Attention Models

Olof Mogren

Chalmers University of Technology

Feb 2016

CHALMERS

Attention Models

- Focus on parts of input
- Improves NN performance on different tasks
- IBM1 attention mechanism (1980's)

CHALMERS

Attention Models

- "One of the most exciting advancements"
  - Ilya Sutskever, Dec 2015

CHALMERS

Arxiv 2016

Multi-Way, Multilingual Neural Machine Translation with a Shared [...]

Incorporating Structural Alignment Biases into an Attentional Neural [...]

Language to Logical Form with Neural Attention

Human Attention Estimation for Natural Images: An Automatic Gaze [...]

Implicit Distortion and Fertility Models for Attention-based [...]

Survey on the attention based RNN model and its applications in [...]

From Softmax to Sparsemax: A Sparse Model of Attention and [...]

A Convolutional Attention Network for Extreme Summarization [...]

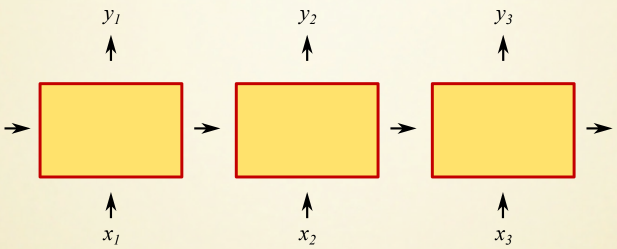
Learning Efficient Algorithms with Hierarchical Attentive Memory

Attentive Pooling Networks

Attention-Based Convolutional Neural Network for Machine [...]

CHALMERS

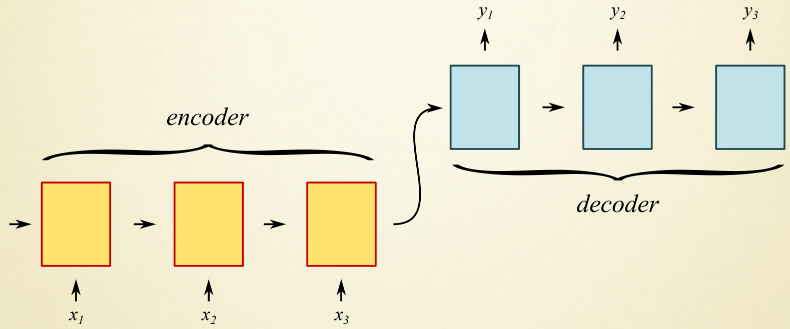
Modelling Language using RNNs



- Language models: $P(word_i | word_1, ..., word_{i-1})$
- Recurrent Neural Networks
- Gated additive sequence modelling:
  LSTM (and variants) details
- Fixed vector representation for sequences

CHALMERS

Encoder-Decoder Framework



- Sequence to Sequence Learning with Neural Networks *Ilya Sutskever, Oriol Vinyals, Quoc V. Le, NIPS 2014*
- Neural Machine Translation (NMT)
- Reversed input sentence!

CHALMERS

NMT with Attention

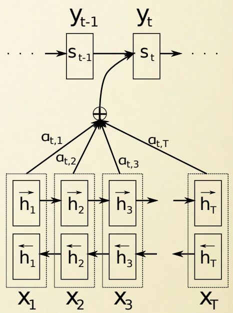
$$p(y_i | y_1, ..., y_{i-1}, x) = g(y_{i-1}, s_i, c_i)$$

$$s_i = f(s_{i-1}, y_{i-1}, c_i)$$

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$$

$$e_{ij} = a(s_{i-1}, h_j)$$



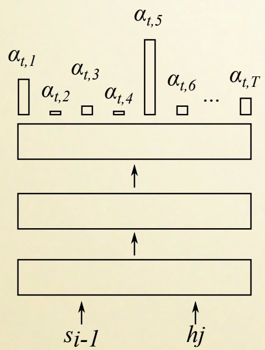
Neural Machine Translation by Jointly Learning to Align and Translate
Bahdanau, Cho, Bengio, ICLR 2015

