LT2318 H21 AICS: Tutorial on Image Captioning

Nikolai Ilinykh[1]
[1]Department of Philosophy, Linguistics and Theory of Science
Centre for Linguistic Theory and Studies in Probability (CLASP)
University of Gothenburg, Sweden
{name.surname}@gu.se

Presented at, November 30, 2021

Outline I: Task

GÖTEBORGS UNIVERSITET

The simplest language-and-vision task that we are going to look at in this course is **image captioning**: describing an image with a single sentence.

- What is included in a standard image captioning model (ICM)?
- How to build an ICM in terms of coding?
- The design tricks for building better ICMs
- How do we *evaluate* ICMs?
- Closer to state-of-the-art ICM: attention and transformers
- A teaser about other language-and-vision tasks we will look in this course

Outline II: Models

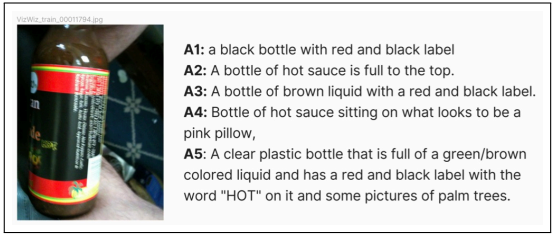
GÖTEBORGS UNIVERSITET

We are going to work mostly with **sequence-to-sequence** models (RNNs, LSTMs) for natural language generation (NLG) in multi-modal setting.

- What does *seq-to-seq* mean and what type of models and/or problems can be solved with such models?
- We will focus a lot on recurrent neural networks, the key concepts and mechanisms behind such networks
- Example RNN: character-level seq-to-seq model

CLASP

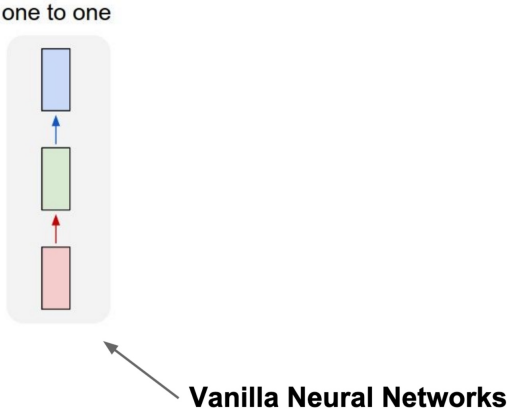
Example of the Project on Image Captioning



The key parts of the projects include:

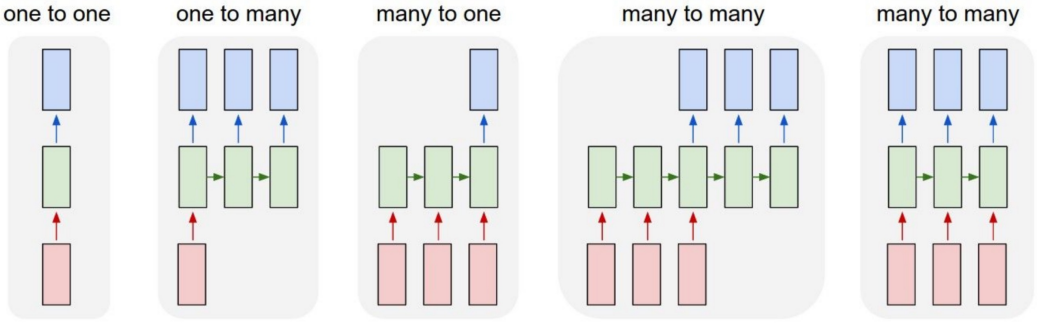
- **VizWiz** (Gurari et al., 2020): the dataset of images taken by people who are either blind or visually impared
- **Image Captioning model with Attention** (Xu et al., 2015)
- Evaluation: report **accuracy, loss, NLG evaluation metrics** (e.g., BLEU (Papineni et al., 2002)), analyse attention on the image

Basics of Sequence Networks[1]

