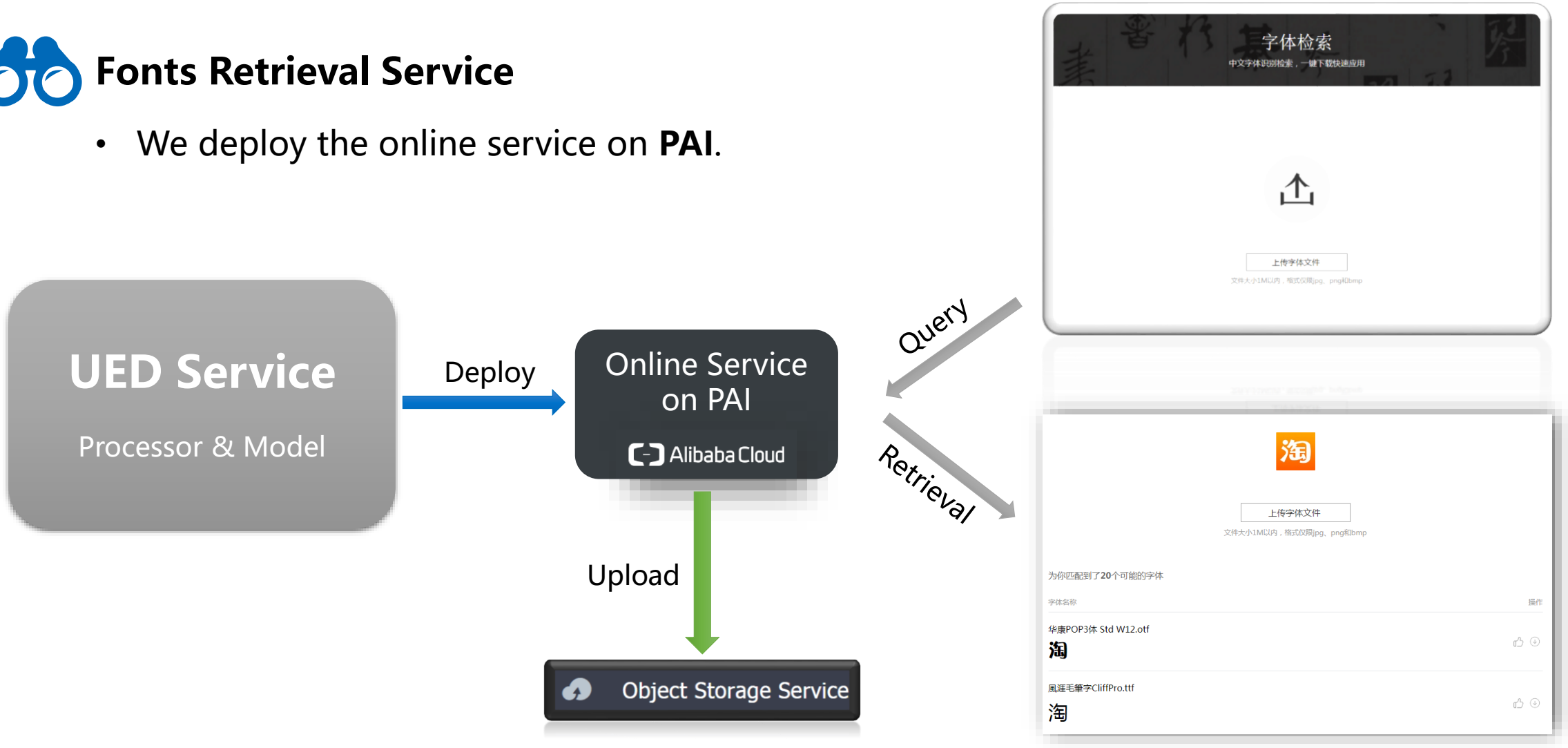
- 1411 font classes:

Model	test-top20	test-top10	test-top5	test-top1	Iteration
ResNet-18	0.98685	0.97055	0.90355	0.4923	205000
GoogleNet-4c	0.9873	0.9697	0.90205	0.48925	370000
VGG-16	0.98405	0.96275	0.88445	0.461	280000

UED Project

Fonts Retrieval Service

- We deploy the online service on **PAI**.



WiKi: http://gitlab.alibaba-inc.com/jason.xb/online_processor/wikis/home

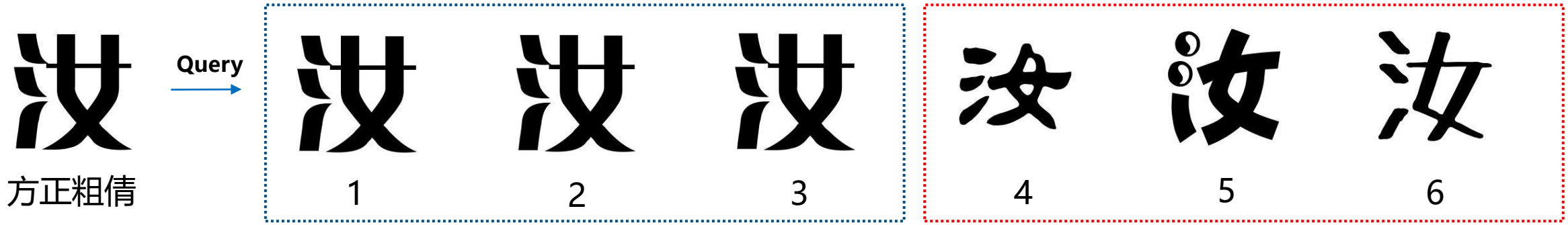
UED Project

Classification or Retrieval ?

