

Learning to Describe Multimedia

Kyunghyun Cho

New York University

Courant Institute (Computer Science) and Center for Data Science

Facebook AI Research

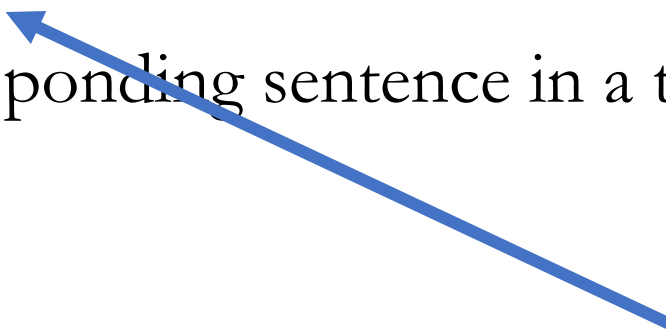
Kyunghyun Cho, Aaron Courville, Yoshua Bengio.

Describing multimedia content using attention-based encoder-decoder networks.

IEEE Transactions on Multimedia. 2015.

Machine Translation

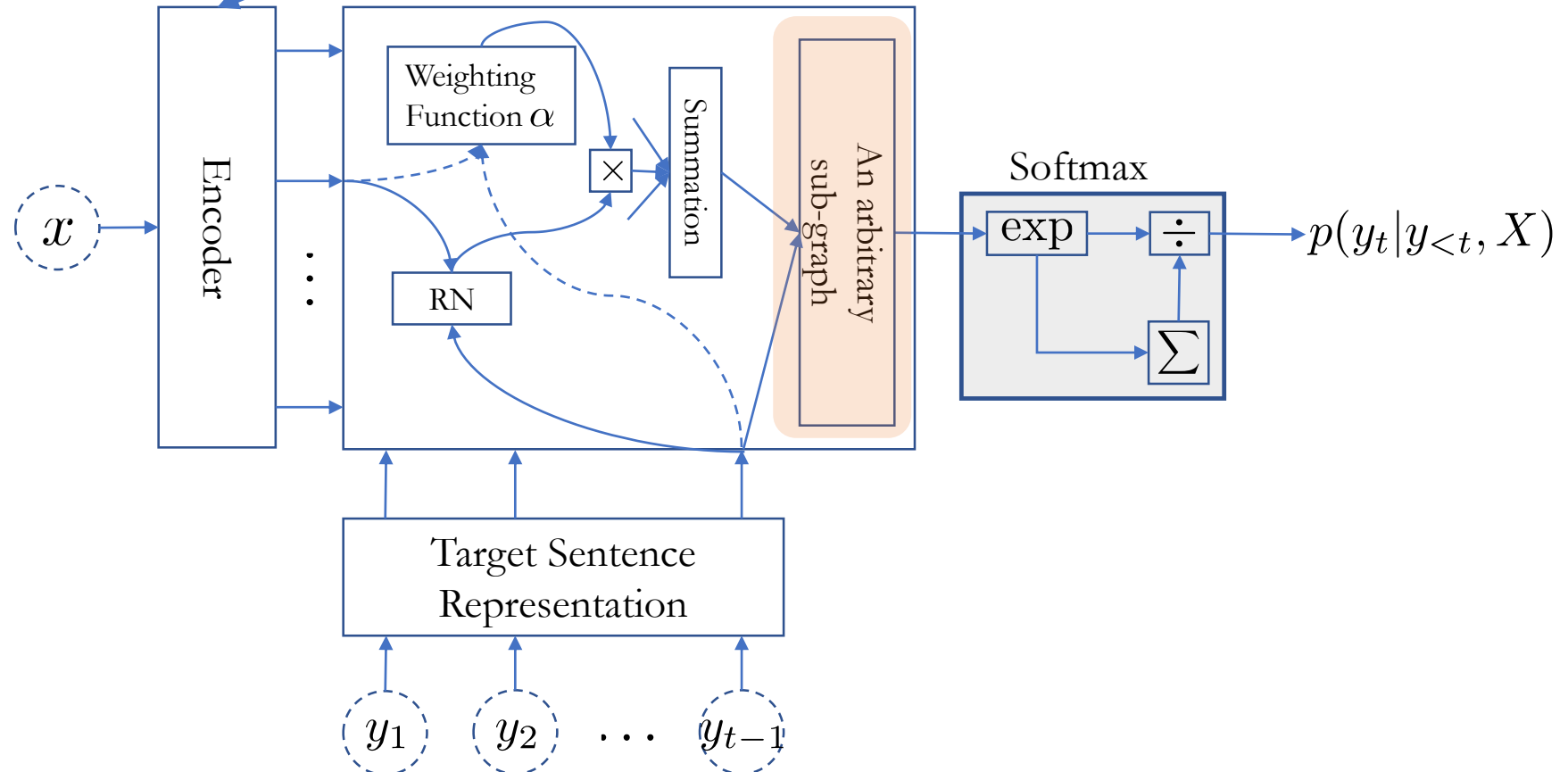
- Input: **a sentence written in a source language**
- Output: a corresponding sentence in a target language



