

Lab 13-2: Image Captioning

Department of Computer Science,
National Tsing Hua University, Taiwan

2020.12.03

Outline

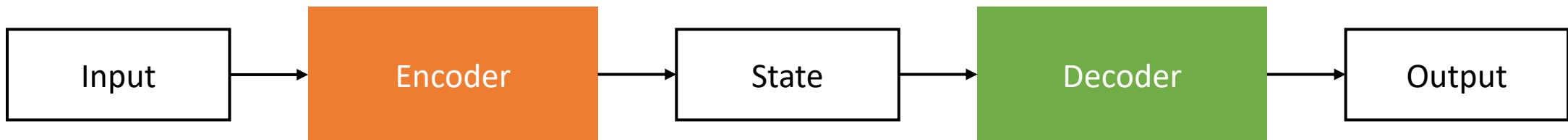
- Encoder-Decoder model
- Attention-based
- Assignment

Outline

- Encoder-Decoder model
- Attention-based
- Assignment

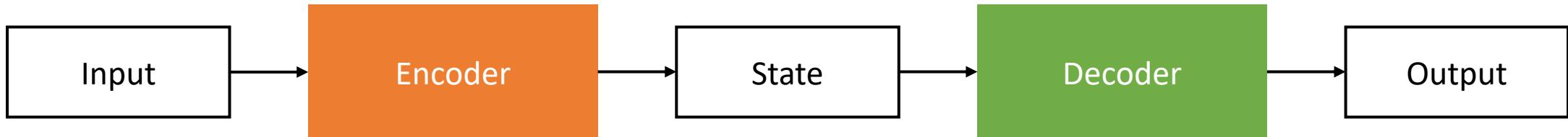
Encoder-Decoder Model

- Lab13-1 - Neural Machine Translation
 - Encoder RNN: reads the source sentence and transforms it into a rich fixed-length vector representation
 - Decoder RNN: uses the representation as the initial hidden state and generates the target sentence



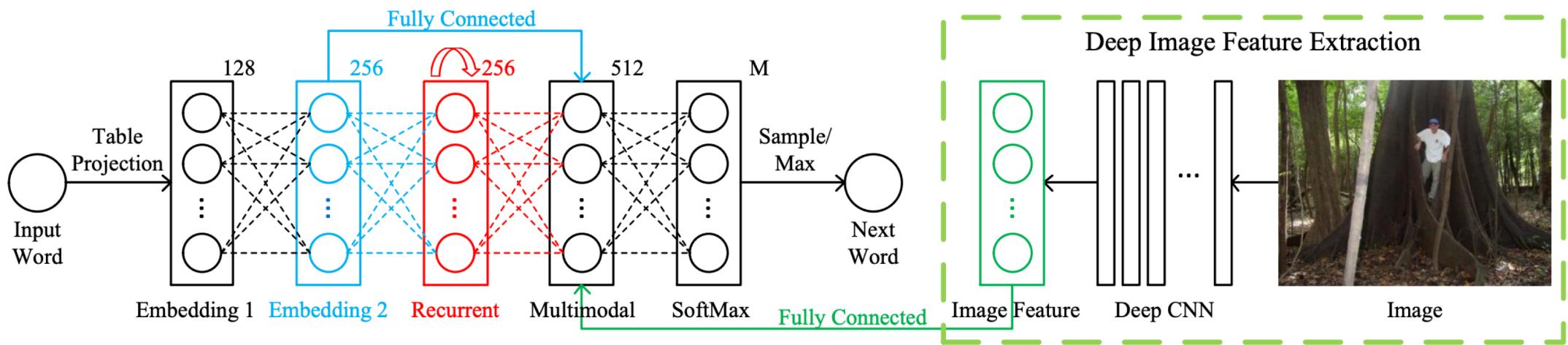
Encoder-Decoder Model

- Image Captioning
 - Encoder CNN: reads the **images** and transforms it into a rich fixed-length vector representation
 - Decoder RNN: uses the representation as the initial hidden state and generates the target sentence