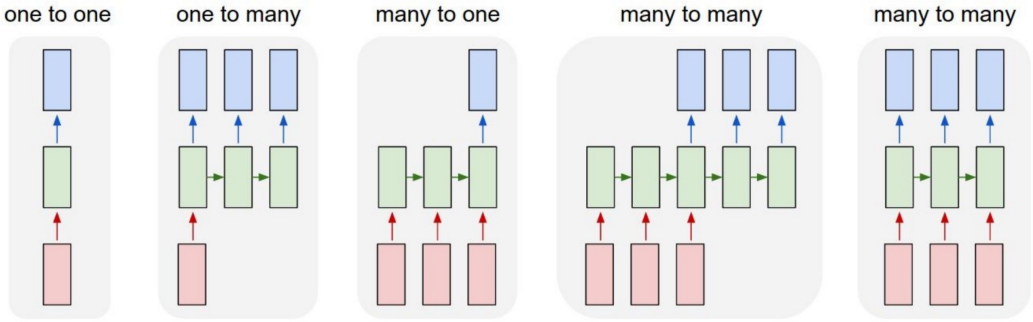


[1] http://karpathy.github.io/2015/05/21/rnn-effectiveness/

Basics of Sequence Networks

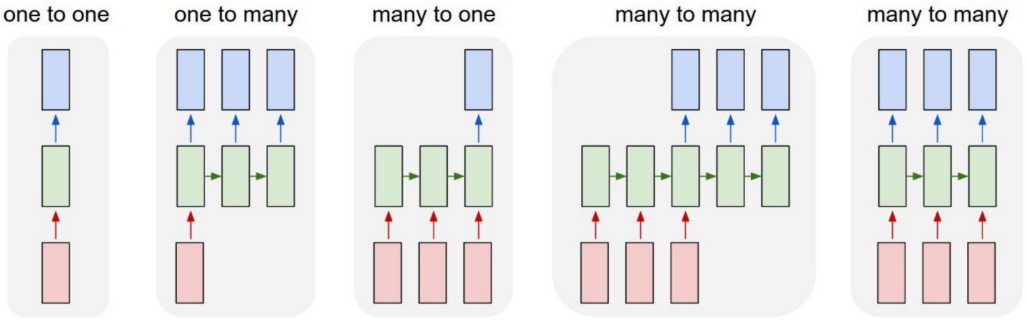


Basics of Sequence Networks



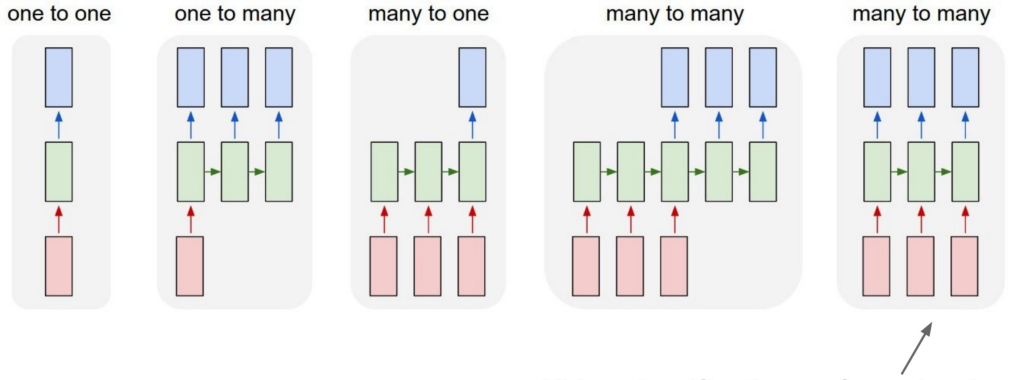
CLASP

Basics of Sequence Networks

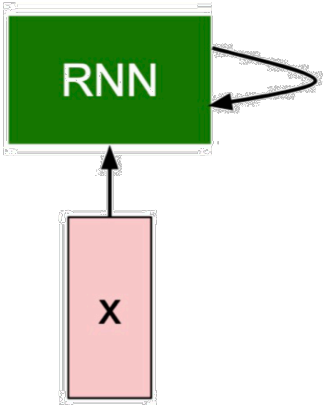


e.g. **Machine Translation**
seq of words -> seq of words

Basics of Sequence Networks

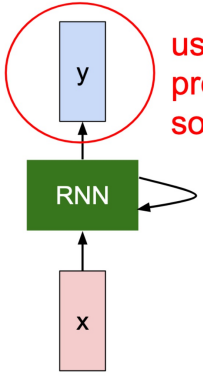


Recurrent Neural Network[2]



[2] http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture10.pdf

Recurrent Neural Network



usually want to predict a vector at some time steps

We can process a sequence of vectors **x** by applying a **recurrence formula** at every time step:

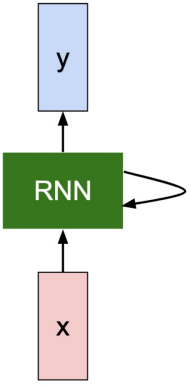
$$h_t = f_W(h_{t-1}, x_t)$$

new state

old state

input vector at some time step

some function with parameters W



Recurrent Neural Network: inner workings

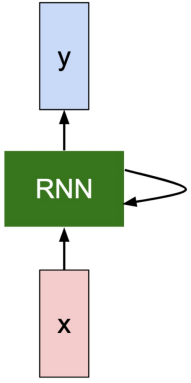
GÖTEBORGS UNIVERSITET

- You can think of hidden state as the memory of the RNN: it encapsulates information from all previous states into one single representations; during training, the model learns this state
- Intuition: in order to predict the next word, we need to take into account what has been stored in the "memory" of the model, e.g., in its hidden state
- Why do we need hidden state and why don't we use the model's output? The output state is a concatenation of all hidden states up to the step t, but to predict the output for the particular current state we use hidden state

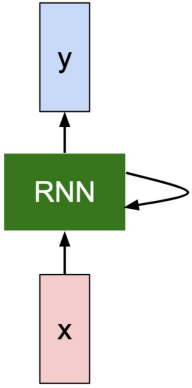
We can process a sequence of vectors **x** by applying a **recurrence formula** at every time step:

$$h_t = f_W(h_{t-1}, x_t)$$

Notice: the same function and the same set of parameters are used at every time step.



The state consists of a single *"hidden"* vector **h**:



$$h_t = f_W(h_{t-1}, x_t)$$

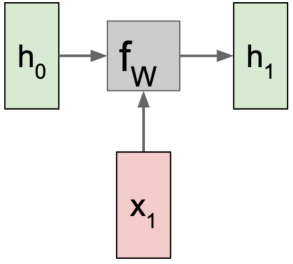
$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t)$$

$$y_t = W_{hy}h_t$$

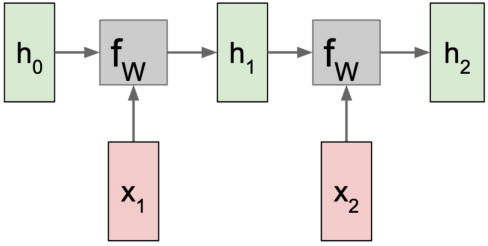
Why non-linearity?[3]

[3] https://stackoverflow.com/questions/9782071/why-must-a-nonlinear-activation-function-be-used-in-a-backpropagation-neural-net

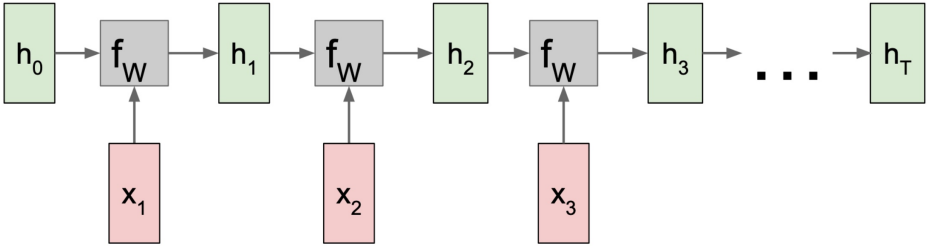
Recurrent Neural Network: Computational Graph



