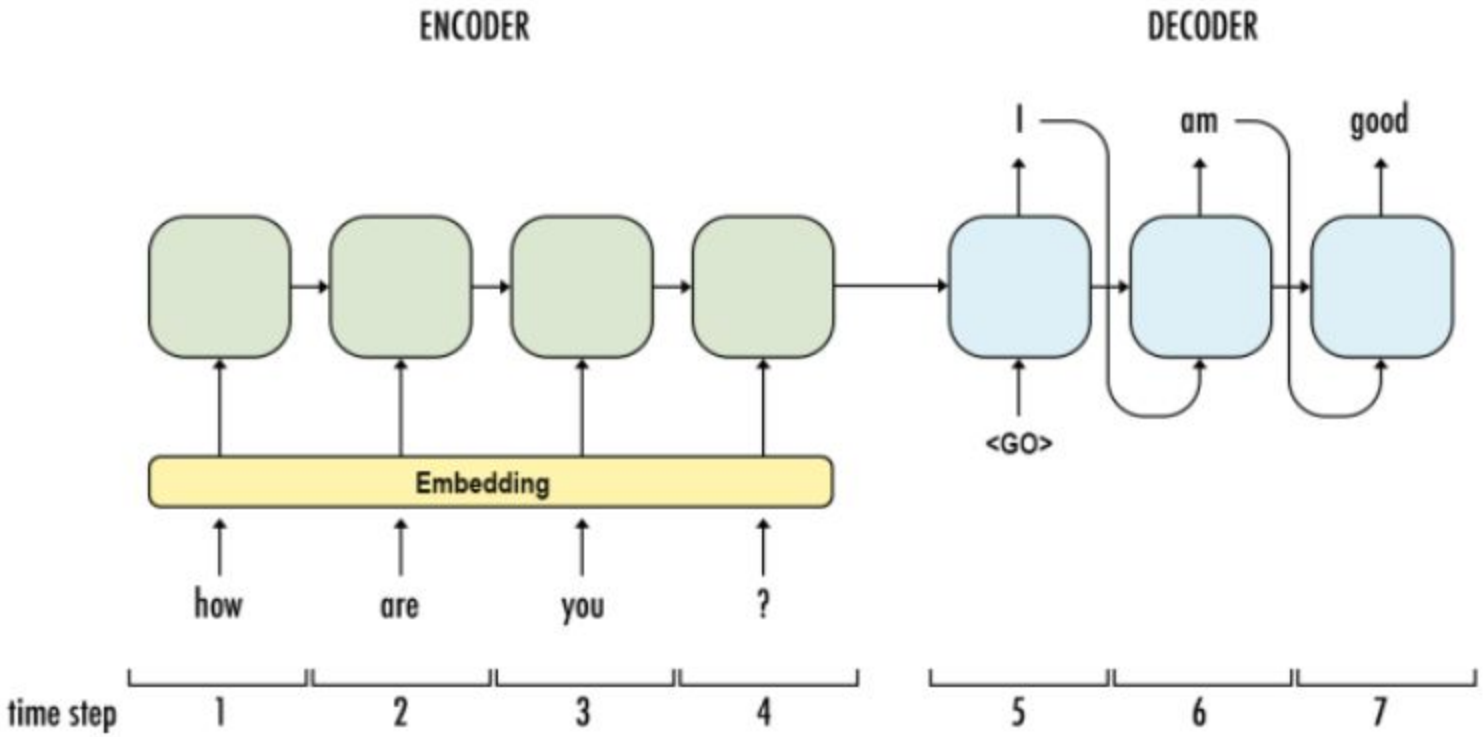
Image Captioning
簡介

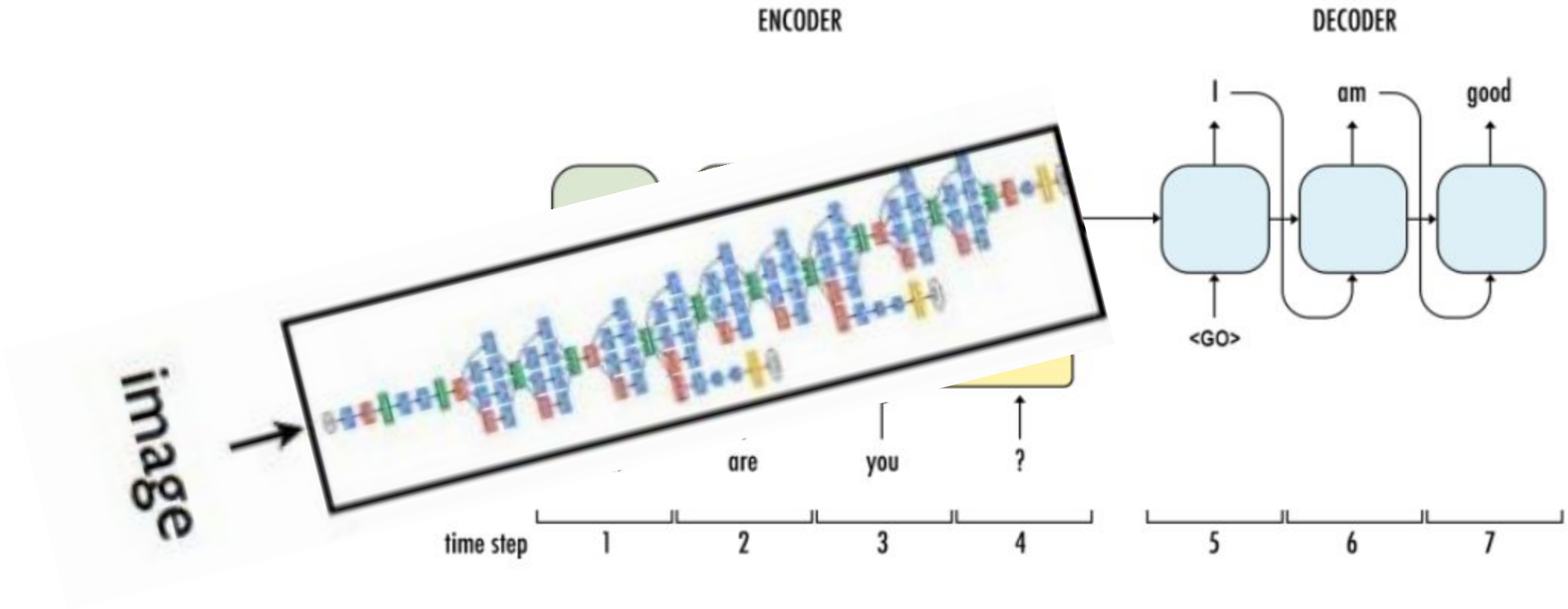
Department of Computer Science, NCTU

TA Ziv(鍾嘉峻)

