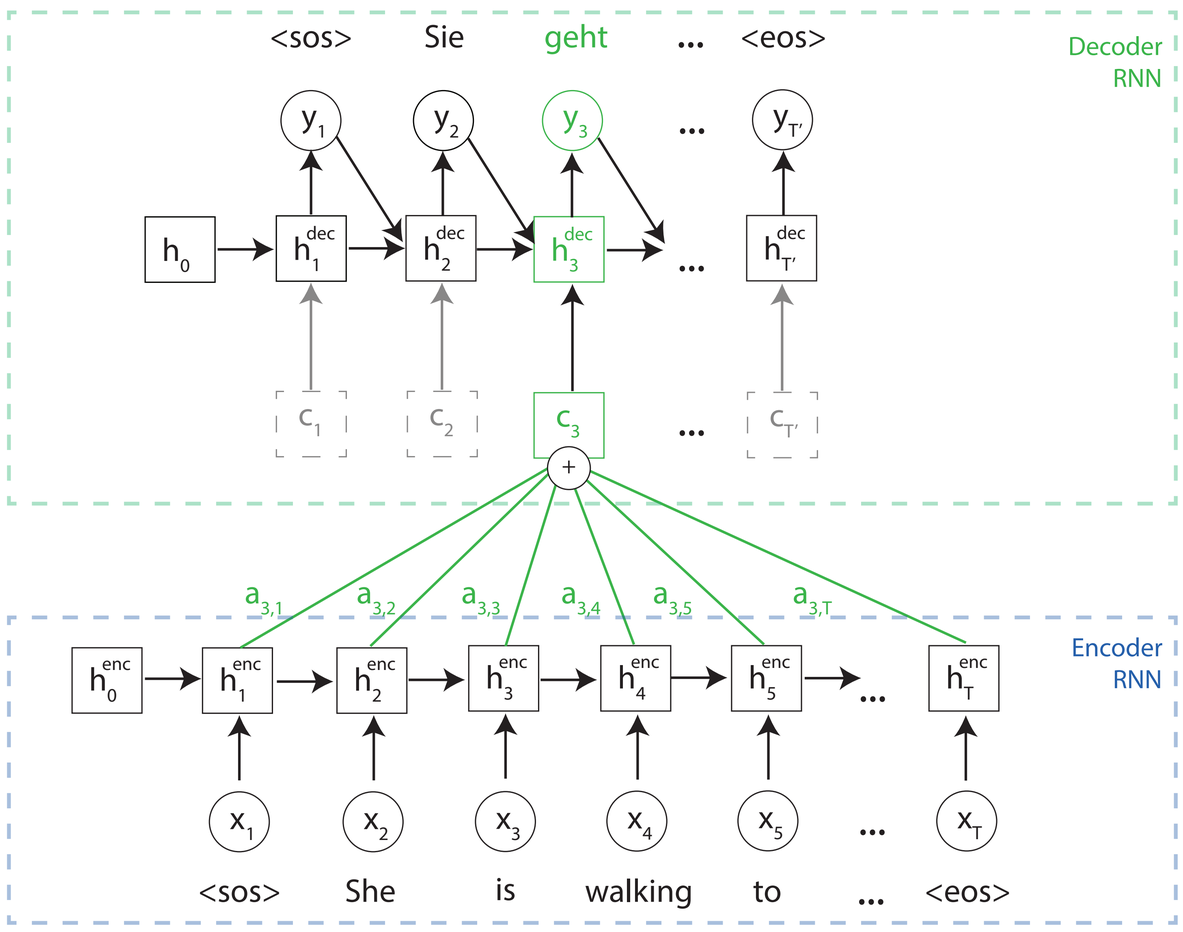
# Attention Mechanisms in Neural Encoder-Decoder Architectures

Sebastian Springenberg
WTM, Hamburg, Germany
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

## Introduction

- Neural Encoder-Decoder architectures allow to map sequences of variable length while being trainable in an end-to-end fashion
- Encoder encodes source sequence into a context vector
- Decoder decodes context vector to produce a target sequence
- Compressing the whole source sentence into a fixed-length context vector can lead to a loss of valuable information
- Attention mechanisms enable the system to dynamically allocate attention to relevant parts of the source sentence, thereby constructing distinct context vectors