Is it necessary for the source to be a natural language sentence?

Description Generation

- Input: arbitrary as long as encoded into a set of continuous vectors
- Output: a corresponding sentence in a target language



Description Generation

- Encouraged by the success of neural machine translation, a lot of new applications were tried in 2015-2016:
 - Image caption generation
 - Video description generation
 - Speech recognition
 - And many others.
- In most of these tasks, the attention-based encoder-decoder has since become *de facto* standard: see Lecture 4.

Image Caption Generation [Xu et al., 2015]

- Input: an image
- Output: an image caption
- Network Architecture
 - Encoder: deep convolution network
 - Decoder: recurrent language model with the attention mechanism.
- Data: image-caption pairs

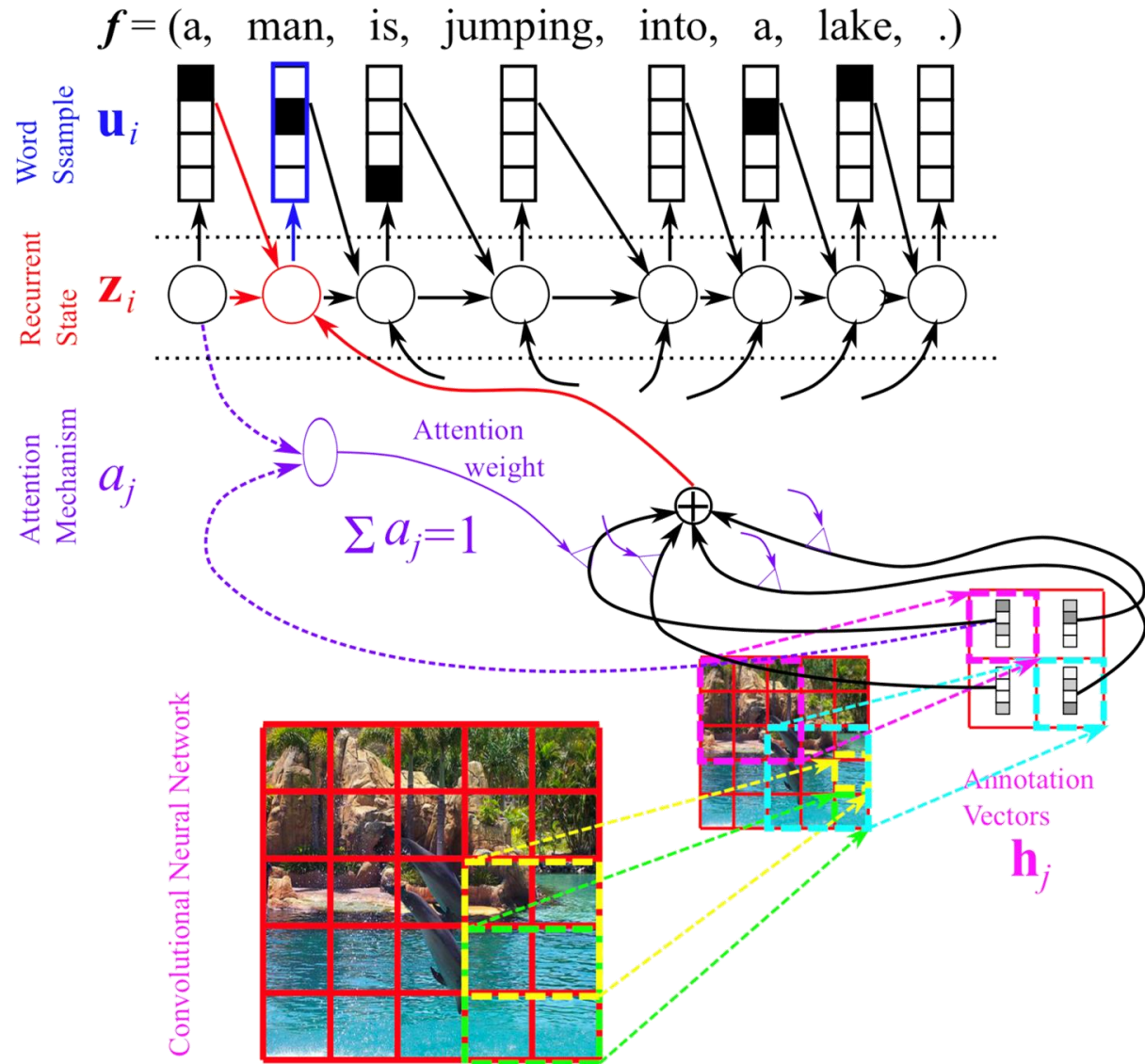


Image Caption Generation



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.

Video Description Generation [Li et al., 2015]

- Input: a short video clip - a sequence of video frames.
- Output: a corresponding description
- Network Architecture
 - Encoder: a deep 2+3D convolutional network
 1. A 2-D convolutional network for each frame
 2. A 3-D convolutional network for the entire clip
 - Decoder: recurrent language model with the attention mechanism.
- Data: clip-description pairs – collected from YouTube.

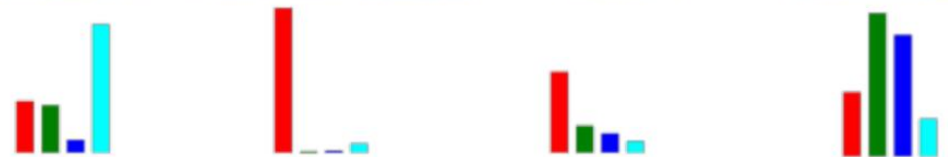
Video Description Generation

- Input: a short video clip - a sequence of video frames.
- Output: a corresponding description
- Attention allows us to inspect the inner-working of the model.
- Some encouraging result in 2015, and a lot of advances have been proposed since then.