Encoder-Decoder



Encoder-Decoder



Show and Tell: A Neural Image Caption Generator

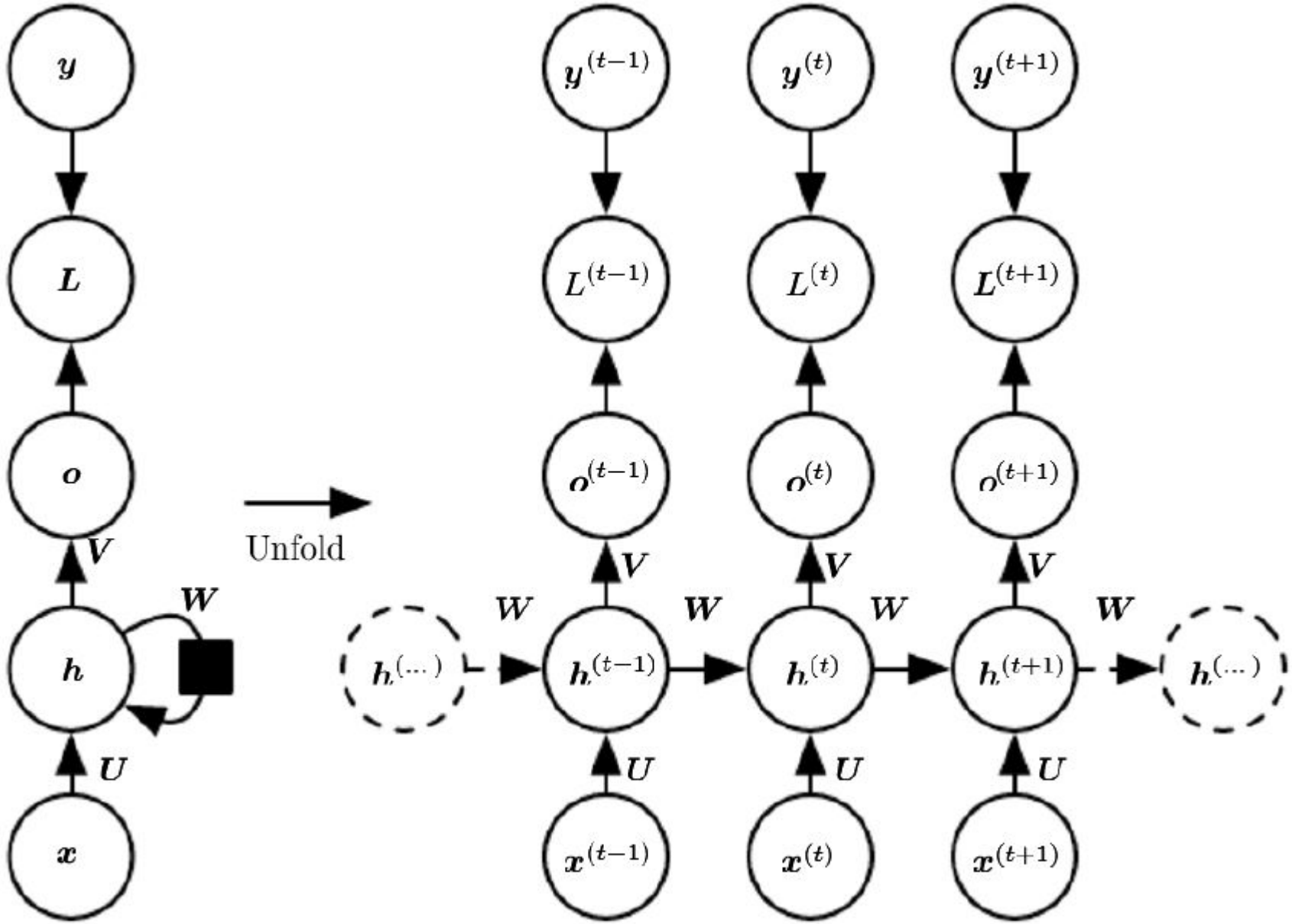
Sequence-to-Sequence

- Basic RNN concepts:

$$\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, ..., \boldsymbol{x}^{(\tau)}$$

$$\boldsymbol{h}^{(t)} = f_{\boldsymbol{\theta}}(\boldsymbol{h}^{(t-1)}, \boldsymbol{x}^{(t)})$$

$$\begin{aligned}
\boldsymbol{h}^{(t)} &= f_{\boldsymbol{\theta}}(f_{\boldsymbol{\theta}}(\boldsymbol{h}^{(t-2)}, \boldsymbol{x}^{(t-1)}), \boldsymbol{x}^{(t)}) \\
&= f_{\boldsymbol{\theta}}(f_{\boldsymbol{\theta}}(\ldots f_{\boldsymbol{\theta}}(\boldsymbol{h}^{(0)}, \boldsymbol{x}^{(1)}), \boldsymbol{x}^{(2)}), \ldots, \boldsymbol{x}^{(t-1)}), \boldsymbol{x}^{(t)}) \\
&= g^{(t)}(\boldsymbol{h}^{(0)}, \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(t)})
\end{aligned}$$


y
L
o
V
W
h
U
x
Unfold
$y^{(t-1)}$
$y^{(t)}$
$y^{(t+1)}$
$L^{(t-1)}$
$L^{(t)}$
$L^{(t+1)}$
$o^{(t-1)}$
$o^{(t)}$
$o^{(t+1)}$
$h^{(\dots)}$
$h^{(t-1)}$
$h^{(t)}$
$h^{(t+1)}$
$h^{(\dots)}$
$x^{(t-1)}$
$x^{(t)}$
$x^{(t+1)}$

Sequence-to-Sequence

- Basic RNN concepts:

$$\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(\tau)} \longrightarrow \boldsymbol{y}^{(1)}, \boldsymbol{y}^{(2)}, \ldots, \boldsymbol{y}^{(\tau)}$$

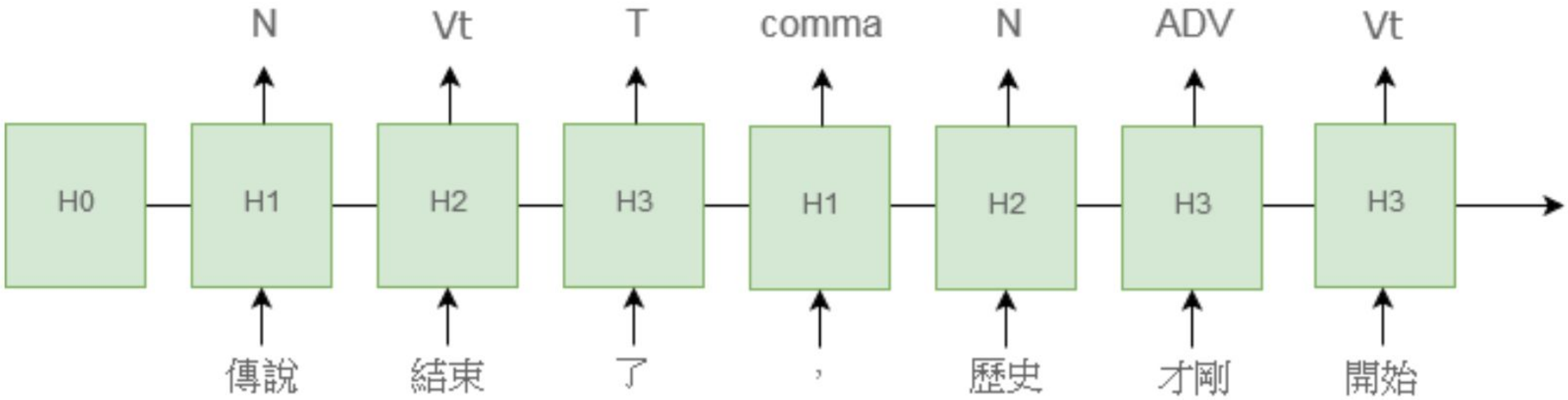
$$\boldsymbol{a}^{(t)} = \boldsymbol{b} + \boldsymbol{W}\boldsymbol{h}^{(t-1)} + \boldsymbol{U}\boldsymbol{x}^{(t)},$$