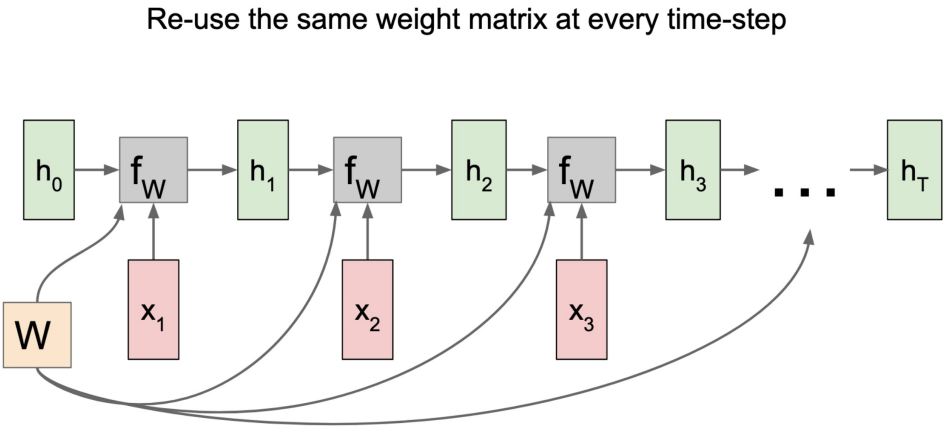
Recurrent Neural Network: Computational Graph



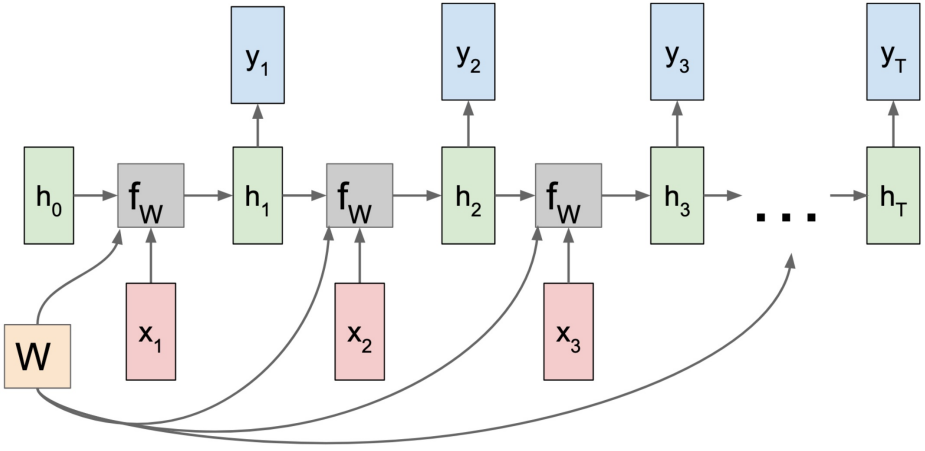
Recurrent Neural Network: Computational Graph



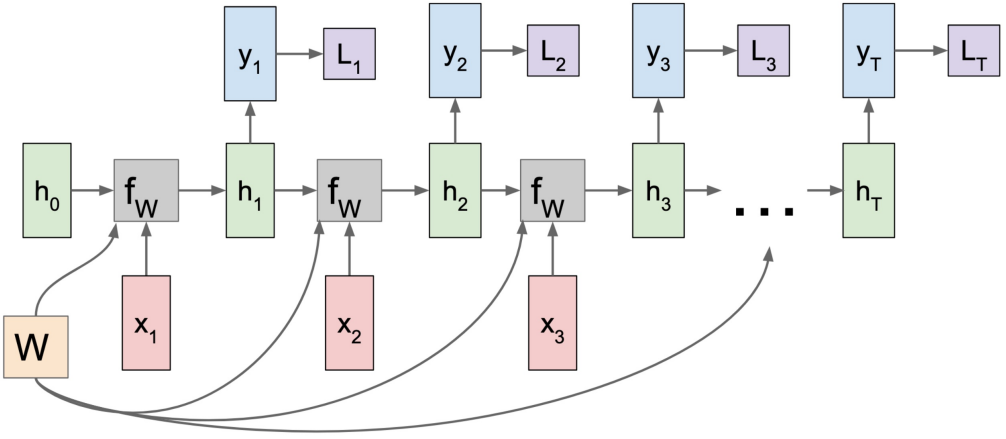
Recurrent Neural Network: Computational Graph



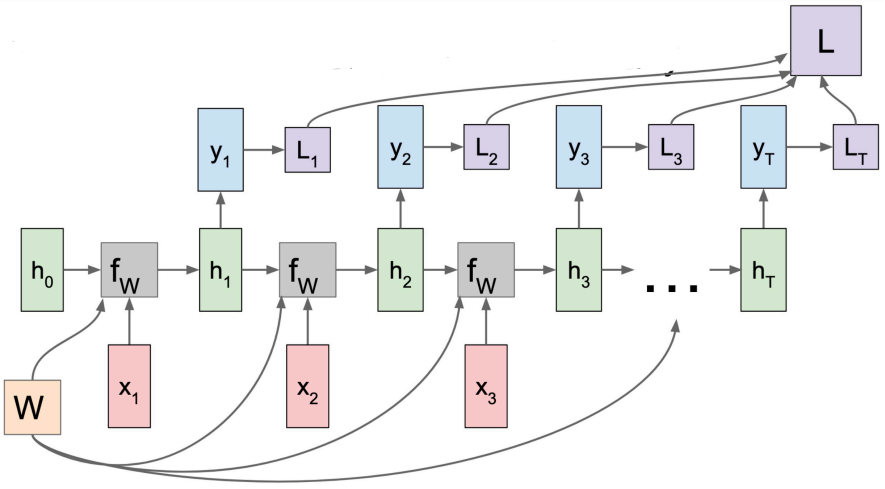
Recurrent Neural Network: Computational Graph