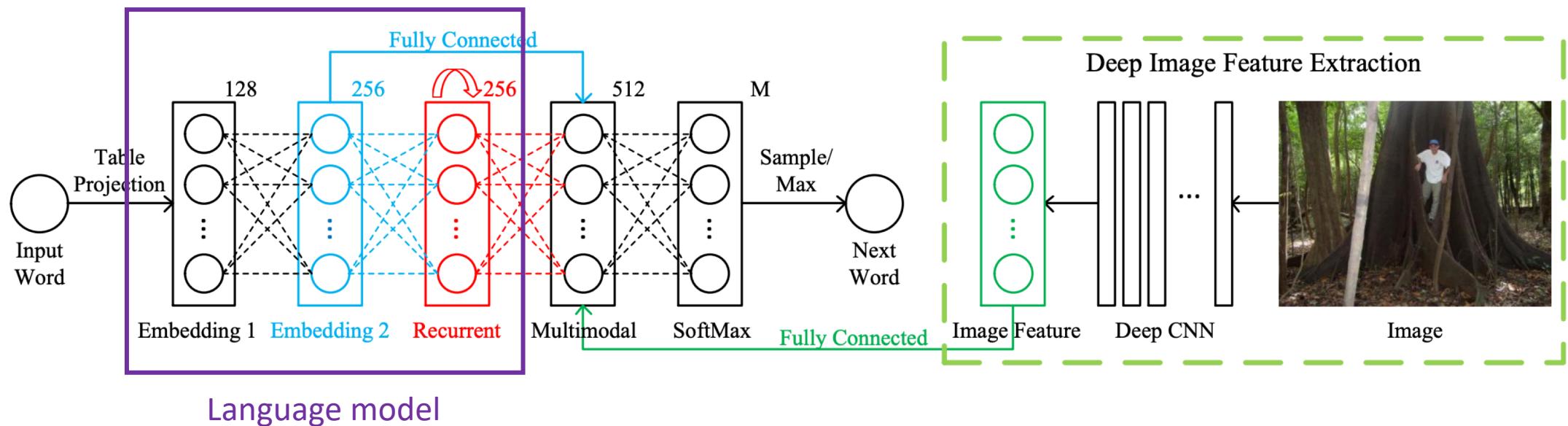
Encoder-Decoder Model

- m-RNN (multimodal RNN)