$$\boldsymbol{h}^{(t)} = \tanh(\boldsymbol{a}^{(t)}),$$

$$\boldsymbol{o}^{(t)} = \boldsymbol{c} + \boldsymbol{V}\boldsymbol{h}^{(t)},$$

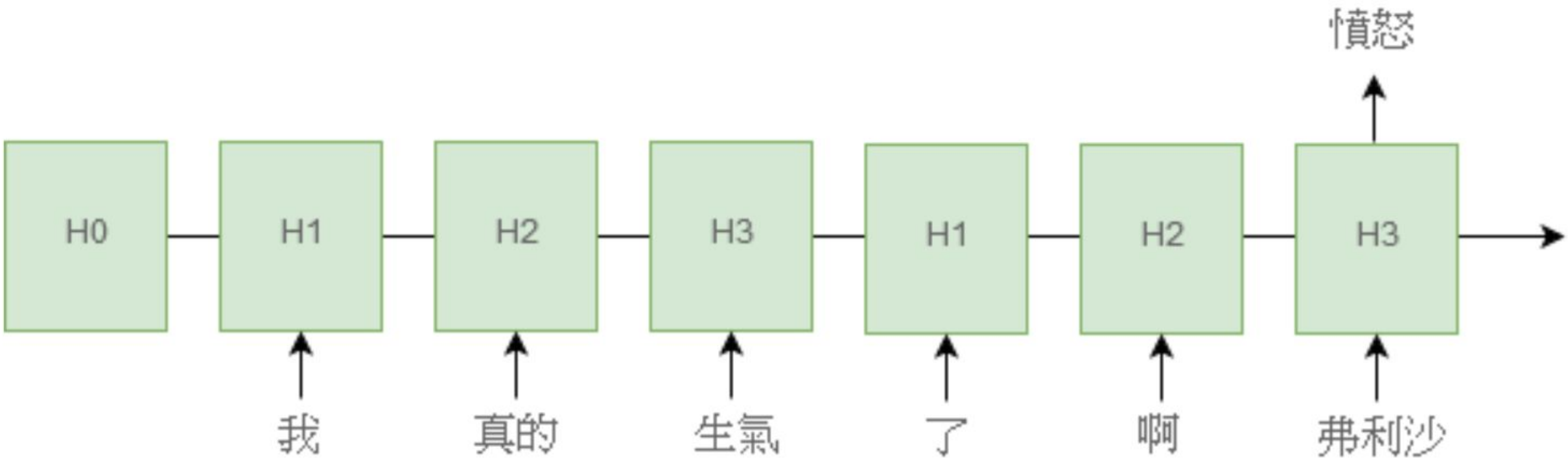
Basic RNN Application

- Label a sentence



Basic RNN Application

- Classification



Basic RNN Application

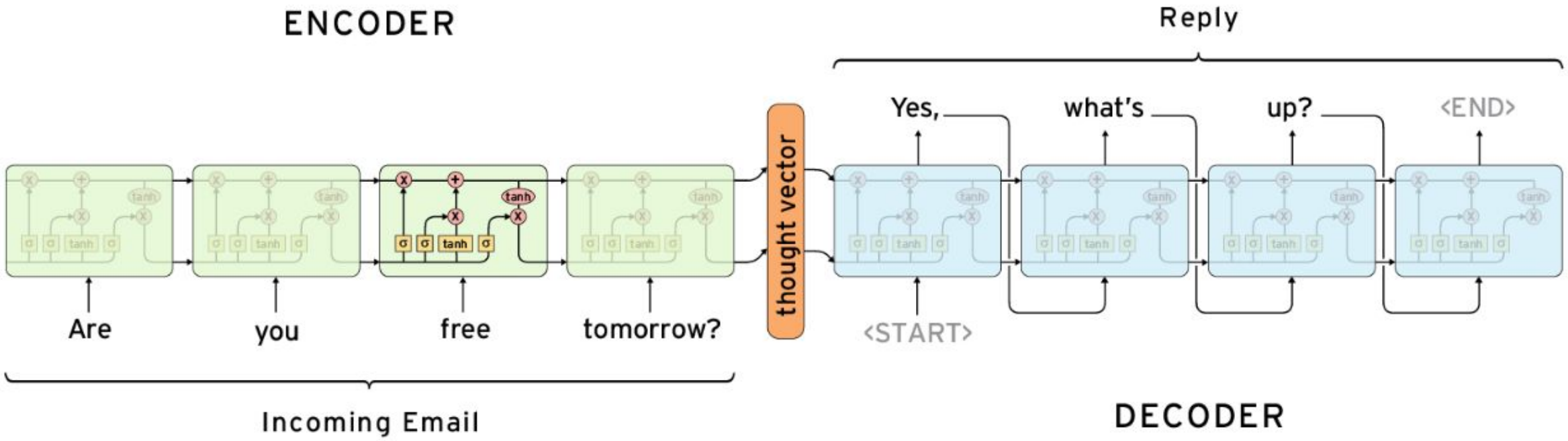
- How about translator ?

- Fine
  - English : How are you?
  - 中文 : 你好嗎?
  - Same length!

- Some …. trouble?
  - English : Elegance is the beauty that never fades.
  - 中文 : 優雅是唯一不會褪色的美。
  - Different length QQQQQQQQ

Sequence-to-Sequence

- Use two RNN models



How about decoder??

- We know that we can use a rnn as a encoder to encode something to a context vector

- How about decoder??
  - We still have one input (a context vector) QQ
  - while True:

```
output = decoder(output)
outputs.append(output)
```

How about decoder??

- Use “while” ??
  - When we stop??

- Using a symbol to let this loop stop
  - We can use “EOS” (End Of Sentence)
  - while output != 'EOS':

```
output = decoder(output)
outputs.append(output)
```

Sentence-to-sentence

- We use M-length as input and get 1 vector
- We use this 1 vector as input and get N-length output
- Appication?
  - Translator
  - Chat Bot
  - Summarizer
  - Poet
  - …..