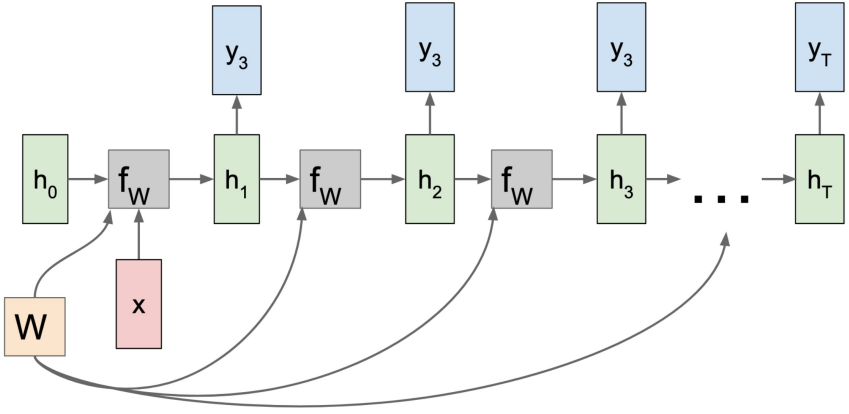
Recurrent Neural Network: Computational Graph



Recurrent Neural Network: Computational Graph



Recurrent Neural Network for Image Captioning

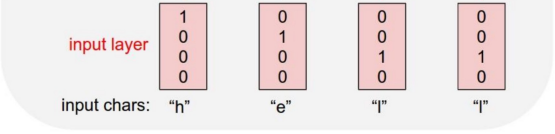


Example RNN: character-level generation

Example:
Character-level
Language Model

Vocabulary:
[h,e,l,o]

Example training sequence:
“hello”



CLASP

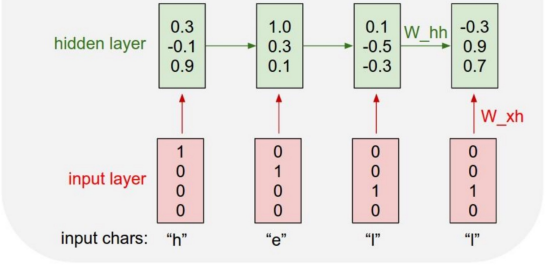
Example RNN: character-level generation

Example: Character-level Language Model

Vocabulary: [h,e,l,o]

Example training sequence: **“hello”**

$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t)$$



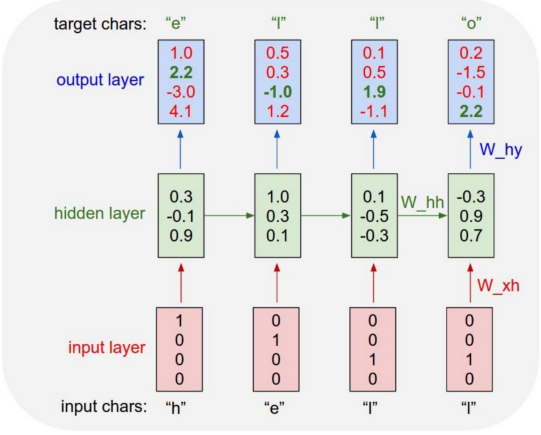
CLASP

Example RNN: character-level generation

Example: Character-level Language Model

Vocabulary: [h,e,l,o]

Example training sequence: **“hello”**

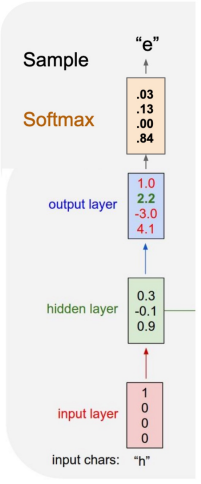


Example RNN: character-level generation

Example: Character-level Language Model Sampling

Vocabulary:
[h,e,l,o]

At test-time sample characters one at a time, feed back to model

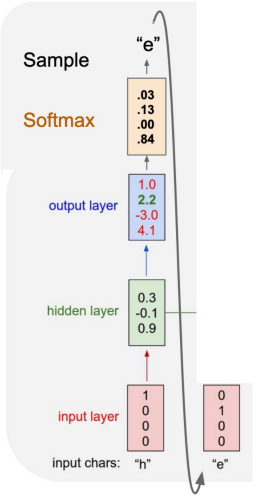


Example RNN: character-level generation

Example: Character-level Language Model Sampling

Vocabulary:
[h,e,l,o]

At test-time sample characters one at a time, feed back to model

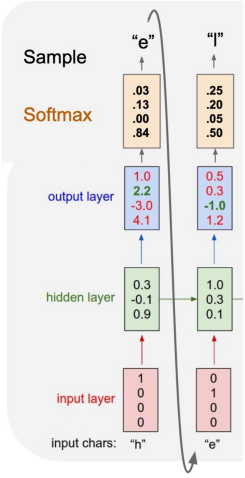


Example RNN: character-level generation

Example: Character-level Language Model Sampling

Vocabulary:
[h,e,l,o]

At test-time sample characters one at a time, feed back to model



Example RNN: character-level generation

Example: Character-level Language Model Sampling

Vocabulary:
[h,e,l,o]

At test-time sample characters one at a time, feed back to model

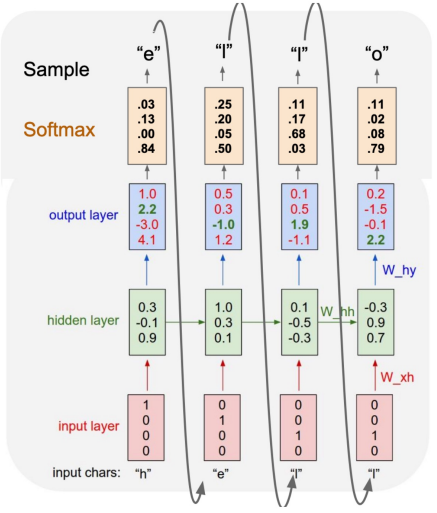
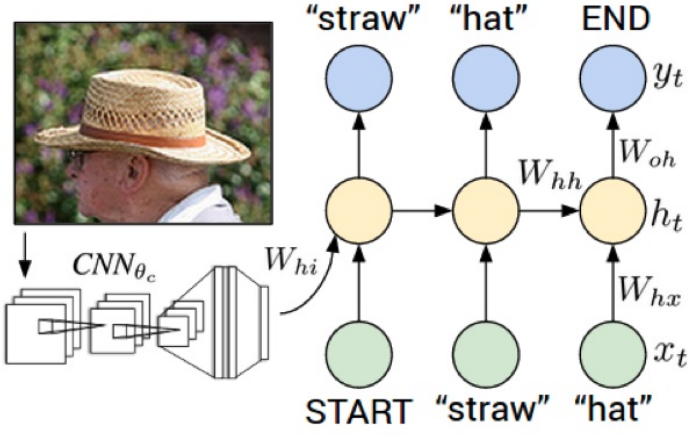