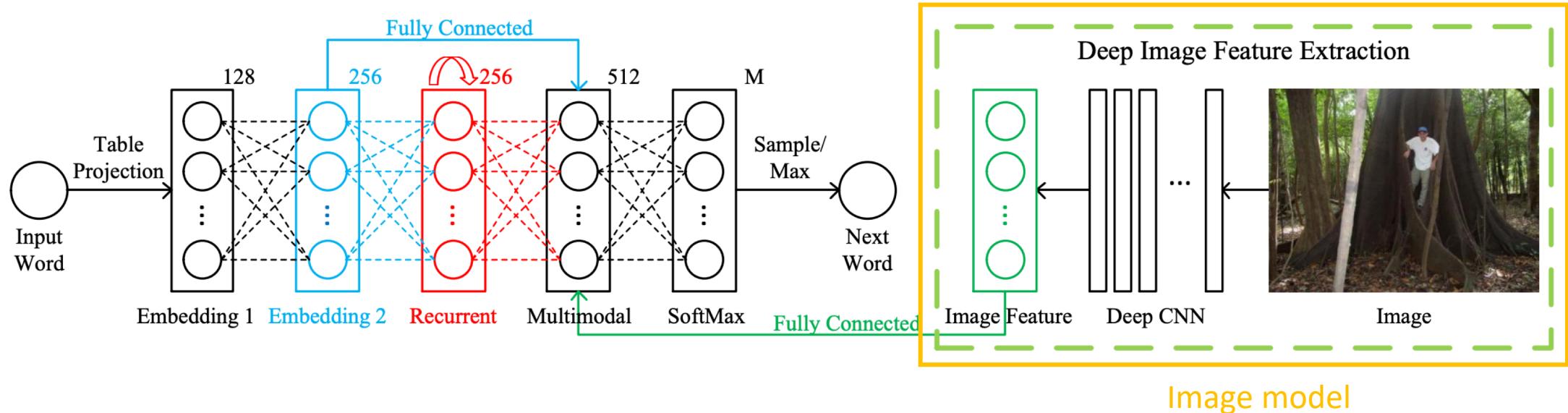
Encoder-Decoder Model

- m-RNN (multimodal RNN)
 - The language model part learns the dense feature embedding for each word



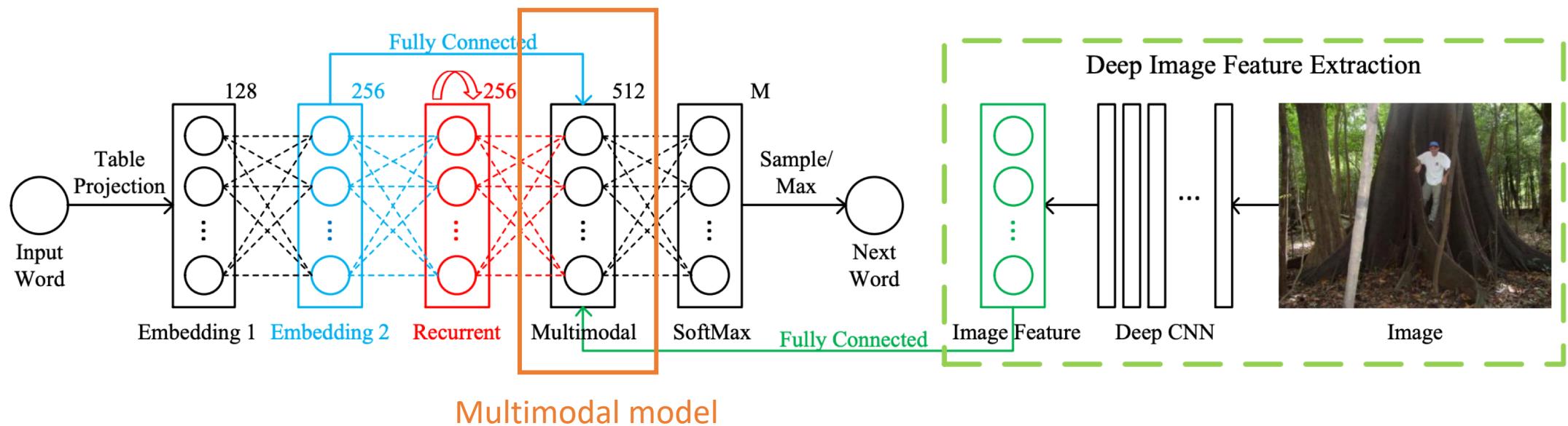
Encoder-Decoder Model

- m-RNN (multimodal RNN)
 - The language model part learns the dense feature embedding for each word
 - The image part contains a deep CNN which extracts image features