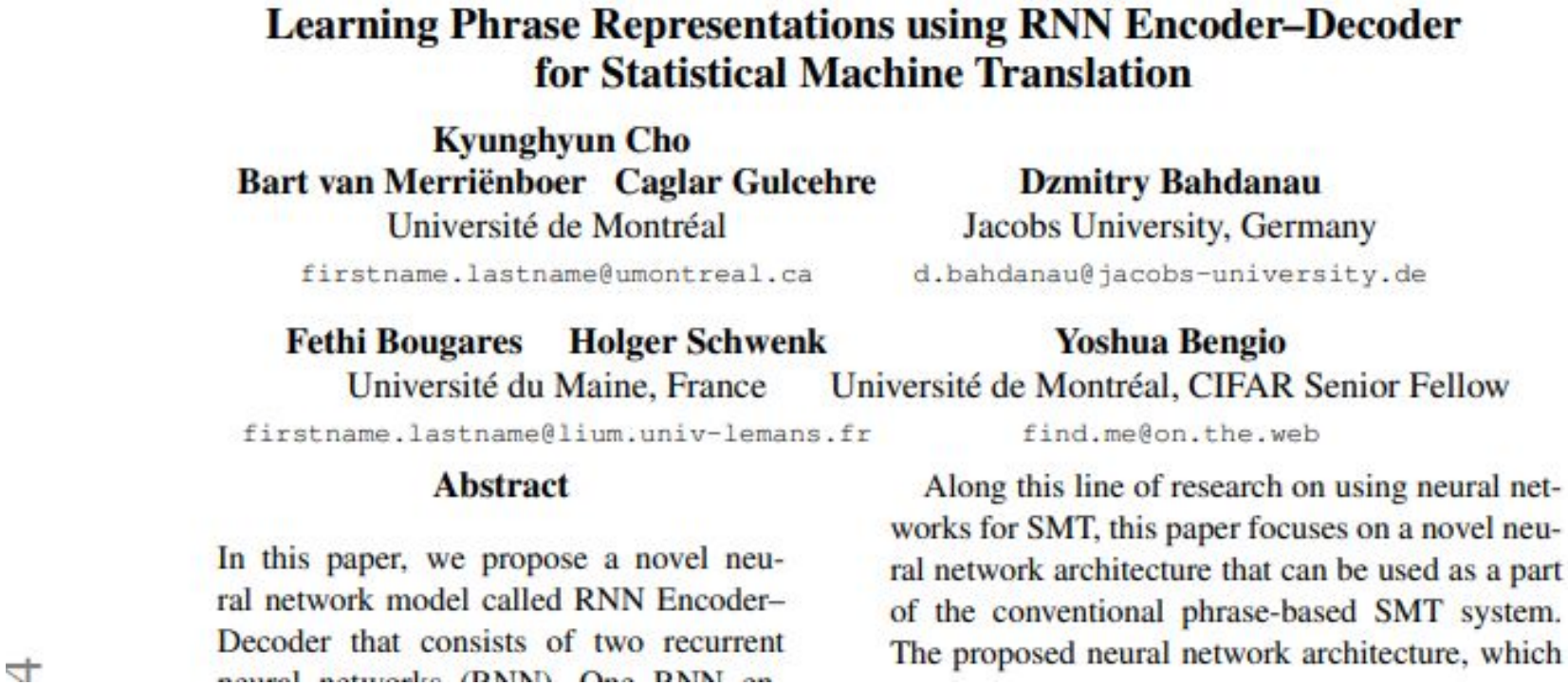
Sentence-to-sentence

- Can we use other architecture as “Encoder” ??
  - Not only use “RNN-based” architecture

- Yes, of course XDDD
  - Or what is the lab3 for XDD

Paper Reference--Sequence-to-sequence

- Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation
  - https://arxiv.org/pdf/1406.1078.pdf



Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation

Kyunghyun Cho
Bart van Merriënboer **Caglar Gulcehre**
Université de Montréal
firstname.lastname@umontreal.ca

Dzmitry Bahdanau
Jacobs University, Germany
d.bahdanau@jacobs-university.de

Fethi Bougares **Holger Schwenk**
Université du Maine, France
firstname.lastname@lium.univ-lemans.fr

Yoshua Bengio
Université de Montréal, CIFAR Senior Fellow
find.me@on.the.web

Abstract

In this paper, we propose a novel neural network model called RNN Encoder–Decoder that consists of two recurrent

Along this line of research on using neural networks for SMT, this paper focuses on a novel neural network architecture that can be used as a part of the conventional phrase-based SMT system. The proposed neural network architecture, which

Paper Reference--Image Caption

- Show and Tell: A Neural Image Caption Generator
  - https://arxiv.org/abs/1411.4555

Show and Tell: A Neural Image Caption Generator

Oriol Vinyals
Google
vinyals@google.com

Alexander Toshev
Google
toshev@google.com

Samy Bengio
Google
bengio@google.com

Dumitru Erhan
Google
dumitru@google.com

Abstract

Automatically describing the content of an image is a fundamental problem in artificial intelligence that connects computer vision and natural language processing. In this paper, we present a generative model based on a deep recurrent architecture that combines recent advances in computer vision and machine translation and that can be used to generate natural sentences describing an image. The model is trained to maximize the likelihood of the target description sentence given the training image. Experiments on several datasets show the accuracy of the model and the fluency of the language it learns solely from image descrip-

