UNIVERSITY OF CALIFORNIA, SAN DIEGO

Deep Learning for Image Understanding

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Electrical Engineering (Signal and Image Processing)

by

Yufei Wang

Committee in charge:

Professor Garrison W. Cottrell, Chair
Professor Nuno Vasconcelos, Co-Chair
Professor Kenneth Kreutz-Delgado
Professor Bhaskar D. Rao
Professor Lawrence K. Saul

2017

The dissertation of Yufei Wang is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Co-Chair

Chair

University of California, San Diego

2017

DEDICATION

To my family.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

The pursuit of Ph. D. is an unforgettable experience, and along the way, there are so many people that I want to thank.

The most important person I would like to thank is of course my Ph. D. advisor, Dr. Garrison W. Cottrell. Gary is very supportive in any way I can imagine, and his insightful ideas and concrete supervision are of great help for me. His serious attitude towards academia influenced my entire graduate study. Not only is Gary a great advisor in research, he is also a mentor and cares like a friend. Whenever I have troubles and doubts, he is always there listening, and ready to give advice.

I also want to thank my co-advisor, Dr. Nuno Vasconcelos, for giving me valuable advice on my research throughout my Ph. D. study. I am honored to have Dr. Bhaskar D. Rao, Dr. Lawrence K. Saul, and Dr. Kenneth Kreutz-Delgado to serve as my doctoral committee. Their insightful discussion and invaluable advice are great asset to my academia career.

I feel thankful to my collaborators from Adobe Research, Dr. Zhe Lin, Dr. Xiaohui Shen, Dr Scott Cohen, Dr. Jianming Zhang, Dr. Radomir Měch, and Dr. Gavin Miller. Through the internships and long time collaboration with Adobe Research, every deep discussion is very valuable to my research. I would like to express my special thanks to Zhe for his mentorship. His inspiration and insights as well as his detailed advice and patient supervision are invaluable to me. Discussion with Xiaohui is always inspiring, and I've learnt a lot from him. I also want to thank Scott for not only giving me valuable academic advice, but also cares me greatly in my personal life, which helped me go through my toughest time. My works greatly benefit from Jianming's insightful suggestions, and Radomir and Gavin's mentorship.

I would like to thank my boyfriend, Si Chen. He was also a graduate student in UCSD. Before that, we studied in the same University in China. For the many years I've known him, he has always been the first person I go to when I have troubles, and he not only comforts me but also gives concrete advice on the steps I can take. With him always being there for me, I know that there is nothing I cannot conquer. When I get too satisfied with my life and get lazy, he is there to remind me of my ambitions and goals; when I have doubts of myself, he always helps me regain the courage to continue fighting. He makes me a better person.

I also want to thank my parents, Suli Wu and Dr. Jianhua Wang, for their unconditional support and love for me throughout my life. Their upright, honest and hardworking personalities shaped my character. They have been strict with me, but they never fail to express how proud they are of me. When it comes to life-changing decisions such as pursuing a Ph. D., they give suggestions and express concerns, but fully support me once I make my decision. They never speak up for their love, but I can feel it every second, and I am grateful for it.

Along the journey, there are a lot of colleagues and friends that I would like to thank for their support on my research and life. My wonderful labmates from Gary's Unbelievable Research Unit (GURU) who make our lab a warm family: Honghao Shan, Ben Cipollini, Tomoki Tsuchida, Vicente Malave, Mohsen Malmir, Panqu Wang, Amanda Song, William Fedus, Yao Qin, Yan Shu, Davis Liang, Sanjeev Rao, Sandy Wiraatmadja, and Angel Zhang. A lot of thanks to my friends and colleagues: Yingwei Li, Shuai Tang, Ning Ma, Mengting Wan, Zhaowei Cai, Weixin Li, and Xiaodi Hou. Also, my friends outside of my research that made the four years of my life colorful: Wei Huang, Yilun Zhang, Lijuan Huang, Pengfei Chen, Zhiyuan Sun, Jiacong Li, Yao Peng, Yi Yang, Jingxin Ye, Gufeng Zhang, Yuan Fang, Chuan Wang, Qiao Zhang, Dongjin Song, Rui Hua, Huan Hu and Qian Yao.