- The classification models output a multiclass categorical probability distribution by the softmax function, and are trained by the Cross Entropy loss function.

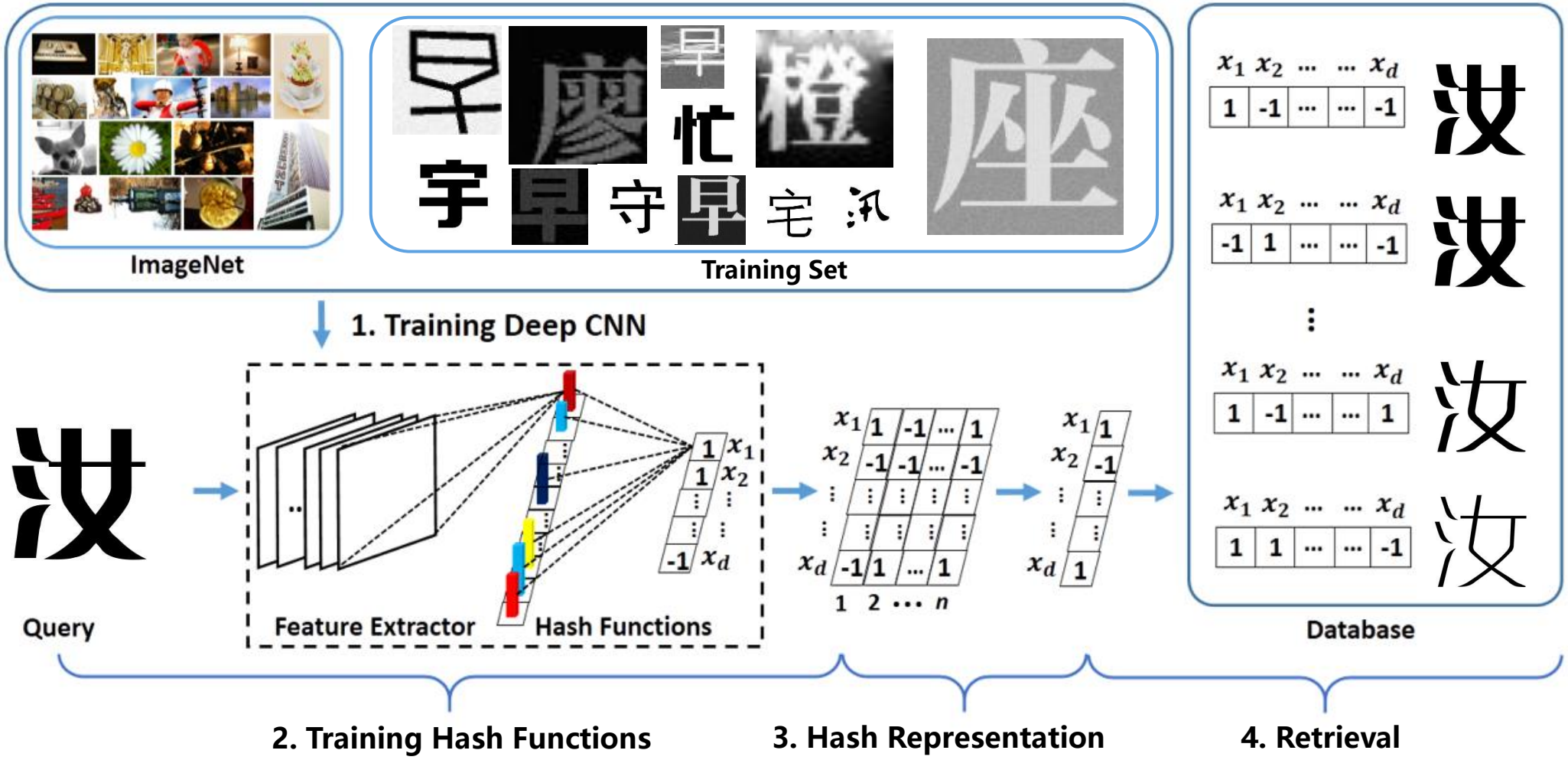
$$y_c=\varsigma(z)_c=\frac{e^{z_c}}{\sum_{d=1}^{C} e^{z_d}} \qquad for\ \ c=1...C \qquad \xi(T,Y)=\sum_{i=1}^{n}\xi(t_i,y_i)=-\sum_{i=1}^{n}\sum_{i=c}^{C} t_{ic}\cdot\log(y_{ic})$$

- We notice that the classification models learn the nonlinear functions from each font class separately, but ignore the inner connection between two similar fonts.