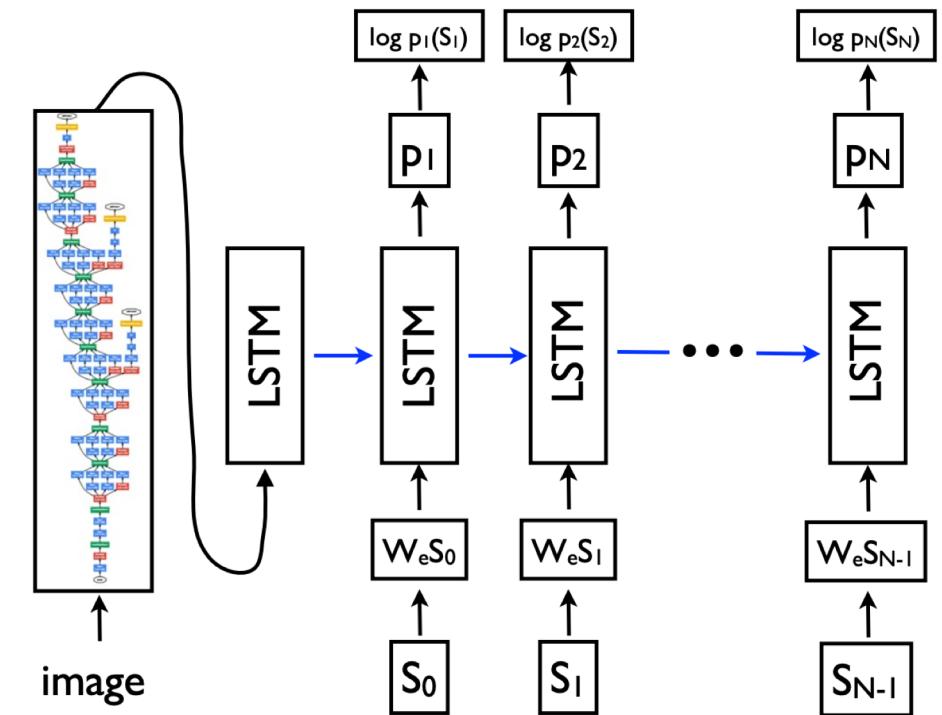
Encoder-Decoder Model

- m-RNN (multimodal RNN)
 - The language model part learns the dense feature embedding for each word
 - The image part contains a deep CNN which extracts image features
 - The multimodal part connects the language model and the deep CNN together by a one-layer representation



Encoder-Decoder Model

- NIC
 - A generative model based on a deep recurrent architecture that combines recent advances in computer vision and machine translation
 - Uses a more powerful CNN in the encoder
 - The image is only input once

