A Multimodal Decision-level Fusion Model for E-Commerce Product Classification

Shuo Wang, Ye Bi, Zhongrui Fan

Ping An Technology

July 30th, 2020

Task 1 Introduction

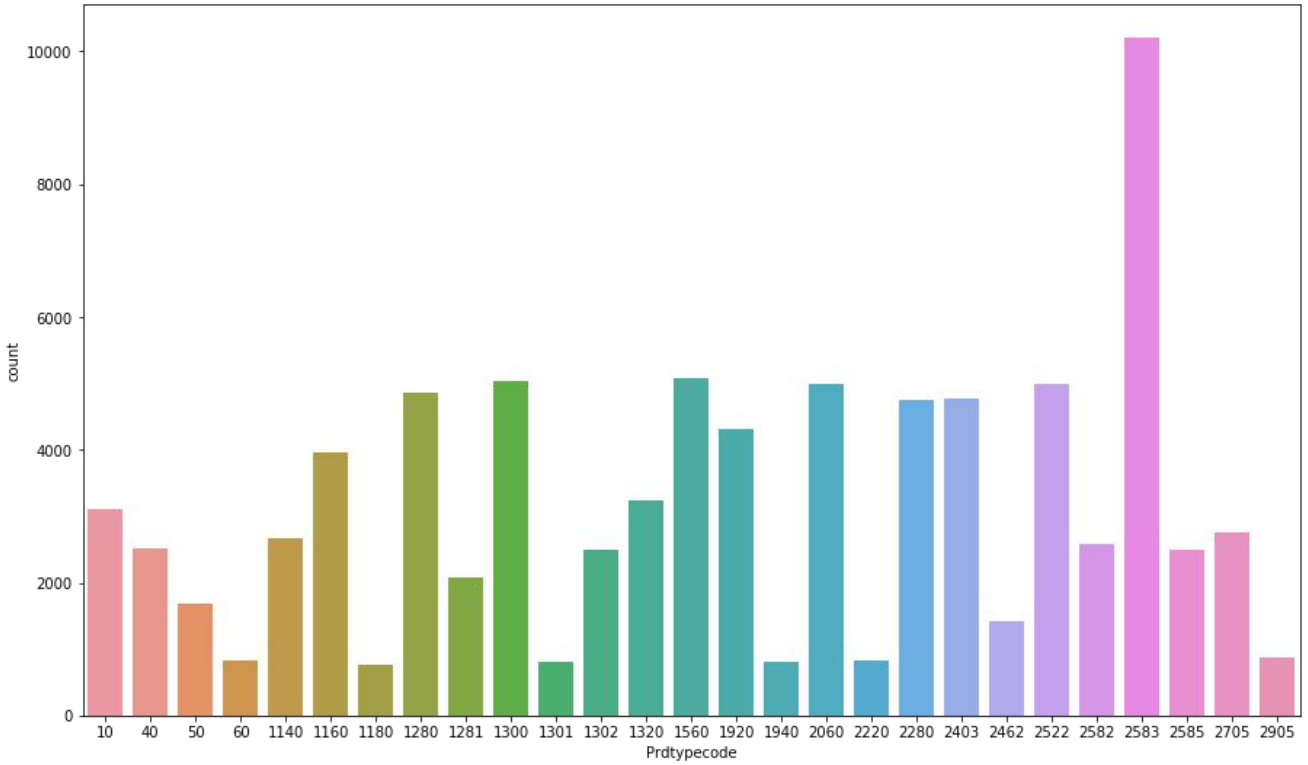
- Given a training set of products and their product type codes, predict the corresponding product type codes for an unseen held out test set of products.