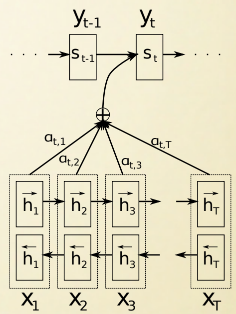
CHALMERS

NMT with Attention

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$$

$$e_{ij} = a(s_{i-1}, h_j)$$

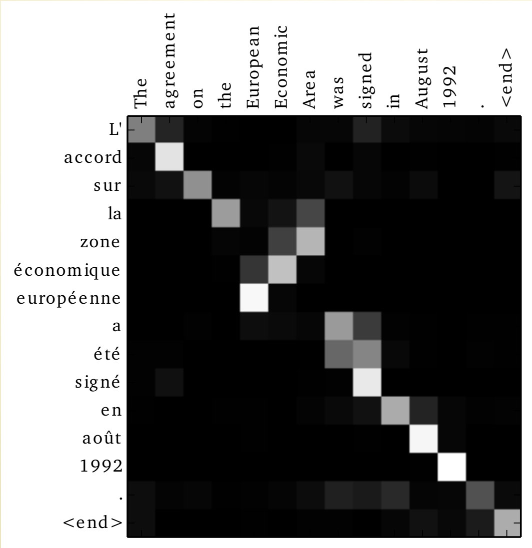




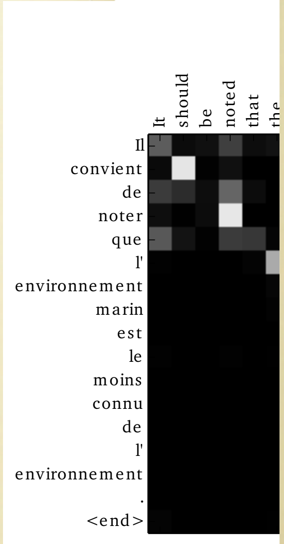
Neural Machine Translation by Jointly Learning to Align and Translate
Bahdanau, Cho, Bengio, ICLR 2015

CHALMERS

Alignment - (more)

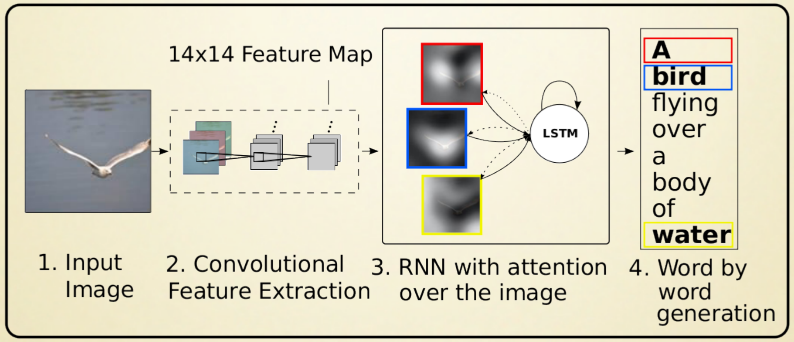


(a)



CHALMERS

Caption Generation



- "Translating" from images to natural language

CHALMERS

Caption Generation

- Convolutional network:
  Oxford net,
  19 layers,
  stacks of 3x3 conv-layers,
  max-pooling.
- Annotation vectors: $a = \{\boldsymbol{a}_1, ..., \boldsymbol{a}_L\}$, $\boldsymbol{a}_i \in \mathbb{R}^D$
- Attention over a.

CHALMERS

Attention Visualization

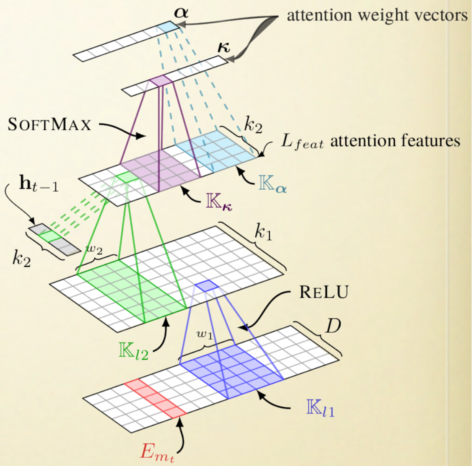
A woman is throwing a frisbee in a park.

