# Abstract

Image classification is often solved as a machine learning problem, where a classifier is first learned from training data, and class labels are then assigned to unlabeled testing data based on the outputs of the classifier. To train an image classifier with good generalization capability, conventional methods often require a large number of human labeled training images. However, a large number of well-labeled training images may not always be available. With the exponential growth of web meta-data, exploiting multimodal online sources via standard search engine has become a trend in visual recognition as it effectively alleviates the shortage of training data. However, the web meta-data such as text data is usually not as cooperative as expected due to its unstructured nature. In particular, how to represent and utilize the web text data from different modality for improved performance of web text-aided image classification is the focus of this thesis.

Our research aims at improving image classification by multimodal fusion of heterogeneous data from heterogeneous sources with the aid the online web resources search. Firstly, we attempt to learn the two modality of data (i.e., images and web text) separately and then combine the bimodal information at the decision level. In particular, low-level text modeling approaches such as class tag occurrence and bag-of-words (BoW) vectorization are utilized to learn text classifiers for decision level fusion. The advantage of decision level (semantic meaning level) fusion is that it is not affected by the vulnerable interaction of heterogeneous data. Meanwhile, we believe that the correlation information between image modality and web text modality is also very important and deserves further exploration. To address this, we also study on feature level multimodal fusion models. When it comes to feature level, how to learn robust representation that can better characterize multimodal data is the keystone. Recently, convolutional neural networks (CNN)-based image representation is dominant in several computer vision tasks. Likewise, deep neural networks (DNNs) greatly accelerate text representation learning methods in the area of natural language processing (NLP). Therefore, we also learn semantic high-level text features and train feature-level multimodal fusion models based on deep neural network training.

In this PhD thesis, I will elaborate web text-aided image classification problem from the aspect of semantic meaning level and feature level. This thesis is organized as follows. Chapter 1 introduces the motivation behind the web resources-aided image

classification. Chapter 2 reviews the related works to this field, including image representation learning, text representation learning and multimodal learning. Chapter 3 investigates semantic meaning-based data fusion for image classification. An adaptive combiner for two separate bimodal classifiers is developed at decision level. This adaptive fusion algorithm is inspired by the multisensory integration mechanism of human whose adaptability is achieved by reliability-dependent weighting of different sensory modalities. In Chapter 4, a novel text modeling namely the semantic matching neural network (SMNN) is proposed, which is quantified by cosine similarity measures between embedded text input and task-specific semantic filters. It is capable of learning semantic features from the associated text of web images. The SMNN text features have improved reliability and applicability, compared to the text features obtained from other methods. Then, the SMNN text features and convolutional neural network (CNN) visual features are jointly learned in a shared representation, which aims to capture the correlations between the two modalities. Improving upon task-specific filters for SMNN, Chapter 5 presents a novel semantic convolutional neural network (s-CNN) model for high-level text representation learning to encode semantic correlation based on task-generic semantic filters. Under CNN architecture, surplus filters in the network may lead to semantic overlaps and feature redundancy issue. To address this issue, the s-CNN Clustered (s-CNNC) models that uses filter clusters instead of individual filters is presented. Interacting with the image CNN models, the s-CNNC models can further boost image classification under a multi-modal framework (mm-CNN), which can be trained end-to-end. In Chapter 6, an adaptive attention network using web text is proposed to aide one-shot image classification. The s-CNN model with pre-trained and task-generic filters plays a role of semantic encoding for the web text data in the proposed attention model. Without any ground truth semantic aid, our model is able to extract useful information from web source data. To address the noise nature of web text, our model is also able to determine whether to attend text-inferred visual features or to original visual features adaptively. The summarization and future prospect of my PhD work is lastly discussed in Chapter 7.

# Publication List

## Journal Articles

- **Dongzhe Wang**, Kezhi Mao: *Task-Generic Semantic Convolutional Neural Network for Web Text-Aided Image Classification.* Revised for Neurocomputing.

- **Dongzhe Wang**, Kezhi Mao: *Learning Semantic Text Features for Web Text Aided Image Classification.* Revised for IEEE Transactions on Multimedia.

## Conference Proceedings

- **Dongzhe Wang**, Kezhi Mao: *Adaptive Web Text-Aided Attention network for One-shot Image Classification.* Submitted to International Conference on Multimedia Modeling. 2019.

- **Dongzhe Wang**, Kezhi Mao: *Multimodal Object Classification using Bidirectional Gated Recurrent Unit Networks.* IEEE International Conference on Data Science in Cyberspace. 2018.

- **Dongzhe Wang**, Kezhi Mao: *Convolutional Neural Networks and Multimodal Fusion for Text Aided Image Classification.* Information Fusion (FUSION), International Conference on. 2017.

- **Dongzhe Wang**, Kezhi Mao: *Adaptive Multimodal Fusion with Web Resources for Scene Classification.* Information Fusion (FUSION), International Conference on. 2016.

- **Dongzhe Wang**, Kezhi Mao: *Improving Scene Classification by Fusion of Training Data and Web Resources.* Information Fusion (FUSION), International Conference on. 2015.