UED Project

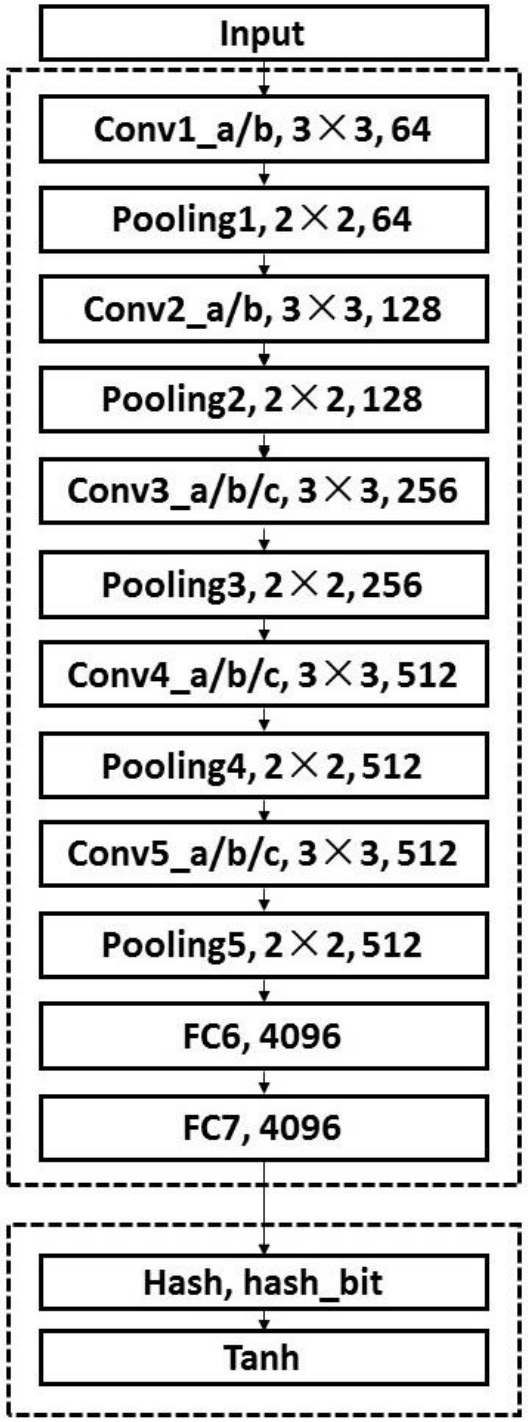
Deep Hash Method — we employ hash coding method in our retrieval model.



UED Project

Proposed Network

- We use the released VGG (Visual Geometry Group) model which is trained on ImageNet dataset for convenience.
- Previous studies show that the 4096-dimensional features of Layer7 perform better than many handcrafted features. In our network, the Layers1−7 are used as the **Feature Extractor Network**.
- We set Hash (full connected) layer and Tanh activation layer behind as the **Hash Network**.



UED Project

Deep Hash Representation

- Hash code of deep features:

$$\boldsymbol{p}_i = h_f(\boldsymbol{f}_i), \qquad h_f : \mathrm{F} \rightarrow (-1,+1)^S$$

$$\boldsymbol{c}_i = h(\boldsymbol{f}_i) = sgn(\boldsymbol{p}_i), \qquad h : \mathrm{F} \rightarrow \{-1,+1\}^S$$

- Hamming distance between two codes:

$$d\left(\boldsymbol{c}_i, \boldsymbol{c}_j\right) = \frac{S - \boldsymbol{c}_i^T \boldsymbol{c}_j}{2}$$

$$\boldsymbol{c}_1 = (1,1,-1) \quad \boldsymbol{c}_2 = (1,-1,1)$$

$$d\left(\boldsymbol{c}_1, \boldsymbol{c}_2\right) = \frac{3 - (1*1 + 1*(-1) + (-1)*1)}{2}$$

$$= \frac{3-(-1)}{2} = 2$$

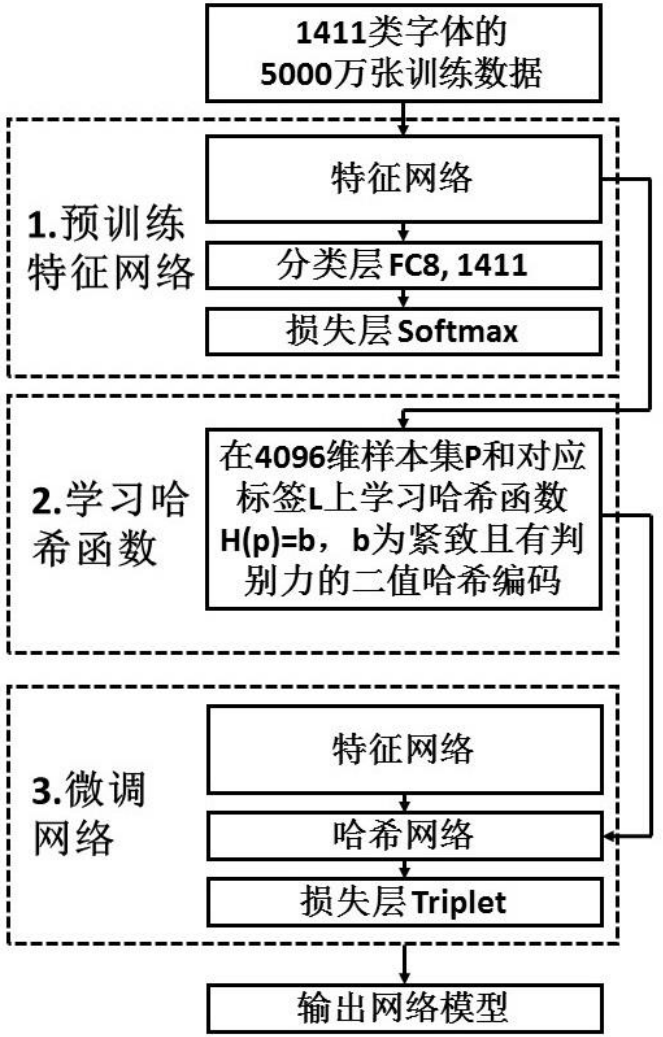
- Hash Representation of Font:

$$\bar{\boldsymbol{c}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{c}_i$$