Image Captioning

Classic literature on neural image captioning:
Mao et al. (2014); Karpathy and Fei-Fei (2015); Vinyals et al. (2015); Donahue et al. (2016); Chen and Zitnick (2014)



CLASP

Image Captioning: how do we do it?

This image is CC0 public domain

test image

Image Captioning: how do we do it?

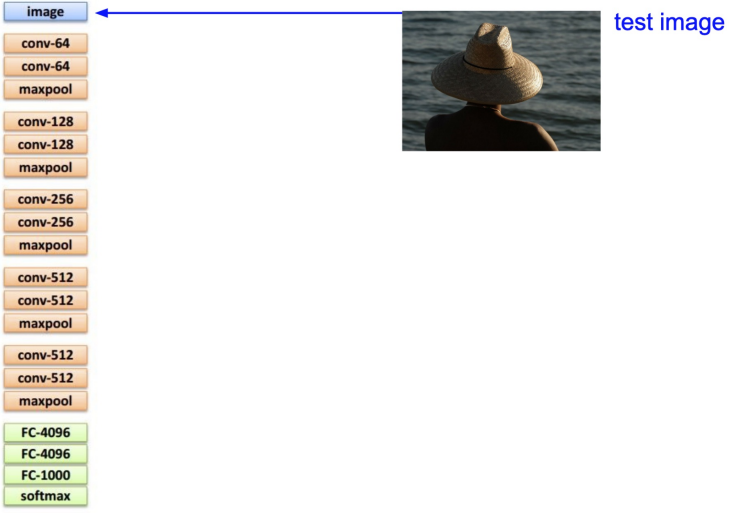
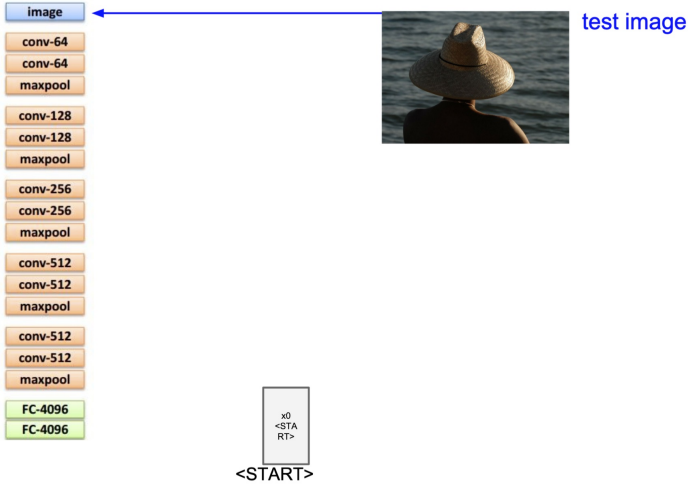


Image Captioning: how do we do it?

test image

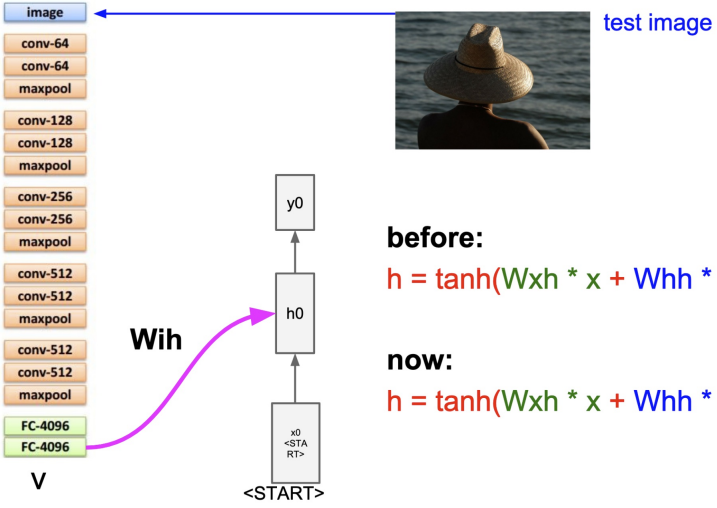
CLASP

Image Captioning: how do we do it?



CLASP

Image Captioning: how do we do it?



before:

h = tanh(Wxh * x + Whh * h)

now:

h = tanh(Wxh * x + Whh * h + **Wih * v**)

CLASP

Image Captioning: how do we do it?

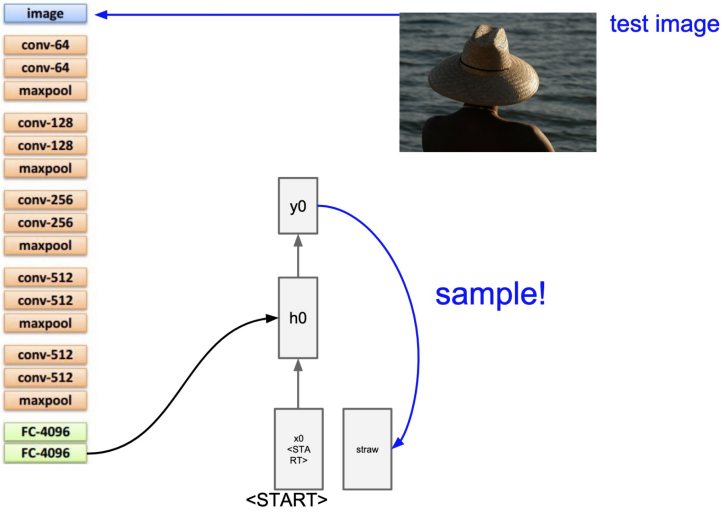


Image Captioning: how do we do it?