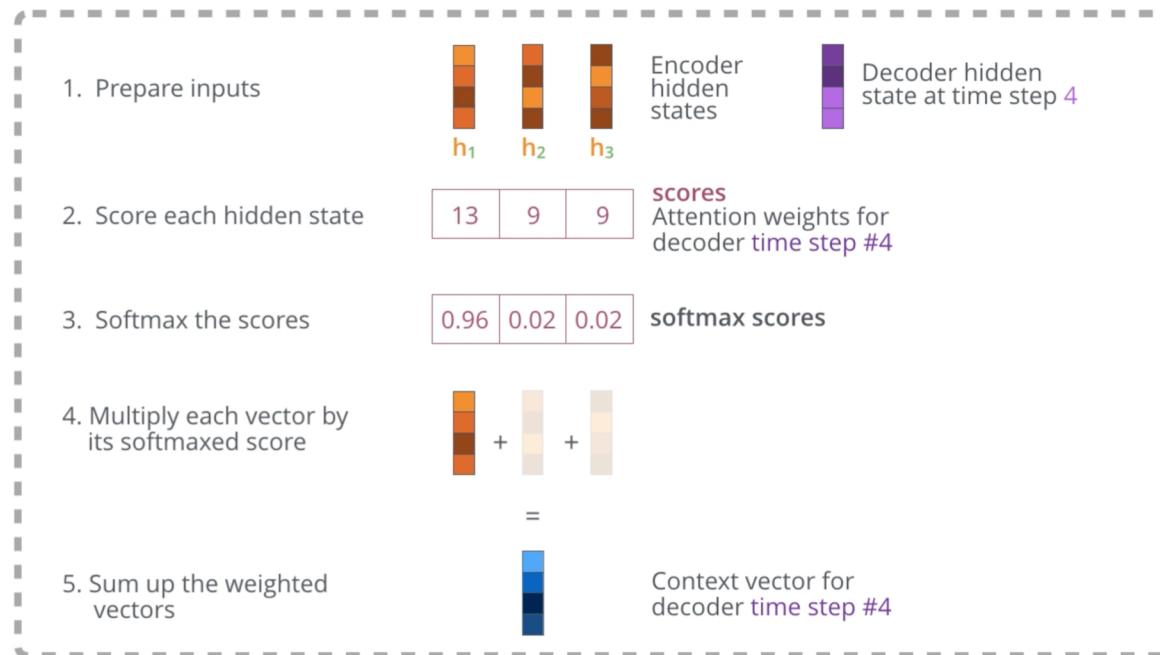


Outline

- Encoder-Decoder model
- Attention-based
- Assignment

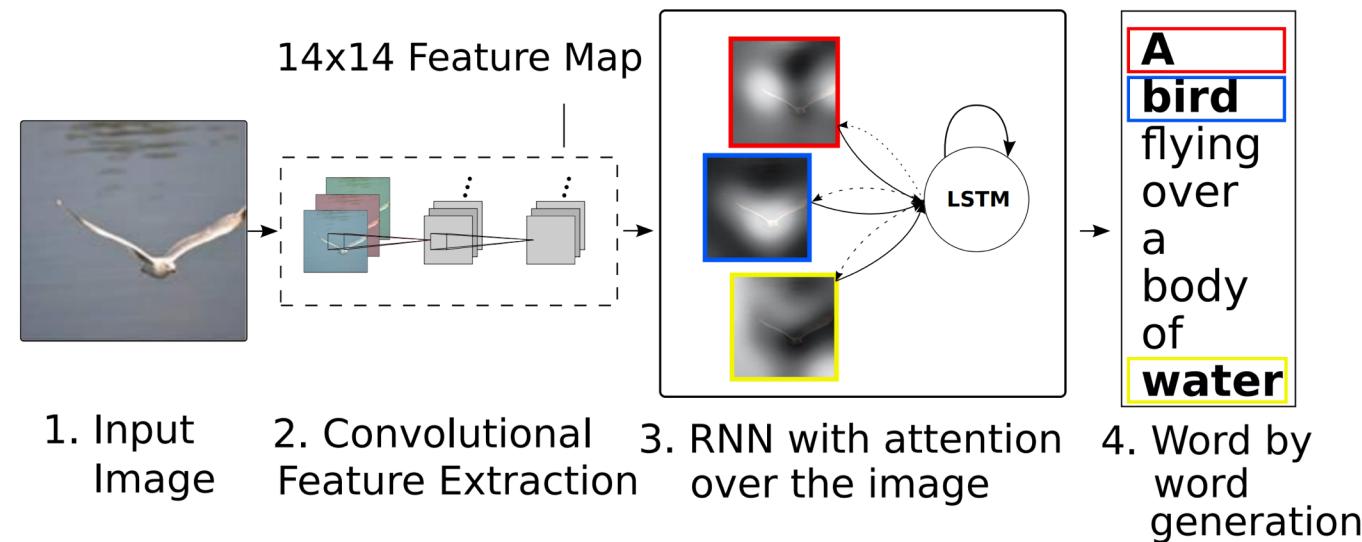
Attention Based

- Attention allows the model to focus on the relevant parts of the input sequence as needed