CHALMERS

Source Code Summarization

- Predict function names given function body
- Convolutional attention mechanism; 1D patterns
- Out of vocabulary terms handled (copy mechanism)

details



A Convolutional Attention Network for Extreme Summarization of Source Code*Allamanis et al. Feb 2016 (arxiv draft)*

CHALMERS

Source Code Summarization

Target			Attention Vectors	λ
m_1	is	α =	<s> { return (m Flags & e Bullet Flag) == e Bullet Flag ; } </s>	0.012
		κ =	<s> { return (m Flags & e Bullet Flag) == e Bullet Flag ; } </s>	
m_2	bullet	α =	<s> { return (m Flags & e Bullet Flag) == e Bullet Flag ; } </s>	0.436
		κ =	<s> { return (m Flags & e Bullet Flag) == e Bullet Flag ; } </s>	
m_3	End	α =	<s> { return (m Flags & e Bullet Flag) == e Bullet Flag ; } </s>	0.174
		κ =	<s> { return (m Flags & e Bullet Flag) == e Bullet Flag ; } </s>	

A Convolutional Attention Network for Extreme Summarization of Source Code*Allamanis et al. Feb 2016 (arxiv draft)*

CHALMERS

Memory Networks

- Attention refers back to internal memory; state of encoder
- Neural Turing Machines
- (End-To-End) Memory Networks: explicit memory mechanisms (out of scope today)

CHALMERS

mogren@chalmers.se

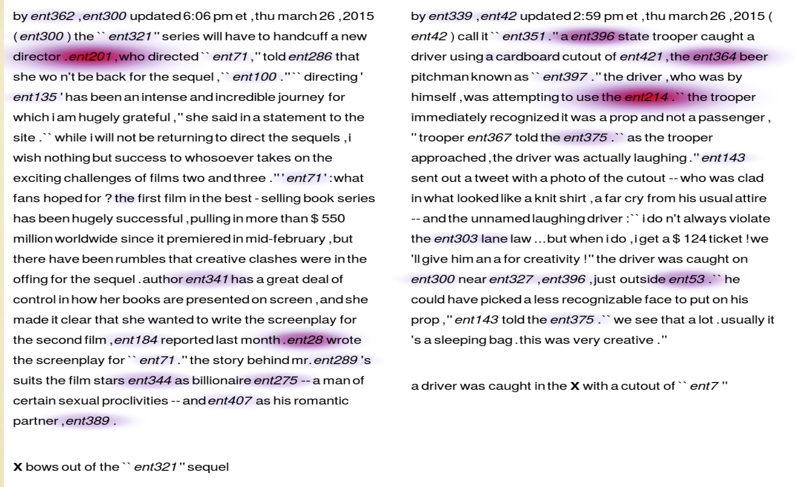
http://mogren.one/

http://www.cse.chalmers.se/research/lab/

CHALMERS

Appendix

CHALMERS



by *ent362* ,*ent300* updated 6:06 pm et ,thu march 26 ,2015 (*ent300*) the `` *ent321* '' series will have to handcuff a new director .*ent201* ,who directed `` *ent71* ,'' told *ent286* that she wo n't be back for the sequel ,`` *ent100* .'' `` directing ' *ent135* ' has been an intense and incredible journey for which i am hugely grateful ,'' she said in a statement to the site .`` while i will not be returning to direct the sequels ,i wish nothing but success to whosoever takes on the exciting challenges of films two and three .'' ' *ent71* ' :what fans hoped for ? the first film in the best - selling book series has been hugely successful ,pulling in more than $ 550 million worldwide since it premiered in mid-february ,but there have been rumbles that creative clashes were in the offing for the sequel .author *ent341* has a great deal of control in how her books are presented on screen ,and she made it clear that she wanted to write the screenplay for the second film ,*ent184* reported last month .*ent28* wrote the screenplay for `` *ent71* .'' the story behind mr. *ent289* 's suits the film stars *ent344* as billionaire *ent275* -- a man of certain sexual proclivities -- and *ent407* as his romantic partner ,*ent389* .

