VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

IMAGE CAPTIONING WITH RECURRENT NEURAL NETWORKS

SEMESTRÁLNÍ PROJEKT
TERM PROJECT

AUTOR PRÁCE Bc. JAKUB KVITA
AUTHOR

BRNO 2015

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

POPIS FOTOGRAFIÍ POMOCÍ REKURENTNÍCH NEURONOVÝCH SÍTÍ

IMAGE CAPTIONING WITH RECURRENT NEURAL NETWORKS

SEMESTRÁLNÍ PROJEKT
TERM PROJECT

AUTOR PRÁCE Bc. JAKUB KVITA
AUTHOR

VEDOUCÍ PRÁCE Ing. MICHAL HRADIŠ, Ph.D.
SUPERVISOR

BRNO 2015

Abstrakt
Výtah (abstrakt) práce v českém jazyce.

Abstract
Výtah (abstrakt) práce v anglickém jazyce.

Klíčová slova
Klíčová slova v českém jazyce.

Keywords
Klíčová slova v anglickém jazyce.

Citace
Jakub Kvita: Image Captioning with Recurrent Neural Networks, semestrální projekt, Brno, FIT VUT v Brně, 2015

Image Captioning with Recurrent Neural Networks

Prohlášení

Prohlašuji, že jsem tento semestrální projekt vypracoval samostatně pod vedením pana Michala Hradiše.

....
Jakub Kvita
December 26, 2015

Poděkování

Zde je možné uvést poděkování vedoucímu práce a těm, kteří poskytli odbornou pomoc.

Contents

Chapter 1

Introduction

Klasicky popis toho co se tady bude dit, jak je to dulezite, atd.

Chapter 2

Neural networks

General idea of neural networks was slowly emerging after World War II. Perceptron, as a single neuron unit, was created in 1958 by Frank Rosenblatt[1], but became popular only after creation of backpropagation algorithm in 1975. At that time neural nets have not reached massive popularity, not because they are not working, but due to small computing power of machines back then and lack of datasets. Recently (after 2000) neural nets became popular again, rebranded as 'Deep Learning', because researchers realized that it is possible and very useful to stack neural nets on top of each other and create deep architectures, which are more practical than shallow ones. During this reinvention neural nets have been successfully applied in multiple fields like computer vision, speech recognition and natural language processing.

Since then various useful architectures and algorithms are now introduced almost every month. There is vast amount of various architectures and algorithms, in this chapter, I will describe only a couple – those used in this thesis.

2.1 Recurrent neural nets

Feedforward neural nets are extremely powerful models, which can be highly parallelized. Despite that, they can be only applied to problems with inputs and outputs, which have fixed dimensionality (e.g. one-hot encoding vectors). This is a serious drawback, as many of the real-world problems are defined as sequences with lengths that are unknown to us in beforehand. Soon recurrent neural networks were introduced and they proved to be very useful to this kind of task. There is vast amount of recurrent neural networks, many not suitable for sequential tasks like Hopfield network, which are very successful in specific tasks, but nevertheless not useful for us now.

We can understand recurrent neural networks as very deep forward nets with shared weights. Layers of this very deep net spread in time, together with the input sequence. This is very innovative idea, which enabled training RNN with backpropagation through time. It also shows that, as very deep networks, they have vanishing or exploding gradient problem. This is a serious issue,which is caused by iterating many times over the weights and the activation function with derivatives > 1 (exploding gradient) or < 1 (vanishing gradient). Gradient then dies out and learning stops. Among others this problem has been solved by the LSTM unit described in part 2.1.1, which is most popular now and following

Mam to tady rozebrat vice?

[1] The perceptron: A probabilistic model for information storage and organization in the brain. Rosenblatt, F. Psychological Review, Vol 65(6), Nov 1958, 386–408.

research resulting in GRU described in part 2.1.2.

Popis toho jak umi pracovat se sekvencema, predikci dalsiho prvku, da se pouzit na spoustu veci, zvuky, ceny na burze, preklady, predikci textu.

2.1.1 LSTM – Long Short-Term Memory

Jak to vyresilo problem vyse. Pridat i rovnice, ktere pouzivam ja, rozebrat dopodrobna.

[6]

2.1.2 GRU – Gated Recurrent Unit

Zminit jako updatovanou verzi

[2] [5]

2.1.3 Text sequences – Word level embeddings and character level

Mozna trochu upravit nazev. (Character level and word level embeddings)

Popis toho jak se pracuje s textem v rnn, ze to je taky sekvence. Character level, word level, embeddings. Popis rozdilu toho jak funguji preklady a generovani dalsiho prvku sekvence.

2.2 Convolutional neural nets

Kratky uvod do toho, kde se pouzivaji, jak se vyvinuly, jednoduchy popis toho jak funguji. Obrazek?