I would like to thank the funders that supported my research, including the Temporal Dynamics of Learning Center (TDLC), NSF (SMA 1041755 to the TDLC, and NSF grant # IIS-1219252 to GWC), UC San Diego (provided a Fellowship to Yufei Wang), Adobe Research (provided gift money to Garrison W. Cottrell).

Chapter 2, in full, is an edited reprint of the material as it appears in the following publication: Wang, Y., Zhe, L., Shen, X., Měch, R., Miller, G., Cottrell, G. W., "Event-specific Image Importance", In *Computer Vision and Pattern Recognition (CVPR)*, 2016. The dissertation

author was a primary researcher and an author of the cited material.

Chapter 3, in full, is an edited reprint of the material as it appears in the following publication: Wang, Y., Zhe, L., Shen, X., Měch, R., Miller, G., Cottrell, G. W. (2017), "Recognizing and Curating Photo Albums via Event-Specific Image Importance", In *British Machine Vision Conference (BMVC)*, 2017. The dissertation author was a primary researcher and an author of the cited material.

Chapter 4, in full, is an edited reprint of the material as it appears in the following publication: Wang, Y., Zhe, L., Shen, X., Cohen, S., Cottrell, G. W., "Skeleton Key: Image Captioning by Skeleton-Attribute Decomposition", In *Computer Vision and Pattern Recognition (CVPR)*, 2017. The dissertation author was a primary researcher and an author of the cited material.

VITA

2013	B. S. in Electrical Engineering, University of Science and Technology of China, China
2017	M. S. in Electrical Engineering (Signal and Image Processing), University of California, San Diego
2017	Ph. D. in Electrical Engineering (Signal and Image Processing), University of California, San Diego

PUBLICATIONS

Wang, Y., Zhe, L., Shen, X., Zhang, J., Cohen, S., “Concept Mask: Large Scale Segmentation from Semantic Concepts”, Under Review in *Computer Vision and Pattern Recognition (CVPR)*, 2018

Wang, Y., Zhe, L., Shen, X., Měch, R., Miller, G., Cottrell, G. W., “Recognizing and Curating Photo Albums via Event-Specific Image Importance”, In *British Machine Vision Conference (BMVC)*, 2017.

Wang, Y., Zhe, L., Shen, X., Cohen, S., Cottrell, G. W., “Skeleton Key: Image Captioning by Skeleton-Attribute Decomposition”, In *Computer Vision and Pattern Recognition (CVPR)*, 2017

Wang, Y., Zhe, L., Shen, X., Měch, R., Miller, G., Cottrell, G. W., “Event-specific Image Importance”, In *Computer Vision and Pattern Recognition (CVPR)*, 2016

Rao, S., Wang, Y., G., Cottrell, G. W., “A Deep Siamese Neural Network Learns the Human-Perceived Similarity Structure of Facial Expressions Without Explicit Categories.” In *Proceedings of the 38th annual conference of the cognitive science society*, 2016.

Wang, Y., and Cottrell, G. W., “Bikers are like tobacco shops, formal dressers are like suits: Recognizing Urban Tribes with Caffe”. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2015.

Tang, A., Lu, K., Wang, Y., Huang, J., Li, H., “A real-time hand posture recognition system using deep neural networks.”, In *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2015

ABSTRACT OF THE DISSERTATION

Deep Learning for Image Understanding

by

Yufei Wang

Doctor of Philosophy in Electrical Engineering (Signal and Image Processing)

University of California, San Diego, 2017

Professor Garrison W. Cottrell, Chair
Professor Nuno Vasconcelos, Co-Chair

Computer vision and image understanding is the problem of interpreting images by locating, recognizing objects, attributes and other higher level features in an image. In this thesis, I seek to tackle this broad problem using deep learning techniques. More specifically, I build deep neural network based models to solve two specific problems to understand images in a high level: album wise image understanding with event-specific image importance score, and description generation for an image.