UED Project

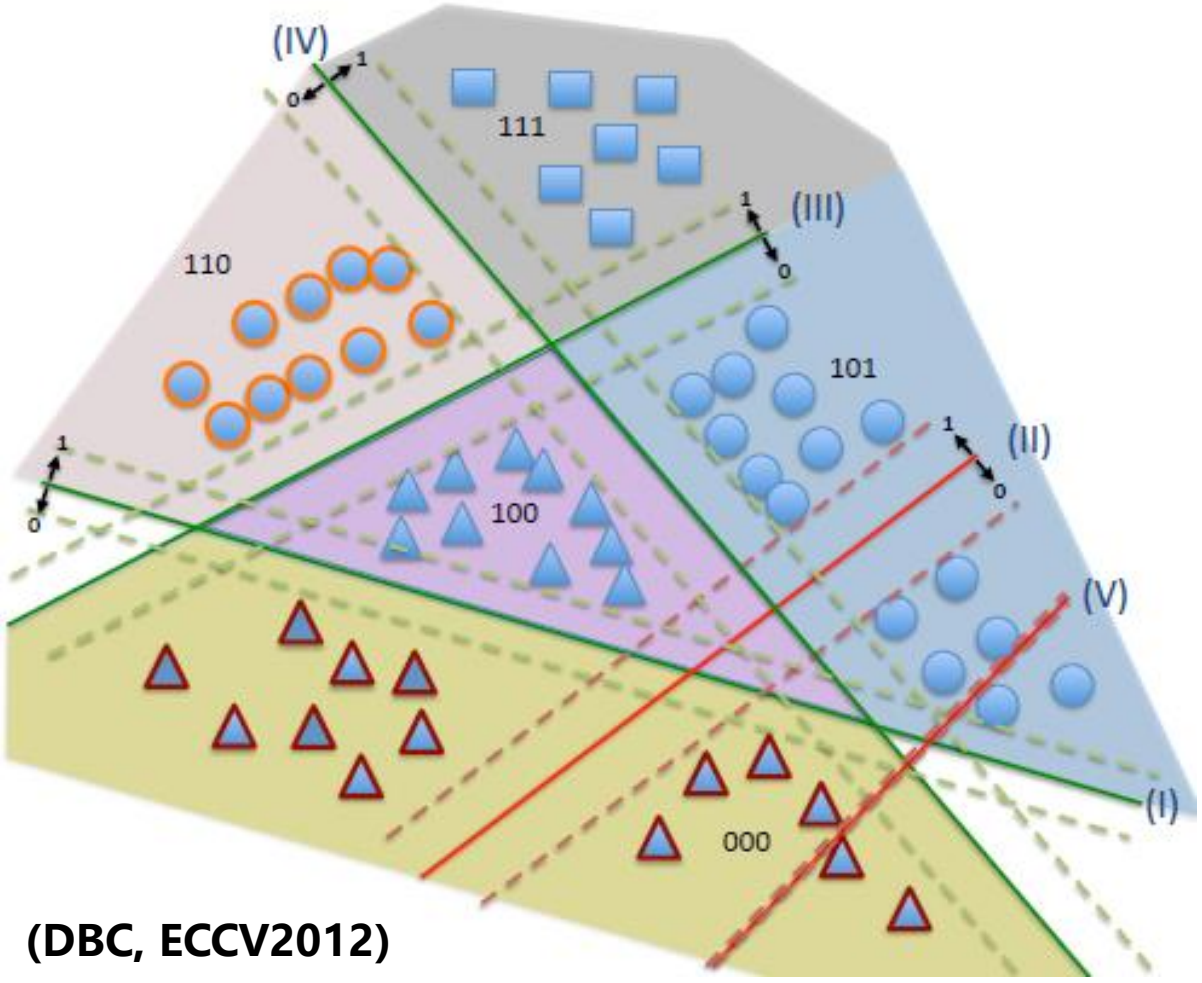
Training Procedure

- The training procedure of our network contains three steps:
  - pre-train the Feature Extractor Network
  - train the Hash Functions
  - fine-tune the Unified Network



UED Project

- Deep features: $\boldsymbol{P} = [\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_N] \in \mathbb{R}^{T \times N}$
- Hash functions: $\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_S] \in \mathbb{R}^{T \times S}$
- Hash coding: $\boldsymbol{B} = sgn(\boldsymbol{W}^T \boldsymbol{P}) = [\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_N] \in \{+1, -1\}^{S \times N}$



(DBC, ECCV2012)

- To learn hash functions, **discrimination** constraint and **stability** constraint are taken into account.

Objective:

$$\min_{\boldsymbol{W}, \boldsymbol{B}} \frac{1}{2} \| \boldsymbol{W} \|^2 + \mu \sum_{i=1}^{S} \sum_{j=1}^{N} \xi_{ij} + g(\boldsymbol{B})$$

$$s.t. \quad \boldsymbol{B}_{ij} (\boldsymbol{w}_i^T \boldsymbol{p}_j) \geq 1 - \xi_{ij},$$

$$\xi_{ij} \geq 0,$$

$$\boldsymbol{B} = sgn(\boldsymbol{W}^T \boldsymbol{P})$$

$$g(\boldsymbol{B}) = \sum_{\substack{p=1, \\ \boldsymbol{b}_i, \boldsymbol{b}_j \in \boldsymbol{B}_p}}^{C} d(\boldsymbol{b}_i, \boldsymbol{b}_j) - \lambda \sum_{\substack{p=1, \\ \boldsymbol{b}_i \in \boldsymbol{B}_p}}^{C} \sum_{\substack{q=1, q \neq p, \\ \boldsymbol{b}_j \in \boldsymbol{B}_q}}^{C} d(\boldsymbol{b}_i, \boldsymbol{b}_j),$$