X bows out of the `` *ent321* '' sequel

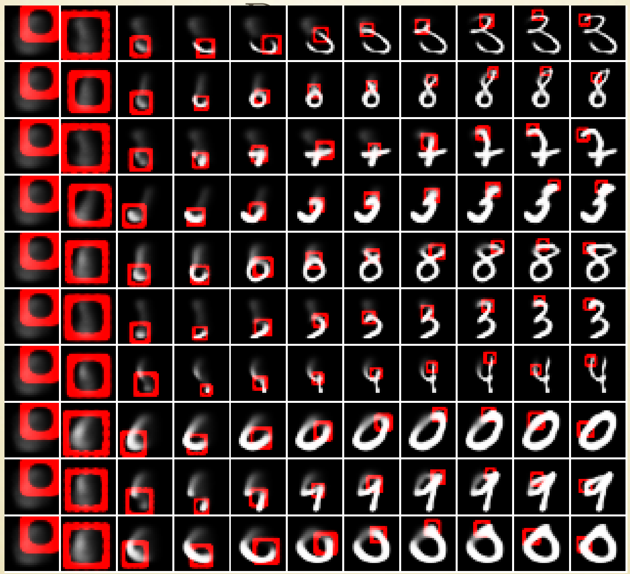
by *ent339* ,*ent42* updated 2:59 pm et ,thu march 26 ,2015 (*ent42*) call it `` *ent351* .'' a *ent396* state trooper caught a driver using a cardboard cutout of *ent421* ,the *ent364* beer pitchman known as `` *ent397* .'' the driver ,who was by himself ,was attempting to use the *ent214* .`` the trooper immediately recognized it was a prop and not a passenger ,'' trooper *ent367* told the *ent375* .`` as the trooper approached ,the driver was actually laughing .'' *ent143* sent out a tweet with a photo of the cutout -- who was clad in what looked like a knit shirt ,a far cry from his usual attire -- and the unnamed laughing driver :`` i do n't always violate the *ent303* lane law ...but when i do ,i get a $ 124 ticket !we 'll give him an a for creativity !'' the driver was caught on *ent300* near *ent327* ,*ent396* ,just outside *ent53* .`` he could have picked a less recognizable face to put on his prop ,'' *ent143* told the *ent375* .`` we see that a lot .usually it 's a sleeping bag .this was very creative .''

a driver was caught in the **X** with a cutout of `` *ent7* ''

