Deep Image Captioning and Neural Machine Translation for Estonian

Tanel Pärnamaa
Institute of Mathematical Statistics, Faculty of Mathematics and Computer Science, University of Tartu

Introduction

The aim of this project is to generate automatic image descriptions in Estonian and is inspired by the recent success of image captioning systems.[1,2,3,4,5] In other words, we want to train a machine to understand images and to reason about them in Estonian.

Method

We go from *image* to *desctiption in English* to *description in Estonian*. The reason is twofold: we do not have high quality image-sentence datasets with Estonian captions (such as Flickr 30K or COCO for English) and because caption generation is like machine translation (translate from an image to a description *vs* transate from one language to another).

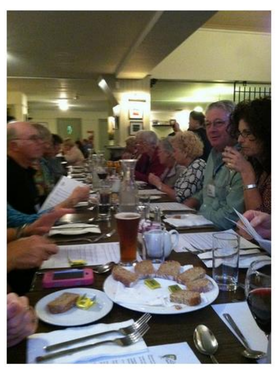
We use a convolutional neural network (CNN) to encode an image and a long short-term memory (LSTM) to decode the image representation to English translation. For that we use a pretrained model.[1] Bidirectional recurrent neural network (RNN) is used to encode an English sentence to a distributed representation and another RNN is used to decode it to Estonian sentence.[6] This model was trained on Europarl and OpenSubs parallel corpus on GPU around 5 days using GroundHog.

Examples

Sample translations of sentences about recent news about Greek debt from BBC News. You can play with the neural machine translation model from here: `www.stat24.ee/tolge`

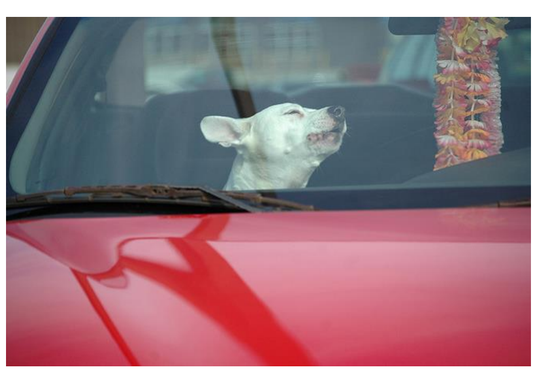
Source sentence	Nobody believes that the Greek debt is sustainable.	Despite the fact that there are differences in perspective, I am absolutely confident that we will soon manage to reach a mutually beneficial agreement, both for Greece and for Europe as a whole.	There is a feeling among many here that this is a historic moment, and supporters speak in terms of big ideas and dreams.
Neural MT	Mitte keegi ei usu, et Kreeka võlg on jätkusuutlik.	Hoolimata asjaolust, et esineb erinevusi, olen ma täiesti kindel, et me saavutame peagi lahenduse, nii Kreeka kui ka Euroopa jaoks tervikuna.	Paljude seas on tunda, et see on ajalooline hetk, ja toetajad kõnelevad suurte ideede ja unistuste osas.
masintolge.ut.ee	keegi ei usu , et kreeka võla jätkusuutlikkus	hoolimata asjaolust , et leidub erinevusi perspektiivi , ma olen täiesti kindel , et me suudame jõuda peagi vastastikku kasulik leping nii kreeka kui ka kogu euroopa jaoks .	seal on tunne paljudest siin , et see on ajalooline hetk ja toetajate rääkida seoses suurte ideede ja unistused .
Google Translate	Keegi ei usu, et Kreeka võlg on jätkusuutlik.	Vaatamata sellele, et esineb erinevusi perspektiivi, ma olen täiesti kindel, et me varsti õnnestub jõuda vastastikku kasulik leping, nii et Kreeka ja Euroopa tervikuna	On tunne, paljude siin, et see on ajalooline hetk, ja pooldajad nii suuri ideid ja unistusi.

Sample generated image captions (the first row shows the generated English caption, then the translation from the neural machine translation model. UT means the translation from *masintolge.ut.ee,* GT means Google Translate.)



`a group of people sitting at a table with food`

Kamp inimesi, kes istuvad laua ääres.
UT: rühm inimesi istudes laua taga koos toiduga
GT: grupp inimesti istub laua koos toiduga

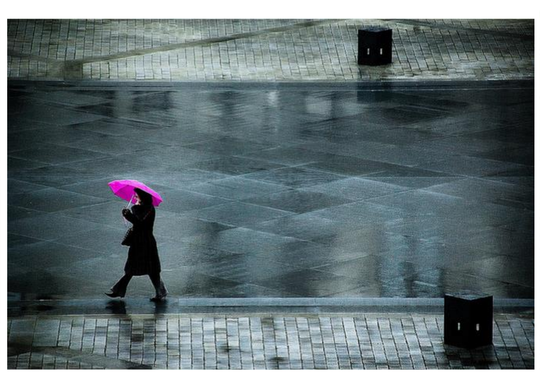


`a dog is sitting on a car looking at the camera`

Koer istub autos, vaatab kaamerasse
UT: Koer istub auto vaadates kaamera
GT: Koer istub auto vaadates kaamera

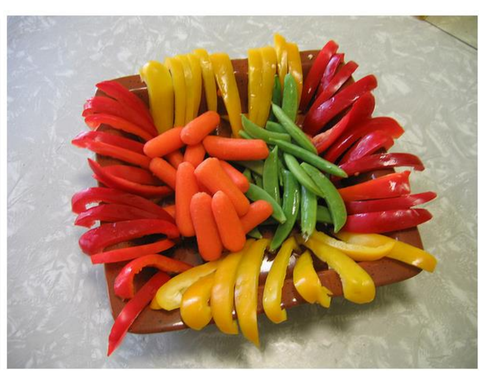
`a large truck is parked in a field`

Suur auto on pargitud põllul
UT: Suur auto on pargitud valdkonnas
GT: suur veoauto on pargitud valdkonnas



`a man is walking down a sidewalk holding an umbrella`

Mees, kes kõnnib tänaval, hoiab vihmavarju käes.
UT: mees kõnnib sätestatakse kõnnitee holding vihmavarju
GT: mees kõnnib mööda kõnniteed, kellel vihmavari



`a bunch of bananas sitting on top of a table`

Kamp banaane istub laua otsas.
UT: Banaane tabli tipus istung
GT: hunnik banaane istud peal tabelis

[1] Deep Visual-Semantic Alignments for Generating Image Descriptions, A. Karpathy, L Fei-Fei, (https://github.com/karpathy/neuraltalk)
[2] Show and Tell: A Neural Image Caption Generator, O. Vinyals *et al*
[3] Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models, R. Kiros *et al*
[4] Long-term Recurrent Convolutional Networks for Visual Recognition and Description, J. Donahue *et al*
[5] Explain Images with Multimodal Recurrent Neural Networks, J. Mao *et al*
[6] Neural Machine Translation by Jointly Learning to Align and Translate, D. Bahdanau *et al, (https://github.com/lisa-groundhog/GroundHog)*