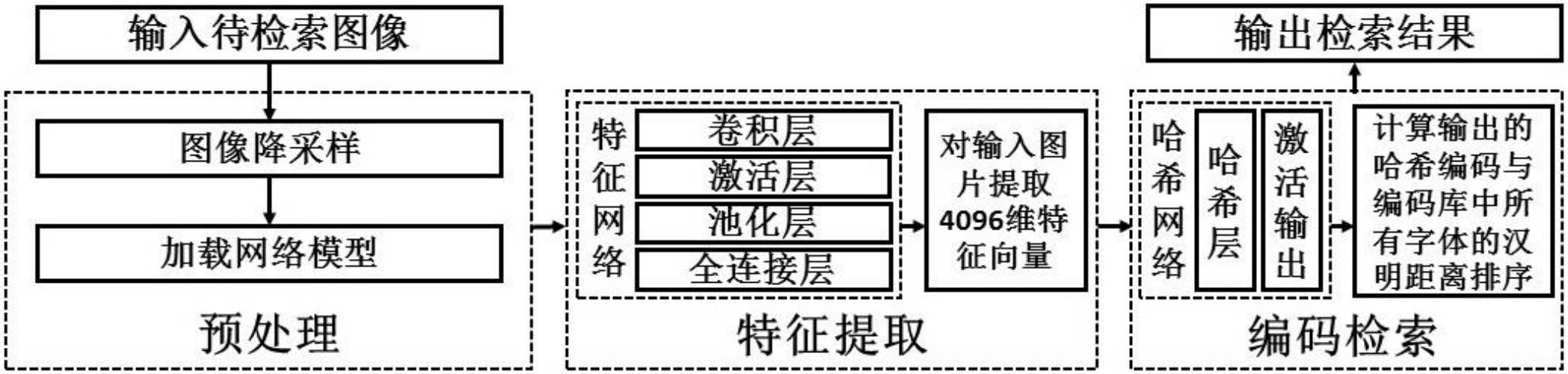
UED Project

Fine-tune

- We use the **Triplet Ranking Loss** similar to *Hamming Distance Metric Learning* (NIPS 2012), to fine-tune our network.
  - sample $\boldsymbol{p}$ is more similar to $\boldsymbol{p}_+$ than to $\boldsymbol{p}_-$ $\longrightarrow$ $d(\boldsymbol{p}, \boldsymbol{p}_+) + \zeta < d(\boldsymbol{p}, \boldsymbol{p}_-)$
  - loss function $L(\boldsymbol{p}, \boldsymbol{p}_+, \boldsymbol{p}_-) = \max\left(d(\boldsymbol{p}, \boldsymbol{p}_+) - d(\boldsymbol{p}, \boldsymbol{p}_-) + \zeta, 0\right)$
  - objective function $\min_{\boldsymbol{W}_h, \boldsymbol{W}} \sum_{i=1}^{C} \sum_{\substack{\boldsymbol{p}, \boldsymbol{p}_+ \in \boldsymbol{F}_i, \\ \boldsymbol{p} \neq \boldsymbol{p}_+}} \sum_{\substack{\boldsymbol{p}_- \in \boldsymbol{F}_j \\ j \neq i,}} \max\left(d(\boldsymbol{p}, \boldsymbol{p}_+) - d(\boldsymbol{p}, \boldsymbol{p}_-) + \zeta, 0\right)$

UED Project

Patent: 《一种基于深度卷积神经网络与哈希编码的字体检索方法和装置》



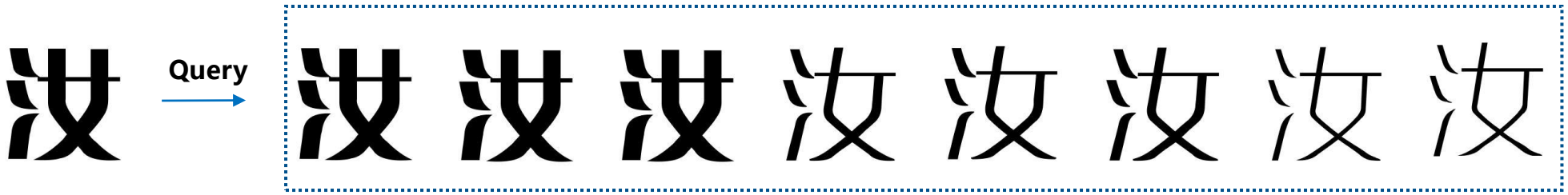
UED Project

Experiment

- We train and evaluate our retrieval model on 506 font classes dataset:

Model	test-top20	test-top10	test-top5	test-top1	Iteration
ResNet-18	-	-	0.92655	0.43925	105000
GoogleNet_4c	-	-	0.92235	0.43945	195000
VGG_16	-	-	0.92645	0.43085	185000
VGG+Hash	**0.977**	**0.965**	**0.929**	**0.474**	280000

- Retrieval results:



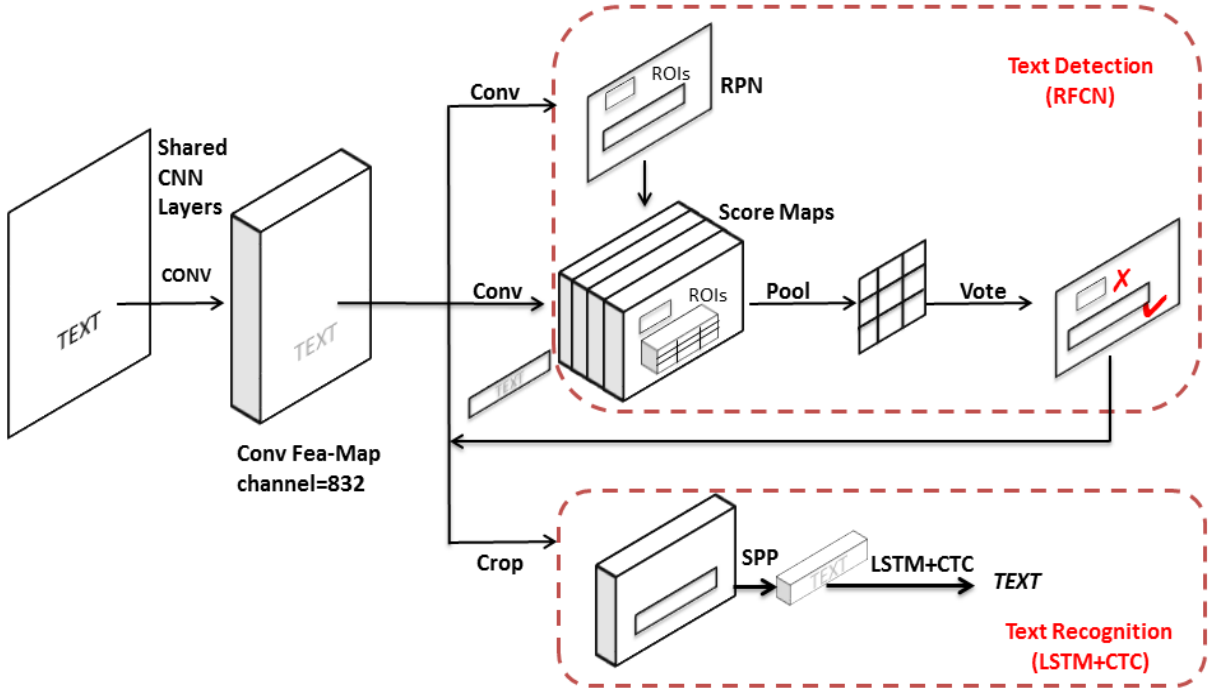

OCR

DL Model Innovation

Detection & Recognition Parameter Sharing for OCR — share the parameter for detection and that of recognition so that the time for recognition can be reduced

Two types of training approach:

- Separate training: detector and recognizer trained separately, use the same parameters for shared layers.
- End2end training: detection loss and recognition loss get summed as the whole loss to update the detector and recognizer, respectively.



Parameter Sharing – evaluation on public datasets

ICDAR2017 RCTW-17

- The dataset is built by **Huazhong University of Science and Technology**.
- 12000 images: 8,000 training, 4,000 testing

Parameter Sharing – separate training

RCTW-17

- We achieve comparable detection results on testing image (2,000 images divided from RCTW-17 train set).

F-measure Rank	Institute	Dataset	F-meausre	Precision	Recall
1	Peking University	Test set (4,229 images)	0.6610	74.38%	59.48%
6	Shanghai Jiao Tong University		0.5760	74.06%	47.12%
9	Sogou Inc.		0.4961	65.43%	39.95%
#	Organizing team (CVPR 2017)		0.5278	76.03%	40.43%
#	Ours	Train subset (2,000 images)	0.5421	75.65% (7219/9542)	42.24% (7219/17089)

- We also test our model on RCTW-17 Test set (**annotations not available**).

Ground truth

Our detection results

Train subset(2,000 images)

Our detection results

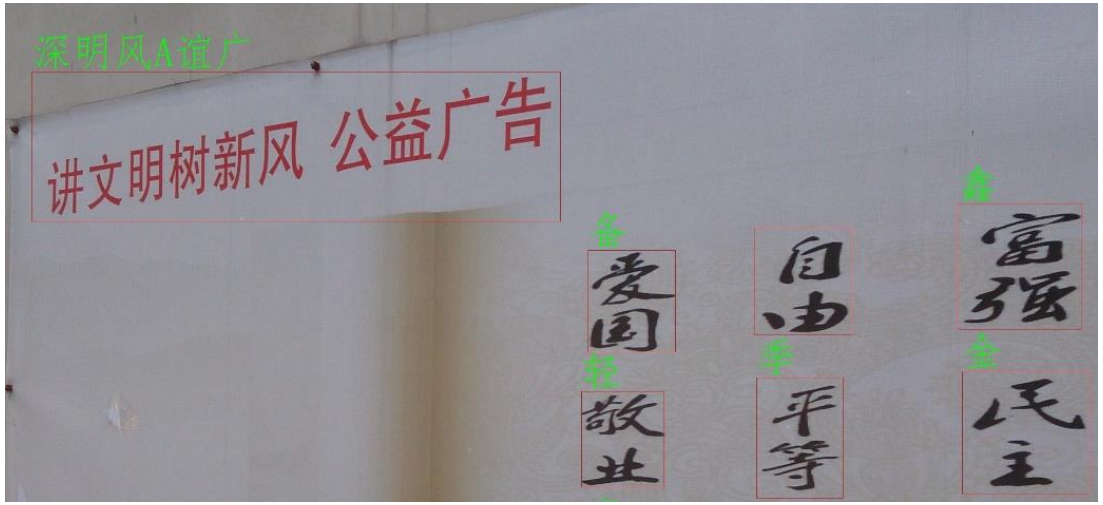
Test set(4,229 images)

Parameter Sharing – separate training

RCTW-17

- Our recognition model performs well on most line samples, but can not recognize characters in declining or vertical samples.