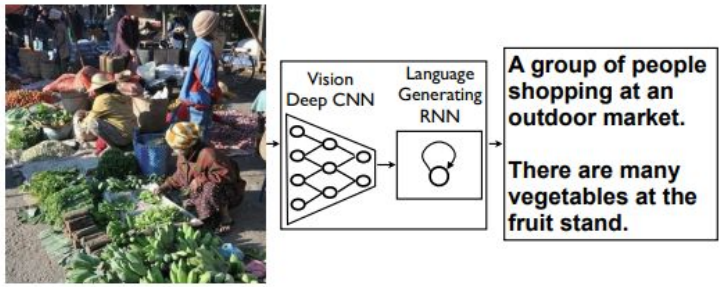
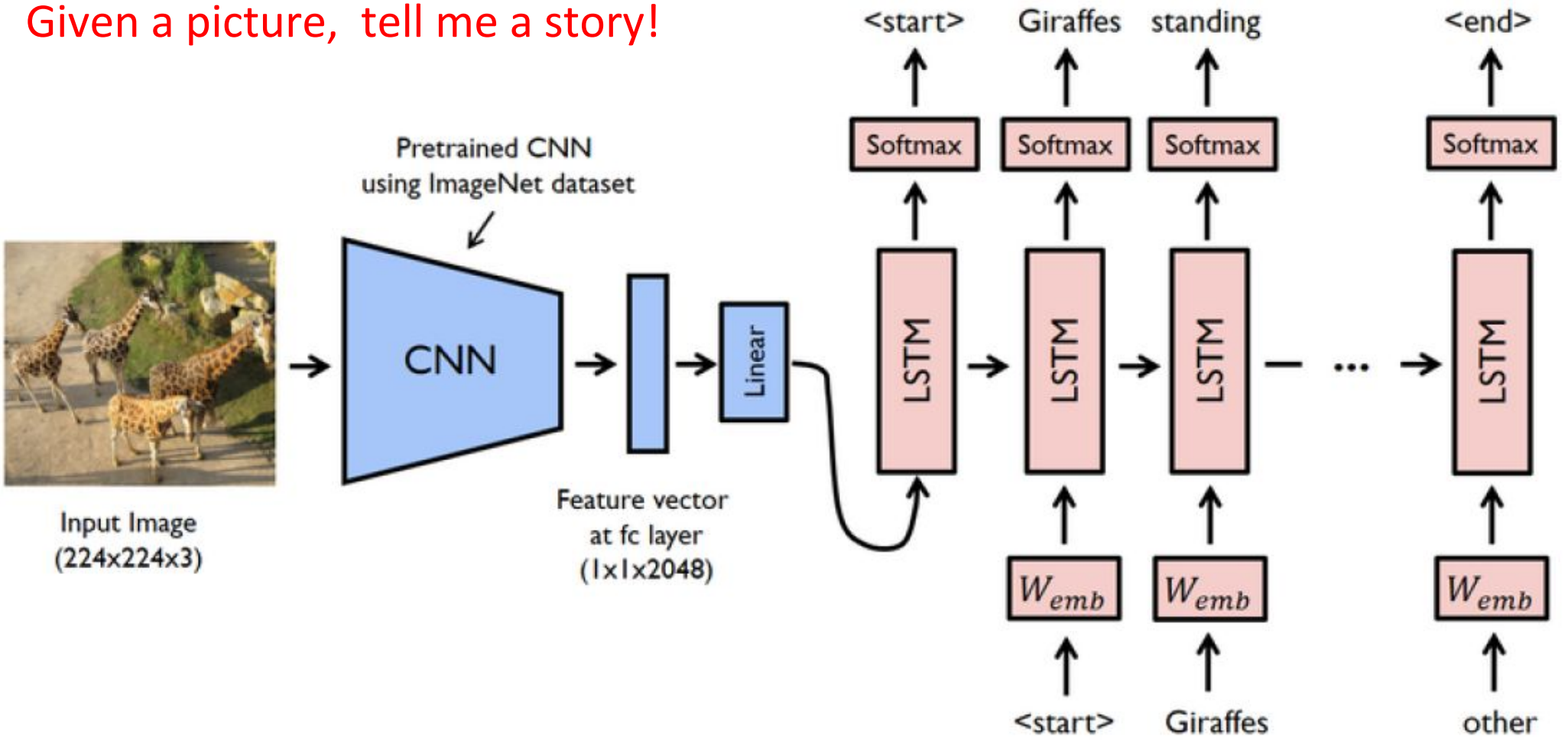


Figure 1. NIC, our model, is based end-to-end on a neural network consisting of a vision CNN followed by a language generating RNN. It generates complete sentences in natural language from an input image, as shown on the example above.