## Neural Machine Translation with Temporal Attention




- Two RNNs as encoder and decoder respectively
- Encoder reading source sentence $x$ while updating hidden state $h^{enc}$:

$$h_t^{enc} = f(h_{t-1}^{enc}, x_t) \tag{1}$$

- Decoder producing target sentence $y$ while updating hidden state $h^{dec}$:

$$h_t^{dec} = f(h_{t-1}^{dec}, y_{t-1}, c_t). \tag{2}$$

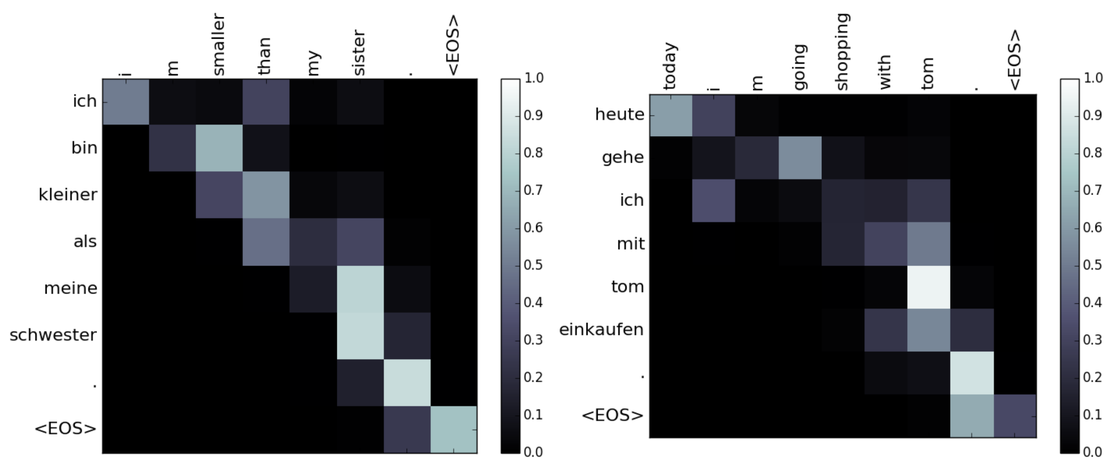
- Context vector $c_t$ calculated by a combination of encoder hidden states weighted by attention weights $a$:

$$c_t = \sum_{i=1}^{T} a_{t,i} * h_i^{enc}, \tag{3}$$

$$a_{t,1\dots T} = softmax(align(h_{t-1}^{dec}, y_{t-1}, h_{1\dots T}^{enc})), \tag{4}$$

where *align* is a forward neural network

## Neural Machine Translation: Attention Weights




## References

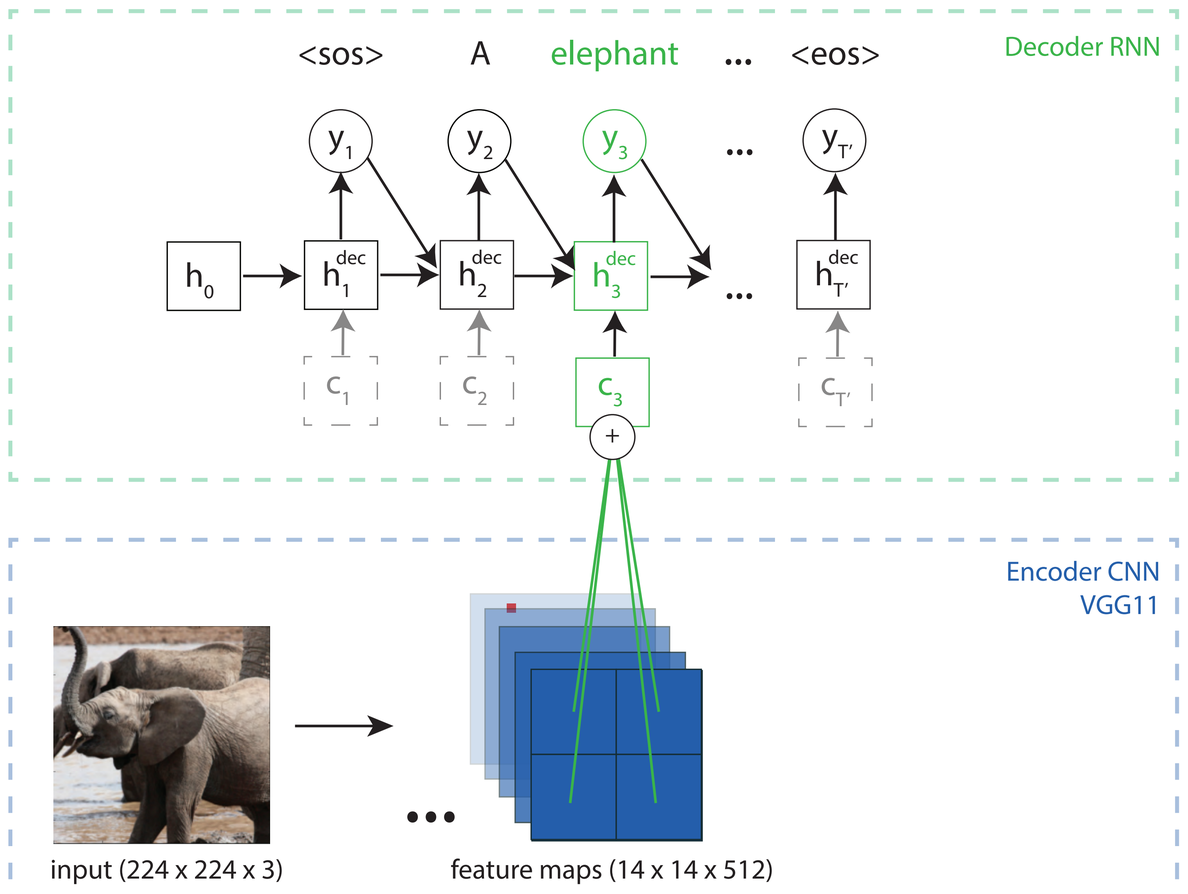
[1] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," 2015.