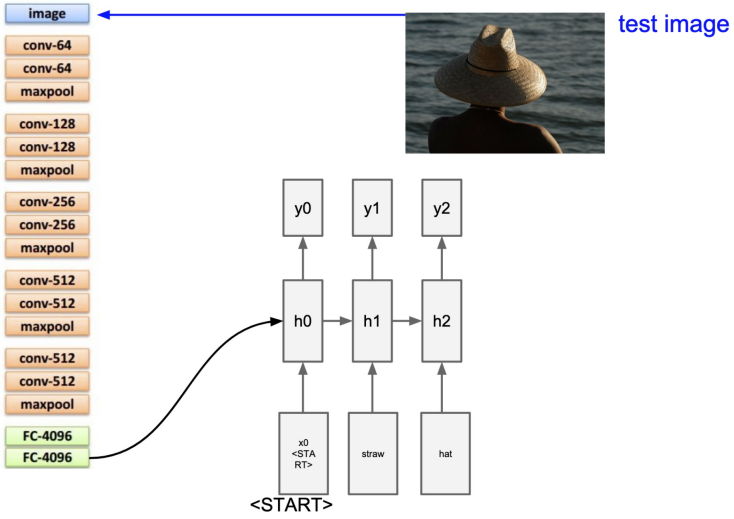
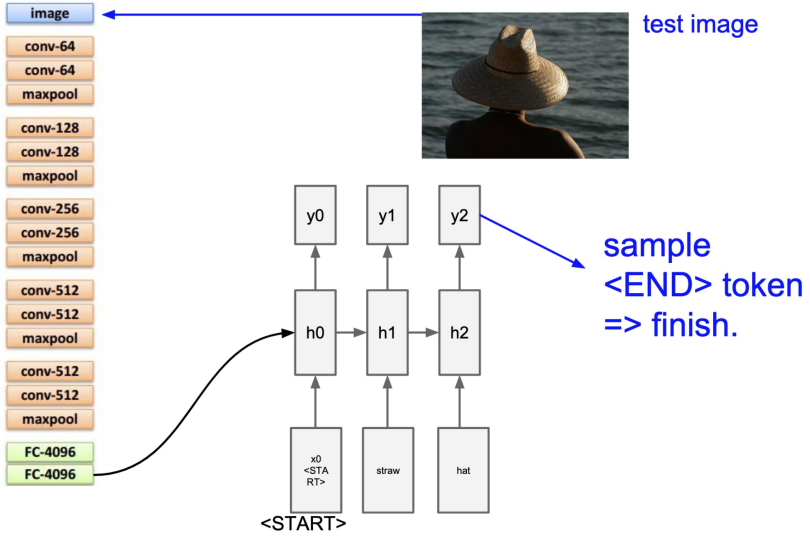


Image Captioning: how do we do it?



CLASP

Image Captioning: good examples

A cat sitting on a suitcase on the floor

A cat is sitting on a tree branch

A dog is running in the grass with a frisbee

A white teddy bear sitting in the grass

Two people walking on the beach with surfboards

A tennis player in action on the court

Two giraffes standing in a grassy field

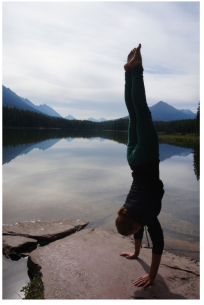
A man riding a dirt bike on a dirt track

CLASP

Image Captioning: bad examples

A woman is holding a cat in her hand

A person holding a computer mouse on a desk



A woman standing on a beach holding a surfboard

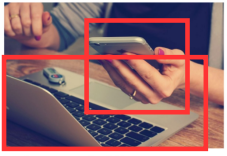
A bird is perched on a tree branch

A man in a baseball uniform throwing a ball

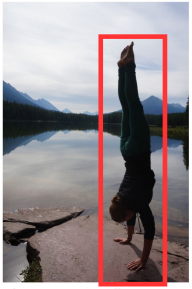
CLASP

Why would the model make such mistakes?

A woman is holding a cat in her hand



A person holding a computer mouse on a desk



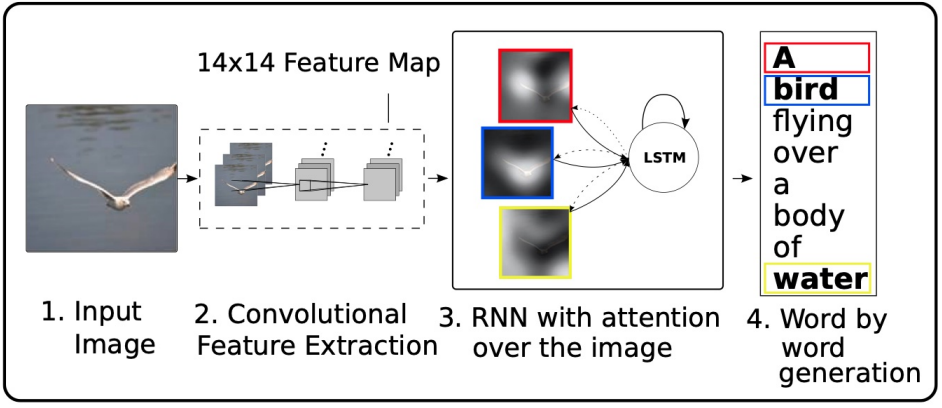
A woman standing on a beach holding a surfboard

A bird is perched on a tree branch

A man in a baseball uniform throwing a ball

- frequency of concepts co-occurring together (e.g., computer mouse - computer desk, tree - bird, cat - woman)
- each image is unique, so **more focus on important parts of the image** might help the model to correctly describe the image

Image Captioning with Attention (Xu et al., 2016)

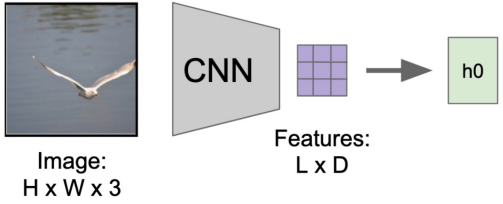


Why do we need attention and what is so interesting about it?