- Multimodal classification task (Text and Image)

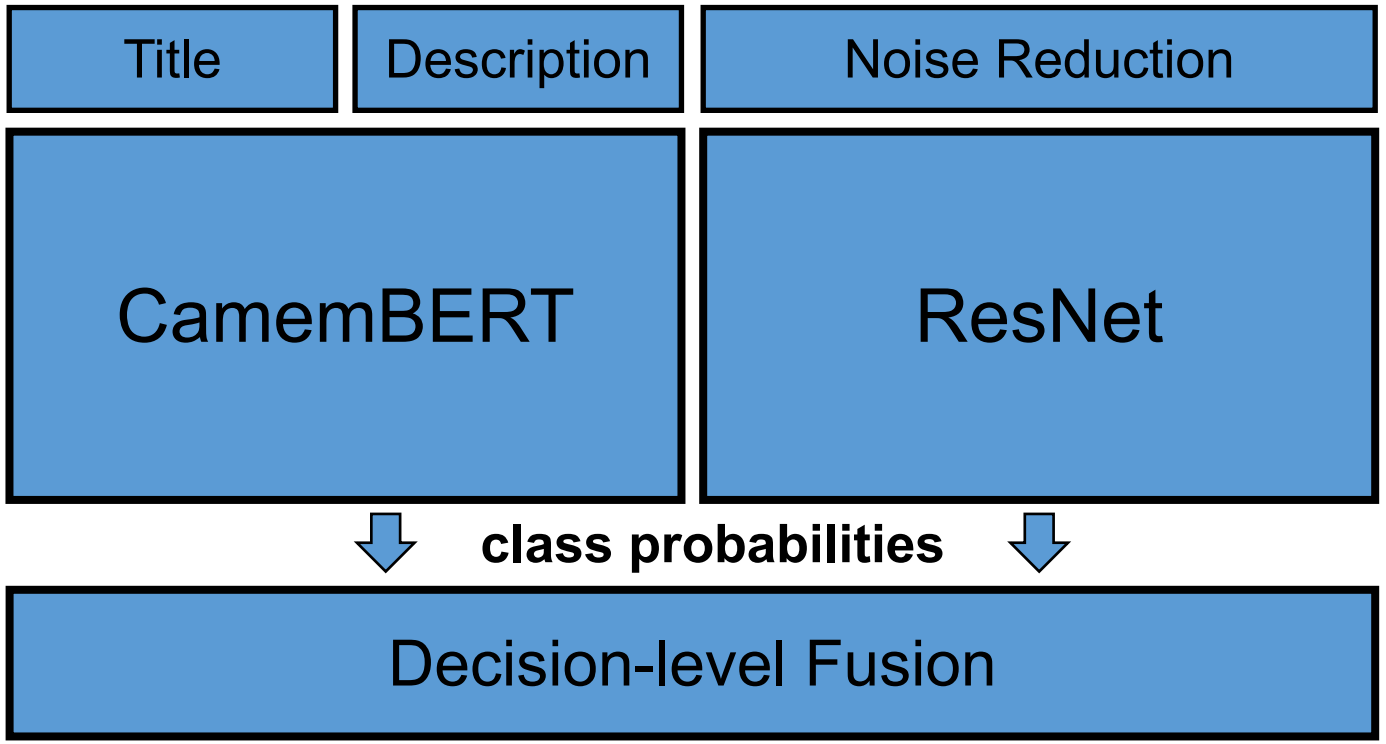
- Metric: Macro-F1

Dataset

- 84,916 samples for training, 937 samples for phase 1 testing and 8435 samples for phase 2 testing
- 27 product categories



Method Overview



Text classifier

- **Text preprocessing:** remove the excessive space and some HTML tags like *L&#39* and <*p*> from product title and description texts
- **CamemBERT:** the state-of-the-art pretrained French language model

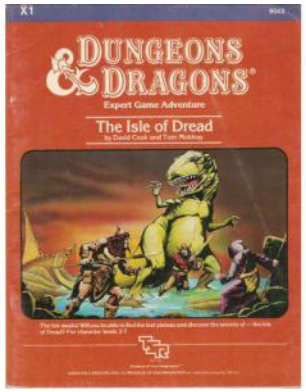
Image classifier

- Noise reduction

- Employ a fine-grained image recognition framework DCL[1]

[1] Yue Chen, Yalong Bai, Wei Zhang, and Tao Mei. 2019. Destruction and Construction Learning for Fine-Grained Image Recognition. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Noise Reduction