[2] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *ICLR 2015*, 2014.

[3] K. Cho, A. Courville, and Y. Bengio, "Describing multimedia content using attention-based encoder-decoder networks," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 1875–1886, 2015.

[4] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015.

## Image Captioning with Spatial Attention

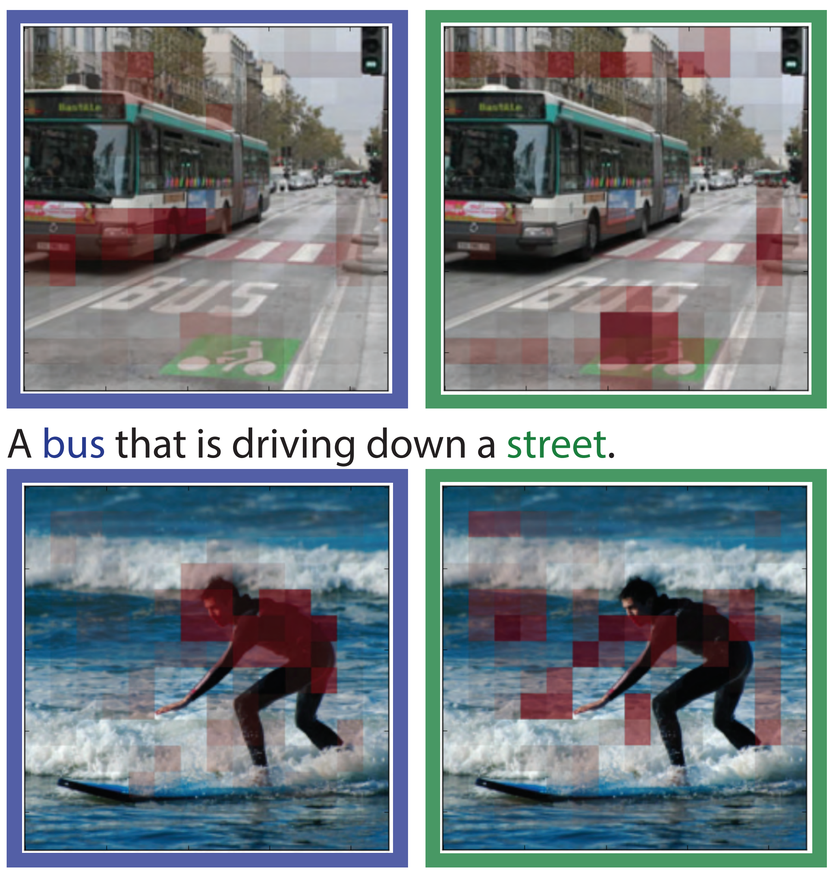



- Pretrained VGG11 CNN as encoder obtaining image features
- Decoder RNN producing image caption by updating hidden state as in equation 2
- Attention weights $a$ calculated by:

$$a_{t,1\dots T} = softmax(align(h_{t-1}^{dec}, y_{t-1}, x_{1\dots T})), \tag{5}$$

where $x$ represents image features of $14 \times 14$ locations

## Image Captioning: Attention Weights




## Conclusion

- Attention mechanisms can help to improve performance in sequence to sequence learning by taking into account the most relevant information
- Can also be seen as a form of dynamic memory allocation
- Can be applied to the mapping of any input / output representations (video to text, audio to text...)
- Hidden representation with attention weights learned by the decoder might be valuable for further transfer learning

5springe@informatik.uni-hamburg.de