I first focus on the understanding of a collection of images in an event album. In an event album, some images are more important or interesting to save or present than others, and I show

that with an event-specific image importance property, we can learn the interestingness of an image given an album, and the performance of the model generated importance score is very close to human preference. I build a siamese network that can predict image importance score given the event type of that image, using novel objective function and learning scheme. Next, to make the process fully automated, I propose an iterative updating procedure for event type and image importance score prediction, that can simultaneously decide the event type of the album and the importance score of every image. It consists of a Convolutional Neural Network that recognizes the event type, a Long-Short Term Memory (LSTM) that uses sequential information for event type recognition, and a siamese network that predicts image importance score.

Furthermore, not just limited to describing an image with a score or by a classified type, I seek the possibility to describe it with a phrase or sentence. I propose a coarse-to-fine LSTM based method that decomposes the original image description into a skeleton sentence and its notable attributes, and demonstrate that in this way the language model can generate better descriptions, with the capability to generate image descriptions that better accommodates user preference.

Chapter 1

Introduction

Computer vision and image understanding is one of the main problem of artificial intelligence. It involves many attempts to help computer “see” images better. Early study for image understanding mostly focused on extracting the low level features, such as feature extraction for edges, corners, and optical flow [47, 14, 52]. The understanding of middle level features such as image segmentation, object detection and recognition then became major focus for many research studies [21, 36, 78, 119, 8]. More recently, with the access to large scale images with high quality annotations through the internet, and the speed up of computing with hardware innovation (GPUs), deep neural networks [43, 69] have brought great innovation into many research areas. Convolutional Neural Networks (CNN) has especially inspired great advance for many problems in image understanding [70, 98, 108, 38, 41]. Recurrent Neural Networks (RNN) and Long-Short Term Memory (LSTM) are widely used in sequence learning, such as machine translation [105] and image captioning [126].

In this thesis, I seek to use deep learning techniques to solve two problems in image understanding. First, I use a siamese network based model and LSTM based model to simultaneously predict album-wise event type and image importance for personal album organization. Second, I propose a coarse-to-fine LSTM based model for image caption generation. In this chapter, I provide relevant background knowledge for the topics relevant to this thesis.

1.1 Deep Learning

Most recently, thanks to the easy access to large scale image set via internet and great efforts researchers take to collect high quality annotations [94], the advance in network architecture [65, 100, 107, 48, 54], and development of faster computing hardware (GPUs), deep learning has been a great success, and has brought large performance boost to many areas in computer vision and image understanding, including object recognition [65, 48], object detection [41, 40, 93, 91], semantic segmentation [98, 64, 17, 130, 22], image captioning [117, 126, 81], and so on.

Deep convolutional neural networks (DCNN) are a type of feed forward network especially designed for image related task. They are advantageous over traditional multilayer perceptron networks in that they are much deeper, with tens or even hundreds of layers, and can learn the image from low level features to very high level features. The basic structure of unit in a DCNN consists of three layers: 1) a two dimensional convolutional layers that learns directly from the input image or from the activation of the previous layers. It preserves the spatial information of the input image and learns translation invariant features; 2) a spatial pooling layer which shrinks the size of features and at the same time enlarges the receptive field of the network; 3) a non-linear activation layer which improves the complexity and expressiveness of the network. With the stack of such units, the network is able to learn different level of features, from the low level features like corner and edges in the early layers, to the high level features like object parts and attributes in the late layers. The output of the stack of units is a high level feature vector representing the input image. In addition to the basic units, there are many variations of the network architecture to enhance the network's ability to interpret images [54, 48, 106, 49, 53].

On top of the feature extraction layers, the features are used for different tasks. For example, for object recognition, the final layer is an aggregated layer over different locations followed by a Softmax layer with cross-entropy loss function, and the output of the layer is the probability distribution of each object category given the input; for semantic segmentation, the output will be probability of each image pixel being in each object/stuff category.

With the use of back-propagation [71], DCNN's can learn the features from the image data directly, and greatly exceeds the performance of human designed features.

Recurrent neural networks (RNN), on the other hand, is different from feed forward networks in that the network not only takes its current input example as input, but also what it has perceived previously. It is designed for understanding a sequence of data, such as texts, handwriting, and spoken words. For each time step of an RNN, it has two sources of input: the present input data, and the output hidden state of the network in the previous time-step. The learning of RNN relies on back-propagation through time [84], the extension of back-propagation.

In this thesis, I seek to use deep learning techniques for image understanding problems.

