TDT4501 Project
Report

Multimodal Deep Learning

Submitted in partial fulfillment of
the requirements for the award of the degree of

Master of Science
in
Computer Science

Submitted by

Roll No	Names of Students
¡Roll no here¿	Markus Lund
¡Roll no here¿	Dag Inge Helgøy

Under the guidance of
Massimiliano Ruocco

Department of Computer and Information Science
Norwegian University of Science and Technology
Trondheim, Norway

Fall Semester 2016

Abstract

¡Abstract here¿

Contents

List of Figures

Chapter 1

Introduction

1.1

Chapter 2

Motivation

Recent advancements made in multi-modal deep learning have opened the door for new use-cases in ranking and classification. Combining images and text into a combined search space give web-stores like Etsy give the customers the ability to search for visual features not already described in the product description. Providing accurate image captions could improve the Internet browsing experience for the blind or provide easy categorizing of images for the user.

While existing approaches to image/text comparison in the visual feature space have shown state-of-the-art results [3], we want to find out whether different iterations of the model that could provide better results, or similar results at a higher efficiency.

Chapter 3

State of the art

3.1 Word embeddings

3.1.1 Word2Vec

Word embeddings have the potential to describe relationships otherwise hidden from the representations like one hot or hashing and has gained a interest over the past few years. Word2Vec [1] is one such word embedding utilizing the strengths of vector representation and deep learning. It spreads words in a multi-dimensional embedding space allowing complex concepts be calculated using algebraic vector calculations. Word2Vec provides a dense and rich word representation which can be used in pre-prosession for deep learning to deal with the curse of dimensionality.

3.1.2 Visual Word2Vec

Recent multi-modal approaches to Word2Vec show consistent improvements in the visually grounded word embeddings. Visual Word2Vec [2] uses deep learning to cluster visual, rather than textual, semantics. For example, a relation between a girl eating and staring at an ice cream might be challenging to represent using only a textual embedding, but the visual relation is apparent. Visual Word2Vec aims to bring such visual relations closer together and word not visually related further apart.

3.2 Multi-modal comparison

3.2.1 Word2VisualVec

Word2VisualVec uses combines the pre-processing of Word2Vec with already existing convolutional neural networks (CNN). By using CNNs but stopping before the output layer it learns a visual vector representation of the image. This representation is then used to as training data on a feed-foreward neural network. By doing this it aims to compare text and images in a visual space.

3.3 Sentence generation

Generating natural language sentences from visual data require its own models and conventional feed-foreward networks fall short of current state-of-the-art models. Many primitive models like rule-based or logic systems have been proposed, but have been found to be narrow in its use and does not convert to different domains. Newer recurrent neural networks (RNNs) have recently had success in generated sequences in machine translation using their internal memory. An implementation of this called Long-Short Term Memory nets have been shown to provide state-of-the-art performance in translation [4, 5] and sentence generation recently sentence generation. [6]

Chapter 4

Challenges

Chapter 5

Ideas

By basing our work on [?] and exploring different structures we hope to find new models that could improve upon state-of-the-art image captioning and tagging. One such model is to reverse the inputs and outputs of our neural network.

Using a more viusally-based word embedding like Viusal Word2Vec in comibnation with other differences could improve pre-processing.

Chapter 6

Experimental part

Chapter 7

Conclusion

Chapter 8

Further work

References

[1] Distributed Representations of Words and Phrases and their Compositionality, https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositiona pdf

[2] Visual Word2Vec (vis-w2v): Learning Visually Grounded Word Embeddings Using Abstract Scenes, https://arxiv.org/pdf/1511.07067.pdf

[3] Word2VisualVec: Cross-Media Retrieval by Visual Feature Prediction https://arxiv.org/pdf/1604.06838v1.pdf

[4] Learning phrase representations using RNN encoder-decoder for statistical machine translation https://arxiv.org/pdf/1406.1078.pdf

[5] Sequence to sequence learning with neural networks https://arxiv.org/pdf/1409.3215v3.pdf

[6] Long short term memory http://deeplearning.cs.cmu.edu/pdfs/Hochreiter97_lstm.pdf