Good case



Bad case

- The detection and recognition model need to be improved, not finished yet.

Parameter Sharing – separate training

IAM (University of Bern)

- The IAM Handwriting Database contains forms of handwritten English text which can be used to train and test handwritten text recognizers and to perform writer identification and verification experiments.

Characteristics

The IAM Handwriting Database 3.0 is structured as follows:

- 657 writers contributed samples of their handwriting
- 1'539 pages of scanned text
- 5'685 isolated and labeled sentences
- 13'353 isolated and labeled text lines
- 115'320 isolated and labeled words

- *Large Writer Independent Text Line Recognition Task*

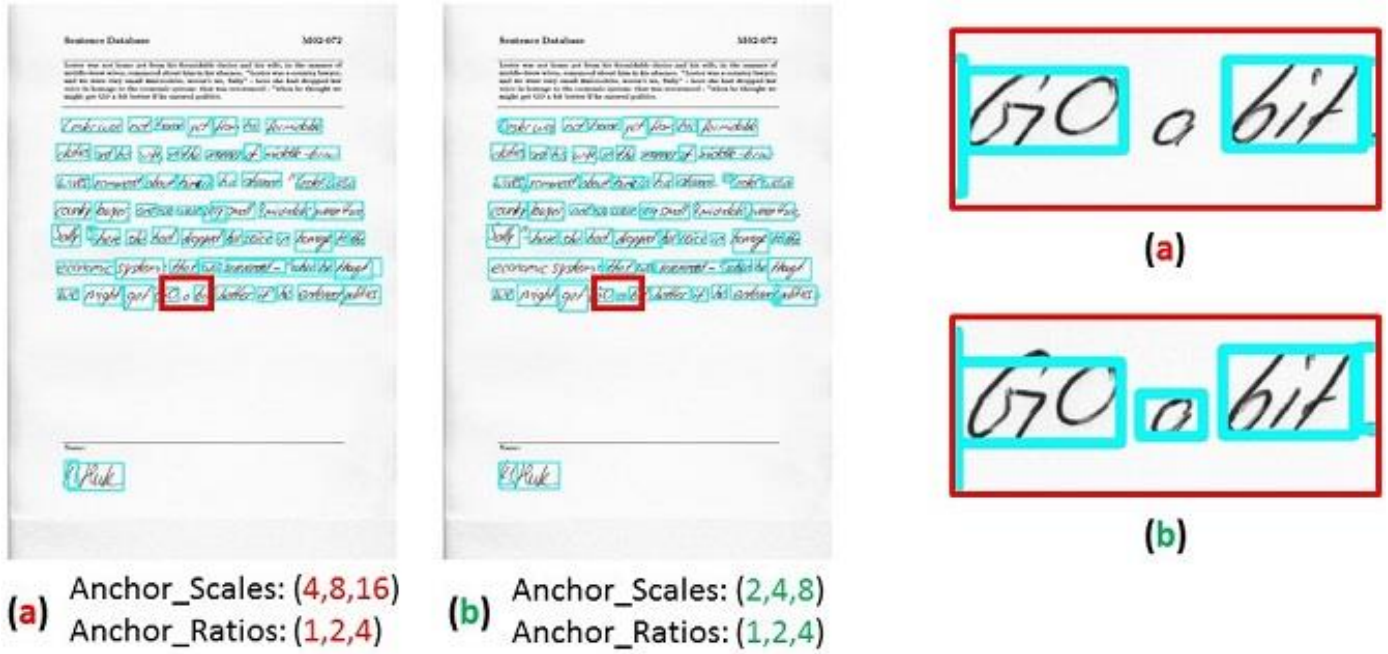
This task consists of a total number of 9'862 text lines.

Set Name	Number of Text Lines	Number of Writers
Train	6'161	283
Validation 1	900	46
Validation 2	940	43
Test	1'861	128
Total	9'862	500

Parameter Sharing – separate training

IAM — Word Level

- We adjust the anchor parameters of the Region Proposal Network (RPN) in R-FCN model for small word samples.



Our results

- We notice that there are a lot of punctuations in IAM, which are easily missed by the detection model, so the recall does not improve much.

