Abstract

In the last decade, due to the rapid development of digital devices, Internet bandwidth and social networks, an enormous amount of multimedia data have been created in the *WWW* (World Wide Web). According to the publicly available statistics, more than 400 hours of video are uploaded to YouTube every minute [You18]. 350 million photos are uploaded to Facebook every day [Fac18]; By 2021, video traffic is expected to make up more than 82% of all consumer Internet traffic [Cis17]. There is thus a pressing need to develop automated technologies for analyzing and indexing those "big multimedia data" more accurately and efficiently. One of the current approaches is *Deep Learning*. This method is recognized as a particularly efficient machine learning method for multimedia data.

Deep learning (*DL*) is a sub-field of *Machine Learning* and *Artificial Intelligence*, and is based on a set of algorithms that attempt to learn representations of data and model their high-level abstractions. Since 2006 *DL* has attracted more and more attention in both academia and industry. Recently *DL* has produced break-record results in a broad range of areas, such as beating human in strategic game systems like Go (Googles AlphaGo [SSS+17]), autonomous driving [BDTD+16], and achieving dermatologist-level classification of skin cancer [EKN+17], etc.

In this *Habilitationsschrift*, I mainly address the following research problems:

Nature scene text detection and recognition with deep learning. In this work, we developed two automatic scene text recognition systems: *SceneTextReg* [YWBM16] and *SEE* [BYM18] by following the supervised and semi-supervised processing scheme, respectively. We designed novel neural network architectures and achieved promising results in both recognition accuracy and efficiency.

Deep representation learning for multimodal data. We studied two sub-topics: visual-textual feature fusion in multimodal and cross-modal document retrieval task [WYM16a]; Visual-language feature learning with its use case *image captioning*. The developed captioning model is robust to generate the new sentence descriptions for a given image in a very efficient way [WYBM16, WYM18].

We developed *BMXNet*, an open-source *Binary Neural Network* (BNN) implementation based on the well-known deep learning framework *Apache MXNet* [CLL+15]. We further conducted an extensive study on training strategy and executive efficiency of *BNN* on image classification task. We showed meaningful scientific insights and made our models and codes publicly available; these can serve as a solid foundation for the future research work.

Operability and accuracy of all proposed methods have been evaluated using publicly available benchmark data sets. While designing and developing theoretical algorithms, we also work on exploring how to apply these algorithms to practical applications. We investigated the applicability of two use cases, namely *automatic online lecture analysis* and *medical image segmentation*. The result demonstrates that such techniques might significantly impact or even subvert traditional industries and our daily lives.