Asi neni potreba davat subsekce na vrstvy, staci popsat jak to funguje vsechno dohromady, jednotlive vrstvy ve vetach v jednom odstavci. Obrazek. V diplomce rozpracovat vic.

Chapter 3

Experiments

Kapitola jen na semestralni projekt. V diplomce ji odstranim.

Jak se to implementuje, jake knihovny se pouzivaji - Caffe, Theano, TensorFlow, Torch. Popsat ze Torch bude v tehle kapitole.

Budu popisovat veci co jsem zkousel implementovat v Torchi.

3.1 Torch

Torch se zrecykluje do diplomky.

Udelat tady tabulku o ruznych balicich co torch ma

Jak funguji rekurentni site v Torchi.

Nacitani modelu z Caffe, ukladani v Torchi...

[1]

3.1.1 nn, nngraph

Linky na knihovny v poznamkach pod carou.

3.1.2 rnn

3.1.3 Other packages

loadcaffe, optim,...

3.2 Predicting next character in sequence

Jak jsem to udelal, co to dela, ukazky.

Reference na Karpathyho char-rnn

[8]

Chapter 4

Image caption generation

Znovu uvod k tomu jak je to dulezite a tentokrat jak na tom lidi pracuji, co je potreba a jak se to hodnoti.

4.1 Related Work

Dat tomu nejake lepsi jmeno, clanky o popisovani obrazku ktere jsem cetl, pouzil.

4.1.1 Show and Tell

[14] [12]

Clanek z Coco od Googlu.

Zminit i strojovy preklad (Sequence to Sequence Learning with Neural Networks), architektura encoder, decoder

4.1.2 Show, Attend and Tell

[15]

Clanek z Coco z Montrealu/Toronta

4.1.3 From Captions to Visual Concepts and Back

[4]

Clanek z Coco od Microsoftu, mrknout se i na pokracovani v druhem clanku

4.1.4 Long-term Recurrent Convolutional Networks for Visual Recognition and Description

[3]

Clanek z Coco z berkeley

4.2 Datasets

COCO, Flicker, popis jake jsou. Asi zrusit sekce, udelat jen tabulku a mensi popis.

4.2.1 MS COCO

[10]

4.2.2 Flickr 30k,8k

[16] [7]

4.2.3 CIDEr datasets

[13]

4.3 Evaluation metrics

BLEU, cIDER, jak se pouzivaji, co delaji...

4.3.1 BLEU

[11]

4.3.2 CIDEr

[13]

4.3.3 METEOR

[9]

Chapter 5

Model

Do semestralniho projektu nebo az na diplomku?

Design modelu, co chci pouzit, jake metody chci zkusit.

Polozit si principialni otazku a zjistit jestli to nejak pomuze, jak to funguje.

5.1 Architecture

Architektura modelu, jake matematicke modely jsem pouzil, bez implementacnich detailu.

5.2 Training details

Popis pomoci jakeho algoritmu jsme trenovali, s jakyma parametrama, minibatches, datasety.

Chapter 6

Conclusion

Udelat jeden zaver pro semestralni projekt, pak ho prepsat pro diplomku.

Bibliography

[1] Torch – Scientific Computing for LuaJIT. http://torch.ch/. [Accessed: 2015-12-24].

[2] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *CoRR*, abs/1406.1078, 2014.

[3] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term Recurrent Convolutional Networks for Visual Recognition and Description. *CoRR*, abs/1411.4389, 2014.

[4] Hao Fang, Saurabh Gupta, Forrest N. Iandola, Rupesh K. Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. From Captions to Visual Concepts and Back. *CoRR*, abs/1411.4952, 2014.

[5] Klaus Greff, Rupesh Kumar Srivastava, Jan Koutník, Bas R. Steunebrink, and Jürgen Schmidhuber. LSTM: A Search Space Odyssey. *CoRR*, abs/1503.04069, 2015.

[6] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, November 1997.

[7] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.

[8] Andrej Karpathy. The Unreasonable Effectiveness of Recurrent Neural Networks. http://karpathy.github.io/2015/05/21/rnn-effectiveness/. [Accessed: 2015-12-30].

[9] Alon Lavie and Abhaya Agarwal. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 228–231, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

[10] Microsoft. Common Objects in Context. http://mscoco.org/dataset/#download. [Accessed: 2015-12-29].

[11] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[12] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to Sequence Learning with Neural Networks. *CoRR*, abs/1409.3215, 2014.

[13] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-Based Image Description Evaluation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[14] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and Tell: A Neural Image Caption Generator. *CoRR*, abs/1411.4555, 2014.

[15] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *CoRR*, abs/1502.03044, 2015.

[16] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.