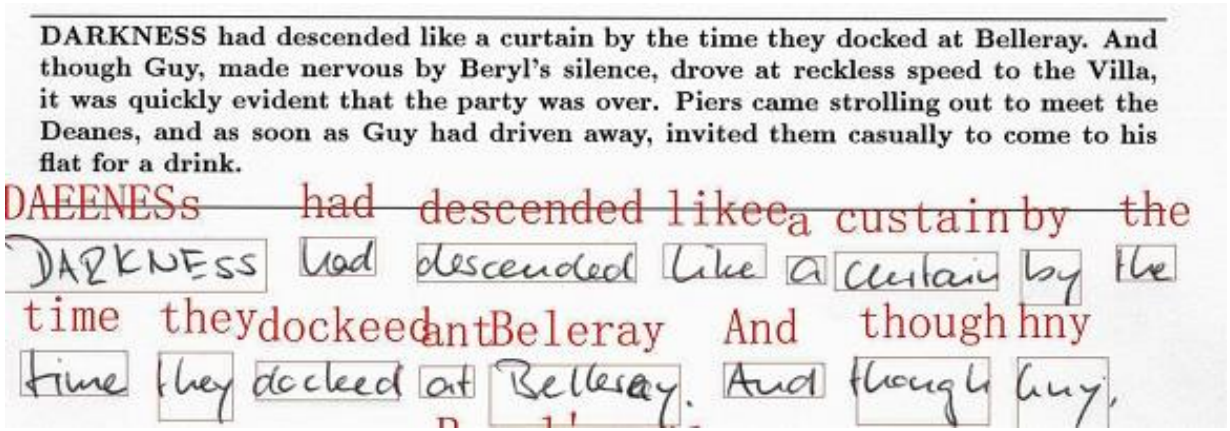
Parameter Sharing – separate training

IAM — Word Level

- We train our LSTM+CTC recognition model on the ground-truth bounding boxes, and obtain comparable results.

Model	Test (17,983 words)		Validation (19,392 words)	
	Word_acc%	Char_acc%	Word_acc%	Char_acc%
MDLSTM-RNNs, NIPS 2016	75.4	92.1	82.3	95.1
Ours (Iter 106,000)	71.38	88.75	75.75	87.82

- Due to the punctuation and biased detection results, our end-to-end word accuracy can only reach 51.84%.



Parameter Sharing – separate training

IAM — Line Level

- In order to recover the missed punctuations, we train our model on line level samples. We also adjust the anchor parameters and obtain accurate detection results.

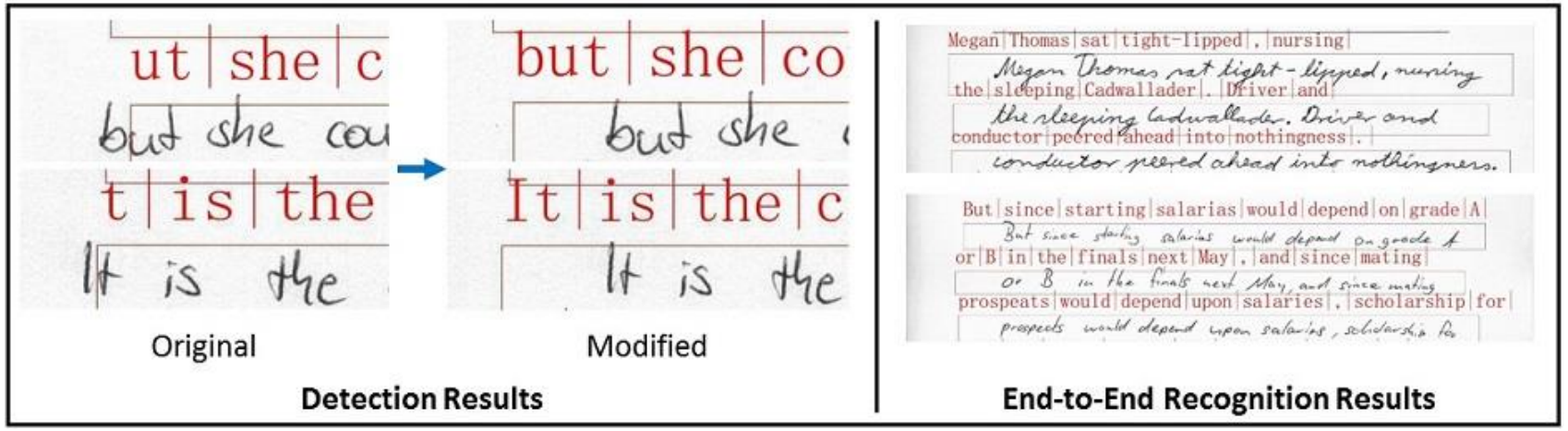
Threshold	ANCHOR_RATIO	ANCHOR_SCALES	Recall	Precision	F-Score	False Negative
IOU>0.5 Score>0.5	(2, 4, 8)	(8, 16, 32)	0.9881	0.9663	0.9771	22
	(4, 8, 16)	(4, 8, 16)	0.9919	0.9887	0.9903	15
	(4, 8, 16)	(8, 16, 32)	0.9940	0.9871	0.9906	11
IOU>0.5 Score>0.1	(4, 8, 16)	(8, 16, 32)	0.9967	0.9411	0.9681	6

- As to recognition model, we adjust the pooling size in the Conv3 and Conv4 to reduce of the loss of information during pooling ROI features.

Parameter Sharing – separate training

UED — Line Level

- When analyzing the errors, we notice that the original detection results might miss characters on the boundary of line samples, so we modify the detection results before inputting them into LSTM.



- Our end-to-end recognition word accuracy can reach **76.27%** (separate training), and our model can save over **30%** of computation time with single thread.
- We are conducting end-to-end model training on the IAM dataset, and we are preparing the paper.

WiKi: https://gitlab.Alibaba-inc.com/sensi.zq/Detection_Recognition_Parameter_Share/wikis/VerificationResult

Summary

UED Intelligent Design

- Fonts Classification CNN Models
- Fonts Retrieval Online Service
- Deep Hash Retrieval Model
- Patent and Paper of Fonts Retrieval Model

DL Model Innovation

- Evaluation Experiments on Public Datasets: RCTW-17、IAM
- Comparable End-to-End Detection & Recognition Results
- Paper of Detection & Recognition Parameter Sharing Model

Thank You

2017.8.10