1. trained multiple ResNet50 image classifiers to compute the predicted product category probabilities for all the training samples in a cross-validation manner.
2. use the open source tool *cleanlab*[1] to find all noisy images
3. remove the top 10% noisy samples from the training set

1. https://github.com/cgnorthcutt/cleanlab

A fine-grained image recognition model

- **Backbone:** ResNet152

- **2-stage training**
  - backbone training (224 x 224)
  - DCL fine-tune (448 x 448)

https://github.com/JDAI-CV/DCL

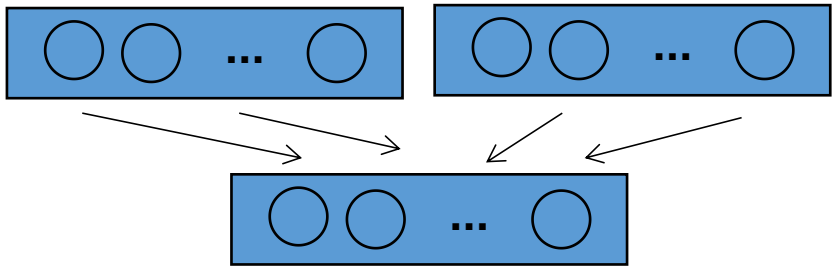
Fusion method

- feature-level fusion
  - fuse text and image feature vectors by concatenating, summation or attention mechanism

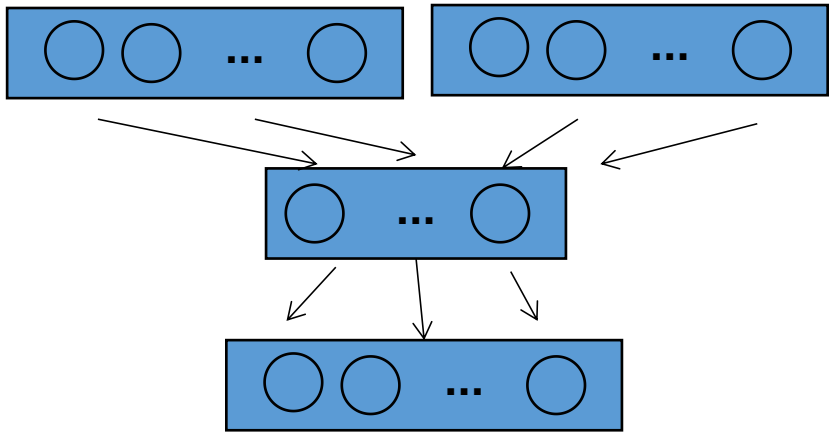
- decision-level fusion
  - a fusion strategy learned from the class probabilities predicted by each modal classifier.

Decision-level fusion

1-layer fusion network



2-layer fusion network



Ensemble Strategy

- Majority voting
- Candidate models
  - different configurations for the text classifier (CamemBERT-base and CamemBERT-large, different learning rates and batch size).
  - different configurations for the image classifier. (Whether to use the clean dataset after denoising, whether to use the DCL training method, models saved at the late training phase)
  - different configurations for the decision-level fusion. (1-layer and 2-layer fusion neural networks).

Experimental settings and Results

- train size : valid size = 9:1
- **text classifier**
  - learning rate: 3e-5, 5e-5
  - batch size: 64, 128
  - epochs: 40
  - optimizer: AdamW
  - warmup rate: 10%
- **image classifier**
  - learning rate: 0.01
  - optimizer: SGD
  - backbone training for 60 epochs and DCL fine-tuning for 20 epochs
- **decision-level fusion**
  - trained on the validation set in 8-fold cross validation manner
  - hidden size: 6 for 2-layer fusion network
  - learning rate: 0.01
  - epochs: 40
  - optimizer: Adam

Method	Phase 1 Test (%)	Phase 2 Test (%)
Uni-image Classifier	69.21	-
Uni-text Classifier	89.93	-
Feature-level Fusion	89.87	-
Decision-level Fusion	90.94	90.17
Ensemble	-	91.44

Thanks for your time.

If you have any questions, please contact us at wangshuotzjz@163.com