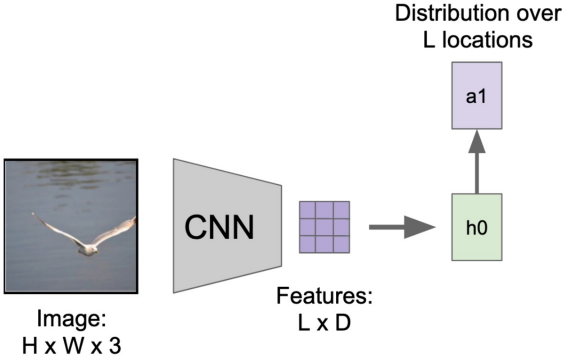
GÖTEBORGS UNIVERSITET

- prior to language-and-vision tasks, attention has been initially introduced in the domain of machine translation (Bahdanau et al., 2016)
- the main idea is that **every word that we say is grounded in the image and we want our model to correctly ground words into objects**
- it is hard to ground relations (e.g., flying) and both language and vision modalities can be more important for grounding of particular word types (Lu et al., 2017; Ghanimifard and Dobnik, 2019; Ilinykh and Dobnik, 2021)
- *project idea sketch*: examine grounding of (spatial) relations in multi-modal set-up

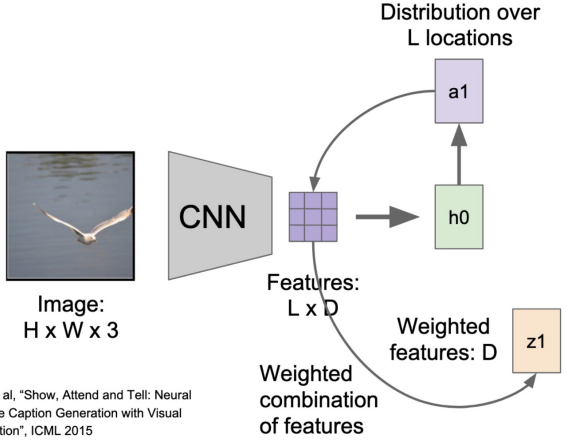
Attention in ICM



Attention in ICM



Attention in ICM

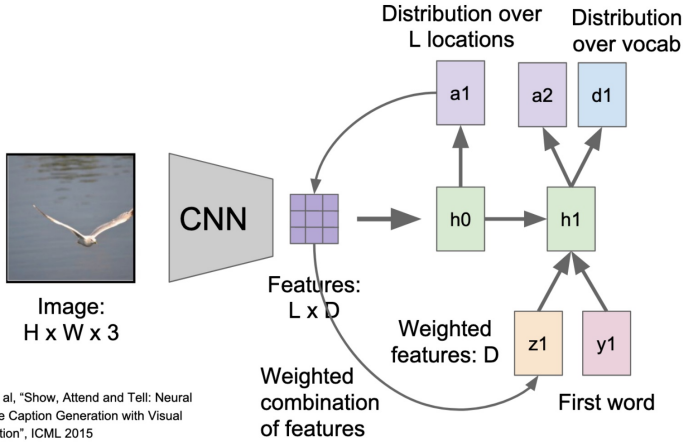


Xu et al, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention", ICML 2015

$$z = \sum_{i=1}^{L} p_i v_i$$

CLASP

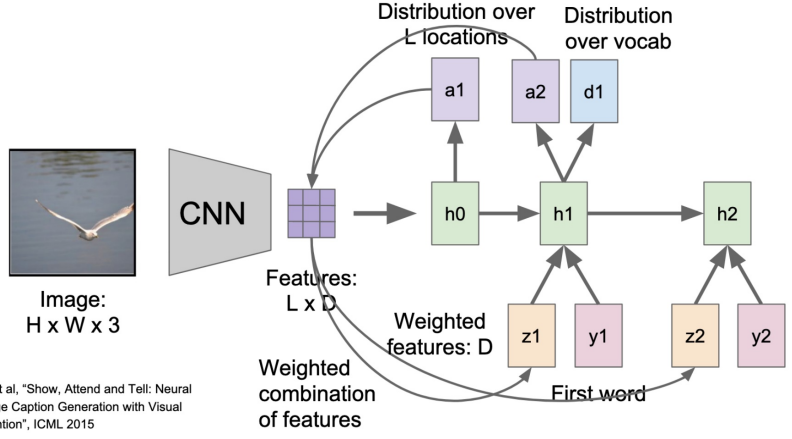
Attention in ICM



:u et al, "Show, Attend and Tell: Neural
nage Caption Generation with Visual
ttention", ICML 2015