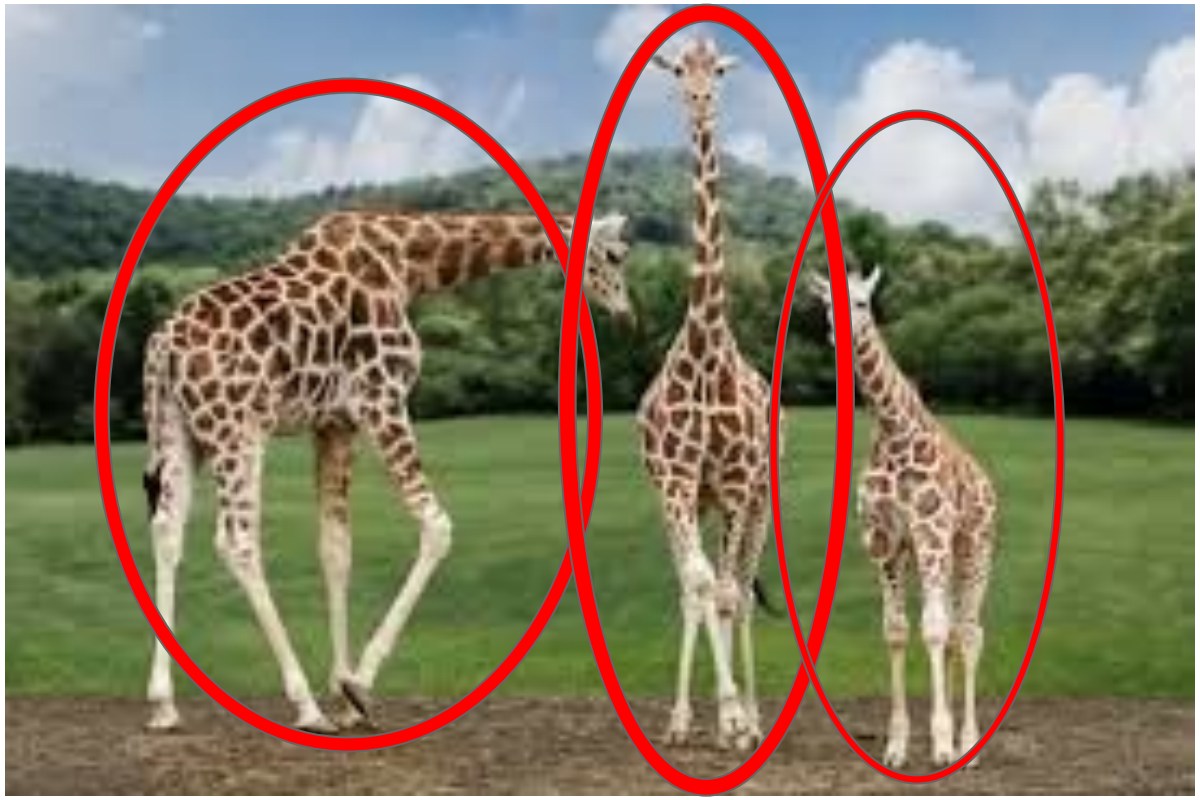
Image-Caption

Given a picture, tell me a story!



Simple Image Caption Problem

- What's wrong about simple "Image cation"
  - E.g We won't "stare" at whole picture
  - We will pick up something and start to describe it



Simple Image Caption Problem

- What's wrong about simple "Image cation"
  - E.g We won't "stare" at whole picture
  - We will pick up something and start to describe it

- How we can improve ?
  - Use attention!!

Attention

- The attention mechanism
  - Want model to paying attention to the salient information that should be focused on when generating a word.

- The means is to give weight to each part of the input information.

Image with Attention

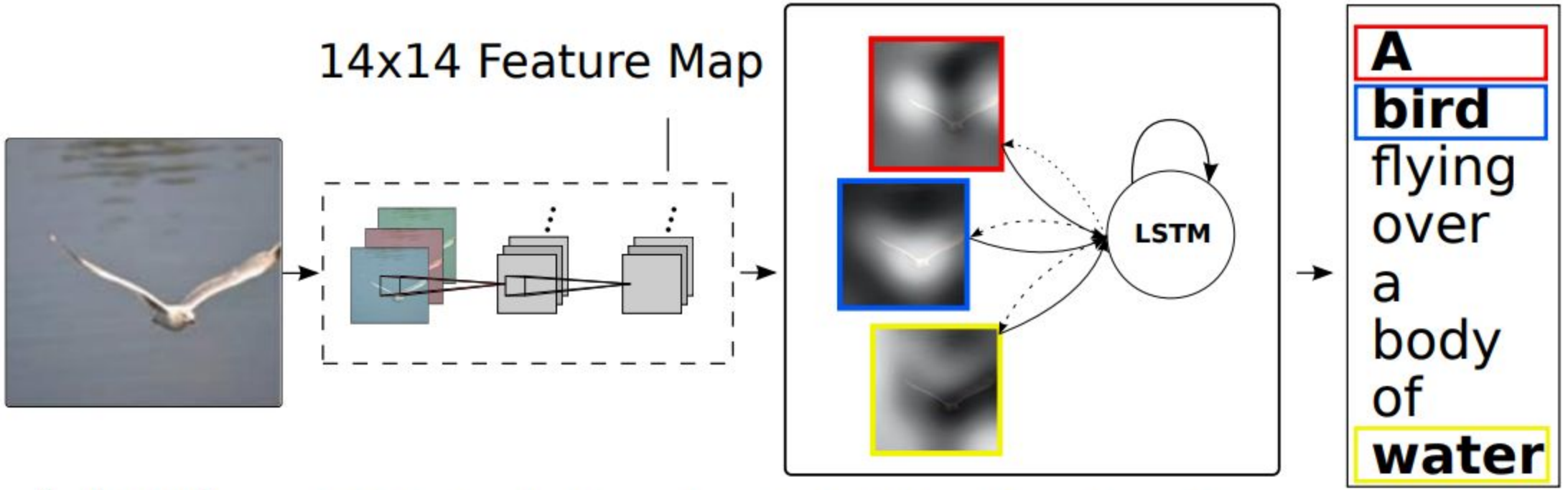


Image with Attention

- Add Attention
  - Compute each area weight at each time “T”
  - E.g,

$$z_t = \sum_{i=1}^{L} \alpha_{t,i} a_i$$

Image with Attention

- Add Attention
  - Compute each area weight at each time “T”