1.2 Album-wise Image Understanding

The first problem I aim to tackle is to understand personal photo albums. A personal photo album is a collection of photos that we take in an event, for example a wedding event, or a trip event. The high level understanding of such photo collection involves two stages: recognizing the event type of the photo album, and suggesting the most important/interesting images in the collection to represent the album or to save for future use.

For event recognition, there are three types of approaches. The most popular approach takes videos as input and uses spatiotemporal features for event recognition [122]. The second approach uses single image as cue to recognize event type. This approach does not use temporal information or relevant frame importance, and only uses object level and scene level features from a single image [73]. In between the two approaches, album-wise event recognition has useful album-wise temporal information, but the images in an album are very sparse in time and is not temporally continuous. Bossard *et al.*[3] found the sequential information of the albums is helpful for learning the event type of the albums, despite their sparsity.

On the other hand, image importance is a complex image property that correlates with various factor, such as aesthetics [23], image interestingness [45, 28], and image memorability [55]. In this thesis, I propose a novel image property named event-specific image importance. To study this property, we collected the CUration of Flickr Events Dataset (CUFED), and let the human annotator to decide the image importance score given an event album. We intentionally gave vague instructions on how annotators decide the importance of an image, to encourage them to rate based on their intuition. We found out that the image importance is indeed highly related to the event type of the album it is in, and although the image importance is a highly subjective property, there is significant consistency across different annotators on the importance score they give in an album.

In this thesis, in Chapter 2, I propose a deep siamese architecture that learns the relative importance score of an image given the album event type it is from, assuming the event type of an album is given in advance. Further, in Chapter 3, I propose an iterative procedure that jointly learns the event type of an album and the importance score for each image. Thus, the two tasks for personal album understanding can be solved simultaneously with our framework.

1.3 Image Captioning

With the advance of image understanding with the development of deep learning, the research on image understanding is not constrained to the interpretation of an image with classification scores or detected tags, and the task of automatically describing the images with a sentence has drawn great attention. The problem is more challenging than conventional computer vision task in that the description generation requires high level understanding of the image beyond simple object recognition. It also requires the organization of a sentence that correctly conveys the notable information in the image.

The dominant approach for image captioning is inspired by the machine translation task [105]. For machine translation, an Encoder-Decoder network is used to map the input sequence to a vector of a fixed dimensionality, and then to decode the target sequence from the vector. The popular network used for encoding/decoding is Recurrent Neural Network (RNN), in which each element of the text sequence share the same unit parameters, and is sequentially fed into the network. RNN can deal with sequences with arbitrary length. Specifically, Long-Short Term Memory (LSTM), a variation of RNN, is commonly used [50]. It is capable to learn long-term dependencies with a cell state.

Similar to machine translation, an image can be viewed as a sentence in the source language, and an Encoder-Decoder network is used to translate it from the source language to the target sentence. Since the source "sentence" is in fact an image in the image captioning task, a CNN is used as Encoder, and LSTM is used as a Decoder.

Despite the great success in image captioning, most of the existing LSTM based methods suffer from two problems: 1) they tend to parrot back the sentences from the training corpus; and 2) the nature of predicting sentence words one by one means the attributes of a sentence is predicted before the object they are referring to, which is counter-intuitive.

To solve these two problems, in Chapter 4, I propose a coarse-to-fine model which decomposes the original caption into two parts: skeleton sentence which contains the main objects and structure in the sentence, and notable attributes for each object in the skeleton sentence.

1.4 Organization of the Thesis

In this thesis, I aim to tackle the two problems in high level image understanding. The rest of the thesis is organized as follows:

In Chapter 2, I introduce the problem of event-specific image importance. I collected a dataset for the study of this image property, and collect annotations of album-wise event type and image-wise importance score for the dataset, using Amazon Mechanical Turk (AMT). With the dataset, we show that the event-specific image importance property is subjective yet learnable. Furthermore, we propose a siamese network based architecture that can learn the image importance score with performance close to human perception, and the model assumes the event type information is know in advance.

In Chapter 3, I further extend our model to learn image importance score with no prior knowledge, by proposing an iterative procedure to learn the two album properties at the same time: album-wise event type recognition and image-wise importance score prediction. We show