Teaching Machines to Read and Comprehend, Dec 2015

Hermann, Kocisky, Greffenstette, Espeholt, Kay, Suleyman, Blunsom

CHALMERS

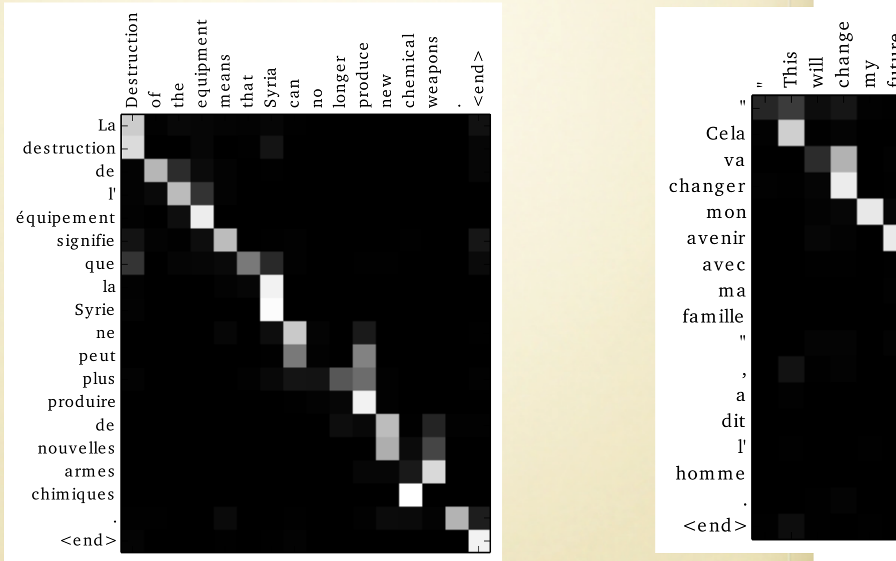


DRAW, A Recurrent Neural Network For Image Generation - 2015

Gregor, Danihelka, Graves, Rezende, Wierstra

CHALMERS

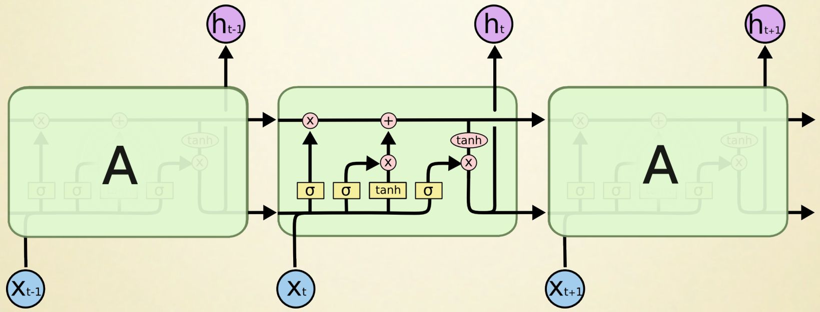
Alignment - (back)



(c)

CHALMERS

LSTM

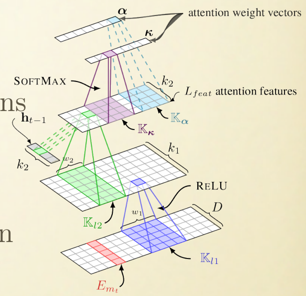


Christopher Olah

back

Source Code Summarization

- $\mathbb{K}_{l1}$: patterns in input
- $\mathbb{K}_{l2}$ (and $\mathbb{K}_{\boldsymbol{\alpha}}, \mathbb{K}_{\boldsymbol{\kappa}}$): higher level abstractions
- $\boldsymbol{\alpha}, \boldsymbol{\kappa}$: attention over input subtokens
- Simple version: only $\mathbb{K}_{\boldsymbol{\alpha}}$, for decoding
- Complete version: uses $\mathbb{K}_{\boldsymbol{\lambda}}$ for deciding on generation or copying



back

A Convolutional Attention Network for Extreme Summarization of Source Code

Allamanis et al. Feb 2016 (arxiv draft)

IBM Model 1: The first translation attention model!

A simple generative model for $p(\mathbf{s}|\mathbf{t})$ is derived by introducing a latent variable $\mathbf{a}$ into the conditional probabiliy:

$$p(\mathbf{s}|\mathbf{t}) = \sum_{\mathbf{a}} \frac{p(J|I)}{(I+1)^J} \prod_{j=1}^{J} p(s_j|t_{a_j}),$$

where:

- $\mathbf{s}$ and $\mathbf{t}$ are the input (source) and output (target) sentences of length J and I respectively,
- $\mathbf{a}$ is a vector of length J consisting of integer indexes into the target sentence, known as the alignment,
- $p(J|I)$ is not importent for training the model and we'll treat it as a constant ϵ.

To learn this model we use the EM algorithm to find the MLE values for the parameters $p(s_j|t_{a_j})$.

back

Soft vs Hard Attention

Soft

- Weighted average of whole input
- Differentiable loss
- Increased computational cost

Hard

- Sample parts of input
- Policy gradient
- Variational methods
- Reinforcement Learning
- Decreased computational cost