+Local+Global: A **man** and a **woman** are **talking** on the **road**

Ref: A man and a woman ride a motorcycle

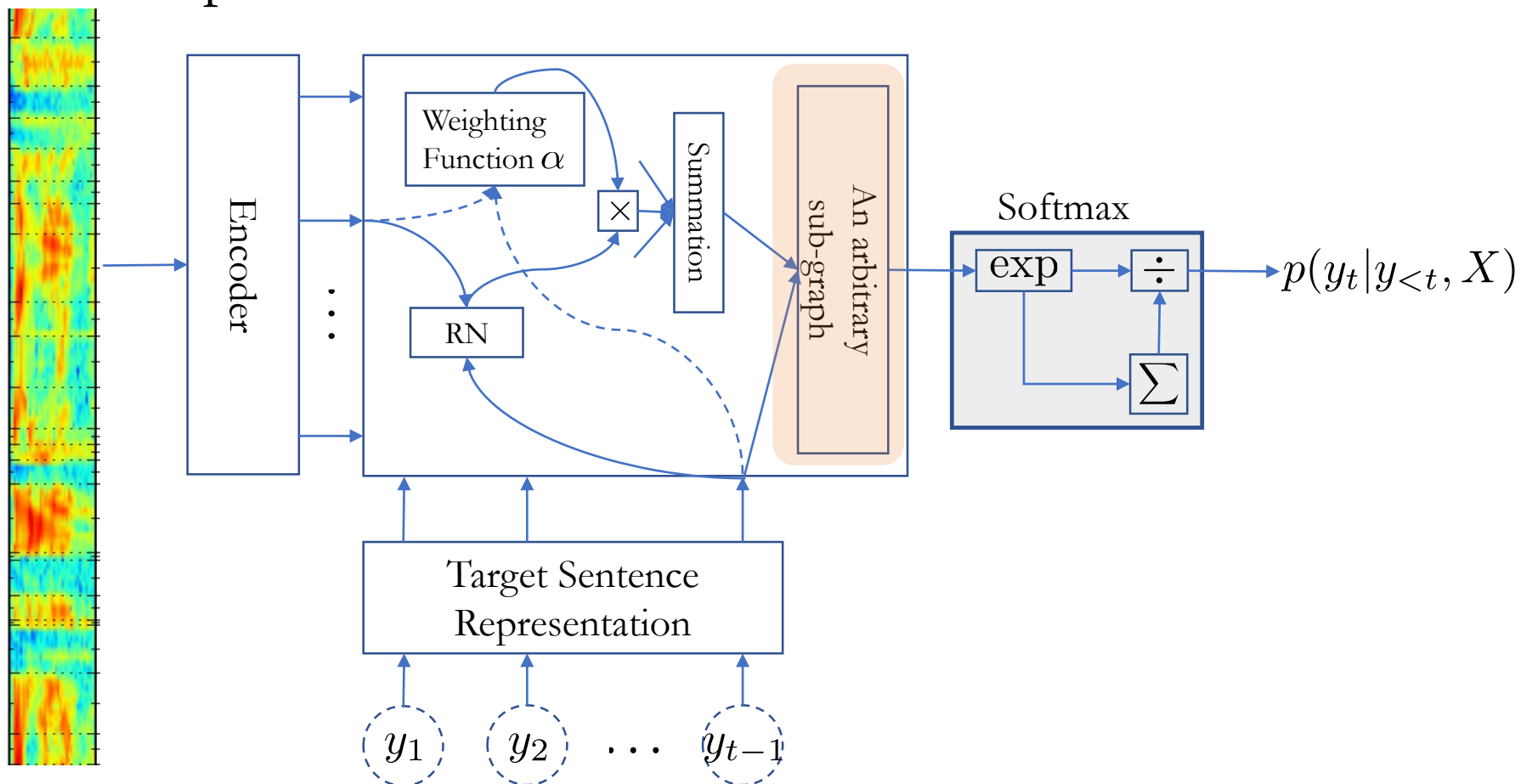


+Local+Global: **Someone** is **frying** a **fish** in a **pot**

Ref: A woman is frying food

Speech Recognition [Chorowski et al., 2015]

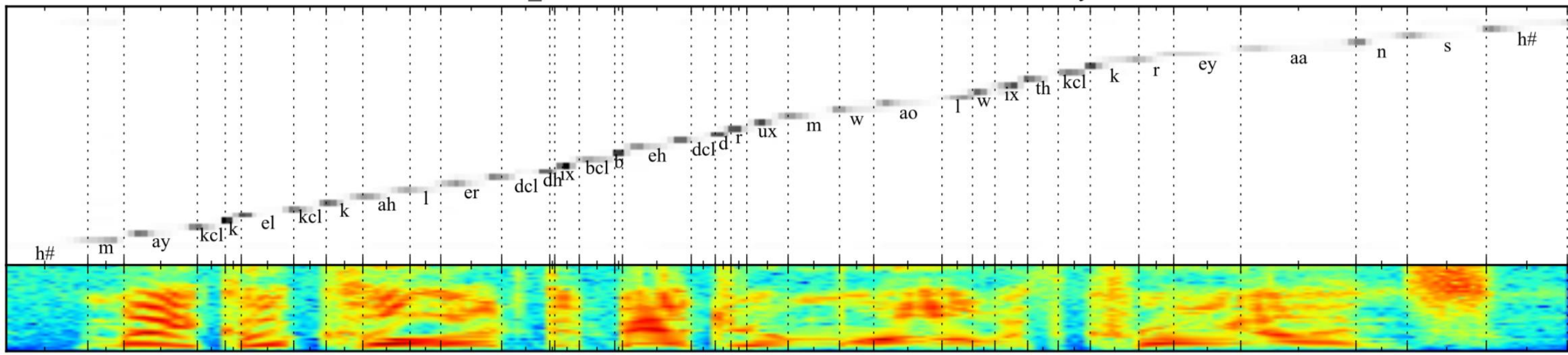
- Input: Speech
- Output: transcription



Speech Recognition

- Input: Speech
- Output: transcription
- Network Architecture
 - Encoder: convolution+recurrent acoustic network
 - Decoder: conditional recurrent language model + attention mechanism

FDHC0_SX209: Michael colored the bedroom wall with crayons.



Since 2015...

- The attention (alignment) mechanism has become a work horse behind various AI models/applications including
 - Neural Turing machines (differentiable neural computer) [Graves et al., 2015&2016], memory networks [Weston et al., 2015; Sukhbaatar et al., 2016], dynamic neural Turing machines [Gulcehre et al., 2017; Miller et al., 2017], ...
 - Reinforcement learning: attentive history selection [Tian et al., 2016], neural episodic control [Pritzel et al., 2017]
 - Generative models: DRAW [Gregor et al., 2016], Image Transformer [Parmar et al., 2018], ...
- Many more on the horizon...