

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
AUTHOR

Bc. JAKUB KVITA

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 NEU- RONOVÝCH SÍTÍ

IMAGE CAPTIONING WITH RECURRENT NEURAL NETWORKS

SEMESTRÁLNÍ PROJEKT

TERM PROJECT

AUTOR PRÁCE

AUTHOR

Bc. JAKUB KVITA

VEDOUcí PRÁCE

SUPERVISOR

Ing. MICHAL HRADIŠ, Ph.D.

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 24, 2015

Poděkování

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

© Jakub Kvita, 2015.

Tato práce vznikla jako školní dílo na Vysokém učení technickém v Brně, Fakultě informačních technologií. Práce je chráněna autorským zákonem a její užití bez udělení oprávnění autorem je nezákonné, s výjimkou zákonem definovaných případů.

Contents

1	Introduction	2
2	Neural networks	3
2.1	Recurrent neural nets	3
2.1.1	LSTM – Long Short-Term Memory	3
2.1.2	GRU – Gated Recurrent Unit	4
2.1.3	Text sequences – Character level and word level embeddings	4
2.2	Convolutional neural nets	4
3	Experiments	5
3.1	Torch	5
3.1.1	nn, nngraph	5
3.1.2	rnn	5
3.1.3	Other packages	5
3.2	Predicting next character in sequence	5
4	Image caption generation	6
4.1	Related Work	6
4.1.1	Show and Tell	6
4.1.2	Show, Attend and Tell	6
4.1.3	From Captions to Visual Concepts and Back	6
4.1.4	Long-term Recurrent Convolutional Networks for Visual Recognition and Description	6
4.2	Datasets	7
4.2.1	MS COCO	7
4.2.2	Flickr 30k,8k	7
4.2.3	CIDEr datasets	7
4.3	Evaluation metrics	7
4.3.1	BLEU	7
4.3.2	CIDEr	7
4.3.3	METEOR	7
5	Model	8
5.1	Architecture	8
5.2	Training details	8
6	Conclusion	9

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 Rosenblatt¹, 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. Mostly because researchers dealt with the problems from before and successfully applied neural nets 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, which are 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 different kinds of neural networks, many not suitable for sequential tasks

Zduraznit problem vanishing a exploding gradientu

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]

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

2.1.2 GRU – Gated Recurrent Unit

Zmínit jako updatovanou verzi

[2] [5]

2.1.3 Text sequences – Character level and word level embeddings

Možná trochu upravit název. (Character level and word level embeddings)

Popis toho jak se pracuje s textem v rnn, že to je taky sekvence. Character level, word level, embeddings. Popis rozdílu toho jak fungují překlady a generování dalšího prvku sekvence.

2.2 Convolutional neural nets

Kratký úvod do toho, kde se používají, jak se vyvinuly, jednoduchý popis toho jak fungují. Obrazek?

Asi není potřeba dávat subsekcce na vrstvy, stačí popsat jak to funguje všechno dohromady, jednotlivé vrstvy ve větách v jednom odstavci. Obrazek. V diplomce rozpracovat víc.

Chapter 3

Experiments

Kapitola jen na semestrální projekt. V diplomce ji odstraním.

Jak se to implementuje, jaké knihovny se používají - Caffe, Theano, TensorFlow, Torch. Popsat ze Torch bude v této kapitole.

Budu popisovat věci co jsem zkoušel implementovat v Torch.

3.1 Torch

Torch se zrecykluje do diplomky.

Udělat tedy tabulku o různých balících co torch má

Jak fungují rekurentní sítě v Torch.

Nacítání modelu z Caffe, ukládání v Torch...

[1]

3.1.1 nn, nngraph

Linky na knihovny v poznámkách pod textem.

3.1.2 rnn

3.1.3 Other packages

loadcaffe, optim,...

3.2 Predicting next character in sequence

Jak jsem to udělal, co to dělá, ukázky.

Reference na Karpatyho 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, Flickr, 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 používají, co dělají...

4.3.1 BLEU

[11]

4.3.2 CIDEr

[13]

4.3.3 METEOR

[9]

Chapter 5

Model

Do semestrálního projektu nebo až na diplomku?

Design modelu, co chci použít, jaké metody chci zkusit.

Položit si principiální otázku a zjistit, jestli to nějak pomůže, jak to funguje.

5.1 Architecture

Architektura modelu, jaké matematické modely jsem použil, bez implementačních detailů.

5.2 Training details

Popis pomocí jakého algoritmu jsme trénovali, s jakými parametry, minibatches, datasy.

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 Merriënboer, Çağlar 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.