CLASP

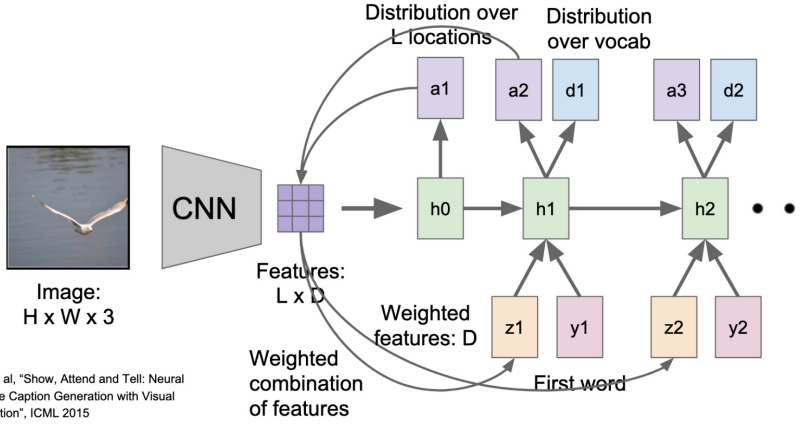
Attention in ICM



Xu et al, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention", ICML 2015

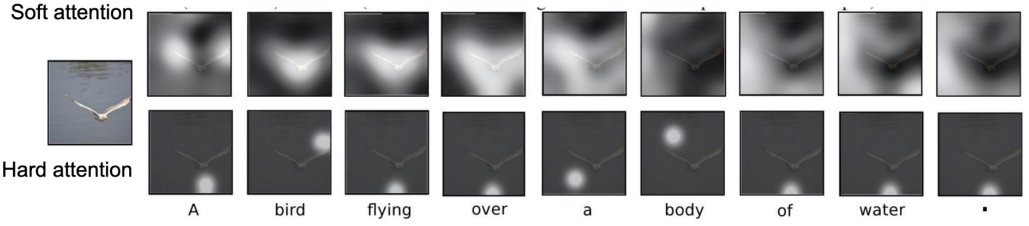
CLASP

Attention in ICM



Xu et al, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention", ICML 2015

Attention in ICM

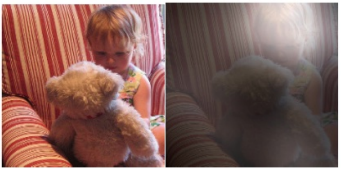


Attention in ICM

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.



A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

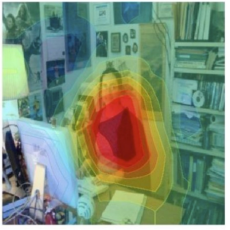
Attention helps with gender bias in captioning models

GÖTEBORGS UNIVERSITET

Wrong

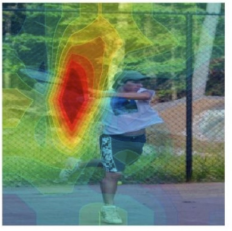
Baseline:
*A **man** sitting at a desk with a laptop computer.*

Right for the Right Reasons



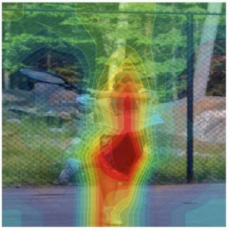
Our Model:
*A **woman** sitting in front of a laptop computer.*

Right for the Wrong Reasons



Baseline:
*A **man** holding a tennis racquet on a tennis court.*

Right for the Right Reasons



Our Model:
*A **man** holding a tennis racquet on a tennis court.*

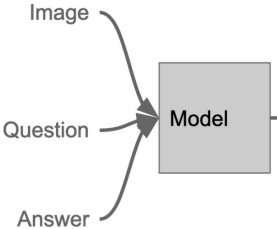
Burns et al. "Women also Snowboard: Overcoming Bias in Captioning Models" ECCV 2018
Figures from Burns et al, copyright 2018. Reproduced with permission.

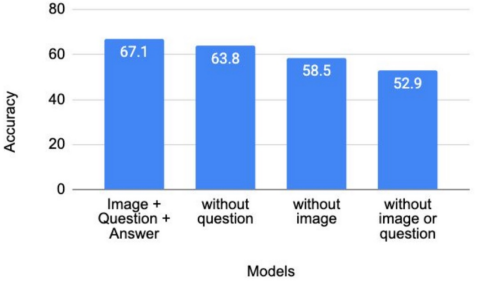
CLASP

Attention helps with dataset biases

What is the dog playing with?

Frisbee





Jabri et al. "Revisiting Visual Question Answering Baselines" ECCV 2016

CLASP

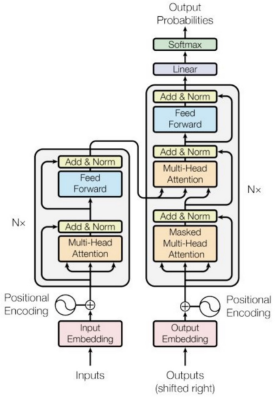
GÖTEBORGS UNIVERSITET

Recently in Natural Language Processing… New paradigms for reasoning over sequences

[“Attention is all you need”, Vaswani et al., 2018]

- New “Transformer” architecture no longer processes inputs sequentially; instead it can operate over inputs in a sequence in parallel through an attention mechanism

- Has led to many state-of-the-art results and pre-training in NLP, for more interest see e.g.
  - “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, Devlin et al., 2018
  - OpenAI GPT-2, Radford et al., 2018



CLASP

Conclusions and Future Content

GÖTEBORGS UNIVERSITET

- image captioning is the core task in language-and-vision field
- standard neural architecture: CNN + LSTM (encoder-decoder, seq-to-seq)
- SOTA: transformers (all inputs are processed simultaneously)
- *Next tutorials*: we will look at decoding, evaluation of captioning models, different modes of image representation, more complex tasks (e.g., visual question answering, visual dialogue, situated interaction)
- **but now**, we are going to code!

Before we start coding...

- An image is worth a thousand words
- How to choose the best way to describe an image? And should we even choose one?

Before we start coding...

GÖTEBORGS UNIVERSITET

- Each of us describes what we see based on our personal individual background
- How do we account for that?

Before we start coding...

GÖTEBORGS UNIVERSITET

- What do we want to describe in the image? Which objects are important or not important?
- When do we want to describe objects? Is there some order of objects that we use?
- How to describe objects? Which attributes we want to mention and which relations?

References I

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate.

Xinlei Chen and C. Lawrence Zitnick. 2014. Learning a recurrent visual representation for image caption generation.

Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Trevor Darrell. 2016. Long-term recurrent convolutional networks for visual recognition and description.

Mehdi Ghanimifard and Simon Dobnik. 2019. What goes into a word: generating image descriptions with top-down spatial knowledge. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 540–551, Tokyo, Japan. Association for Computational Linguistics.

Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. 2020. Captioning images taken by people who are blind. In *ECCV*.

Nikolai Ilinykh and Simon Dobnik. 2021. How vision affects language: Comparing masked self-attention in uni-modal and multi-modal transformer. In *Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR)*, pages 45–55, Groningen, Netherlands (Online). Association for Computational Linguistics.

References II

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3242–3250.

Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. 2014. Explain images with multimodal recurrent neural networks.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France. PMLR.

References III

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2016. Show, attend and tell: Neural image caption generation with visual attention.