$$\hat{\mathbf{z}}_t = \phi(\{\mathbf{a}_i\}, \{\alpha_{ti}\})$$

$$e_{ti} = f_{\text{att}}(\mathbf{a}_i, \mathbf{h}_{t-1})$$

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{L} \exp(e_{tk})}$$

Hard Attention

- Only one area will be choosed to send into decoder
  - S_t = 0 or 1

$$\hat{\mathbf{z}}_t = \sum_i s_{t,i} \mathbf{a}_i$$

Soft Attention

- Use weight to bring all area information into Decoder

$$\mathbb{E}_{p(s_t|a)}[\hat{\mathbf{z}}_t] = \sum_{i=1}^{L} \alpha_{t,i}\,\mathbf{a}_i$$

Result

Figure 3. Examples of attending to the correct object (*white* indicates the attended regions, *underlines* indicated the corresponding word)

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.

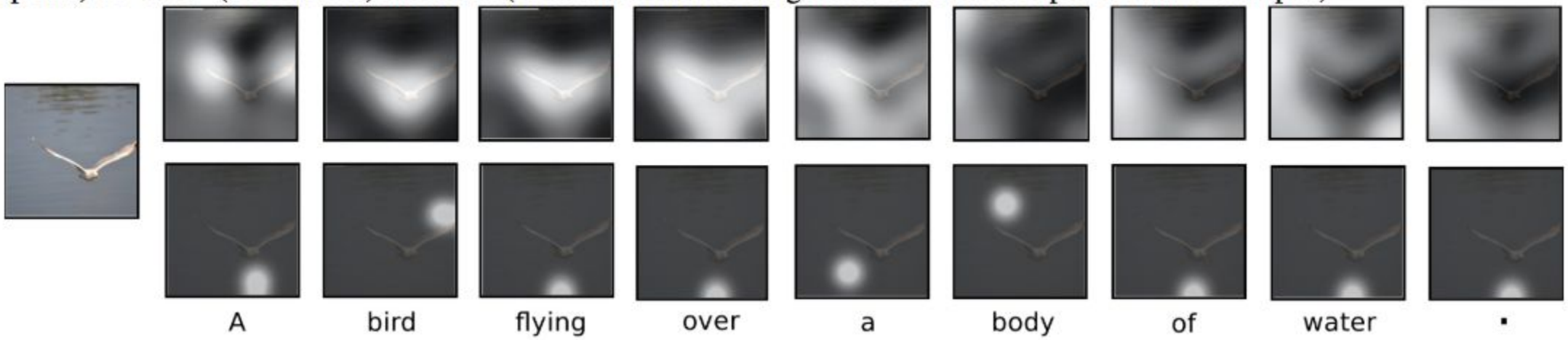
A little girl sitting on a bed with a teddy bear.

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

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

Result

Figure 2. Attention over time. As the model generates each word, its attention changes to reflect the relevant parts of the image. "soft" (top row) vs "hard" (bottom row) attention. (Note that both models generated the same captions in this example.)



Paper Reference -- Attention

- Neural Machine Translation by Jointly Learning to Align and Translate
  - https://arxiv.org/abs/1409.0473

Published as a conference paper at ICLR 2015

NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE

Dzmitry Bahdanau
Jacobs University Bremen, Germany

KyungHyun Cho **Yoshua Bengio***
Université de Montréal

ABSTRACT

Neural machine translation is a recently proposed approach to machine translation. Unlike the traditional statistical machine translation, the neural machine translation aims at building a single neural network that can be jointly tuned to maximize the translation performance. The models proposed recently for neural machine translation often belong to a family of encoder–decoders and encode a source sentence into a fixed-length vector from which a decoder generates a translation. In this paper, we conjecture that the use of a fixed-length vector is a bottleneck in improving the performance of this basic encoder–decoder architecture, and propose to extend this by allowing a model to automatically (soft-)search for parts of a source sentence that are relevant to predicting a target word, without having to form these parts as a hard segment explicitly. With this new approach, we achieve a translation performance comparable to the existing state-of-the-art phrase-based system on the task of English-to-French translation. Furthermore,

Paper Reference -- Image Caption with Attention

- Show, Attend and Tell: Neural Image Caption Generation with Visual Attention
  - https://arxiv.org/pdf/1502.03044.pdf

Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

Kelvin Xu KELVIN.XU@UMONTREAL.CA
Jimmy Lei Ba JIMMY@PSI.UTORONTO.CA
Ryan Kiros RKIROS@CS.TORONTO.EDU
Kyunghyun Cho KYUNGHYUN.CHO@UMONTREAL.CA
Aaron Courville AARON.COURVILLE@UMONTREAL.CA
Ruslan Salakhutdinov RSALAKHU@CS.TORONTO.EDU
Richard S. Zemel ZEMEL@CS.TORONTO.EDU
Yoshua Bengio FIND-ME@THE.WEB

Abstract

Inspired by recent work in machine translation and object detection, we introduce an attention based model that automatically learns to describe the content of images. We describe how we can train this model in a deterministic manner using standard backpropagation techniques and stochastically by maximizing a variational lower bound. We also show through visualization how the model is able to automatically learn to fix its

Figure 1. Our model learns a words/image alignment. The visualized attentional maps (3) are explained in section 3.1 & 5.4