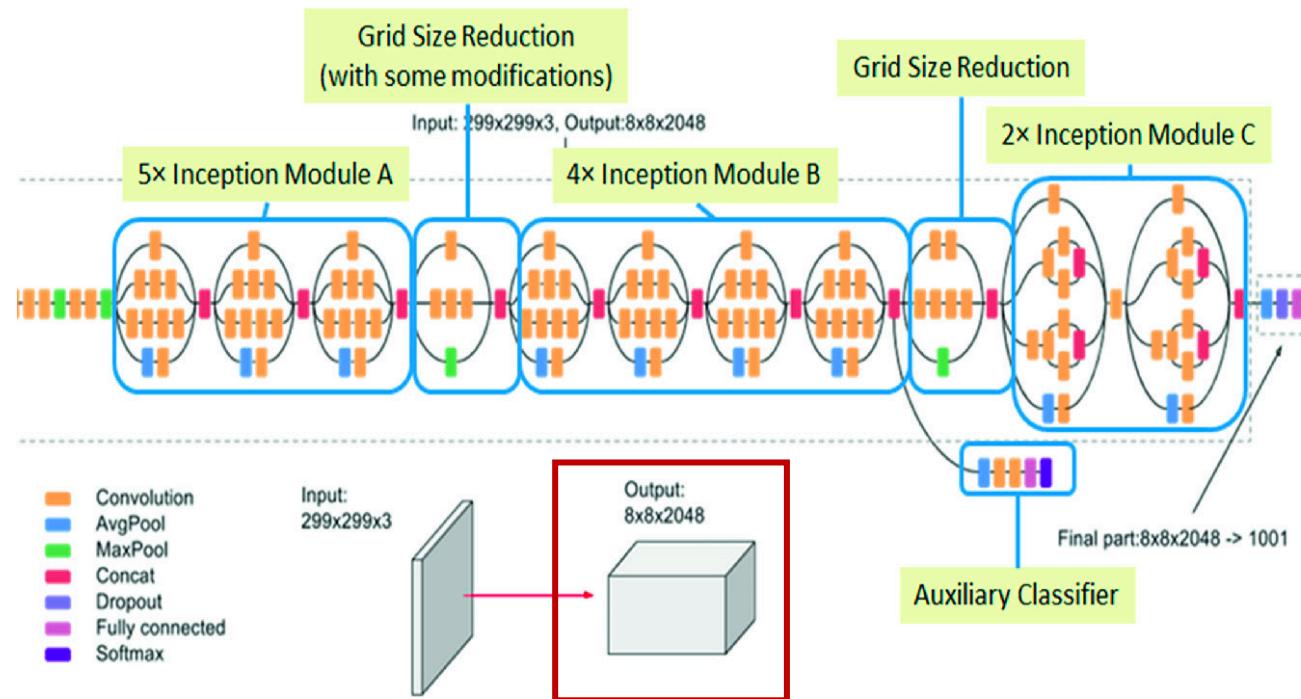
Attention Based

- Attention allows the model to focus on the relevant parts of the input sequence as needed
 - Show, Attend and Tell: Neural Image Caption Generation with Visual Attention



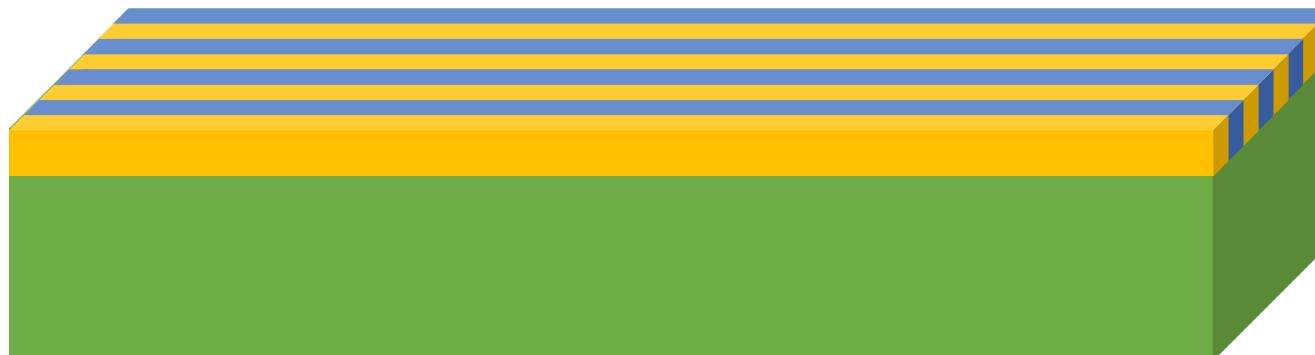
Attention Based

- First, extract the features from image



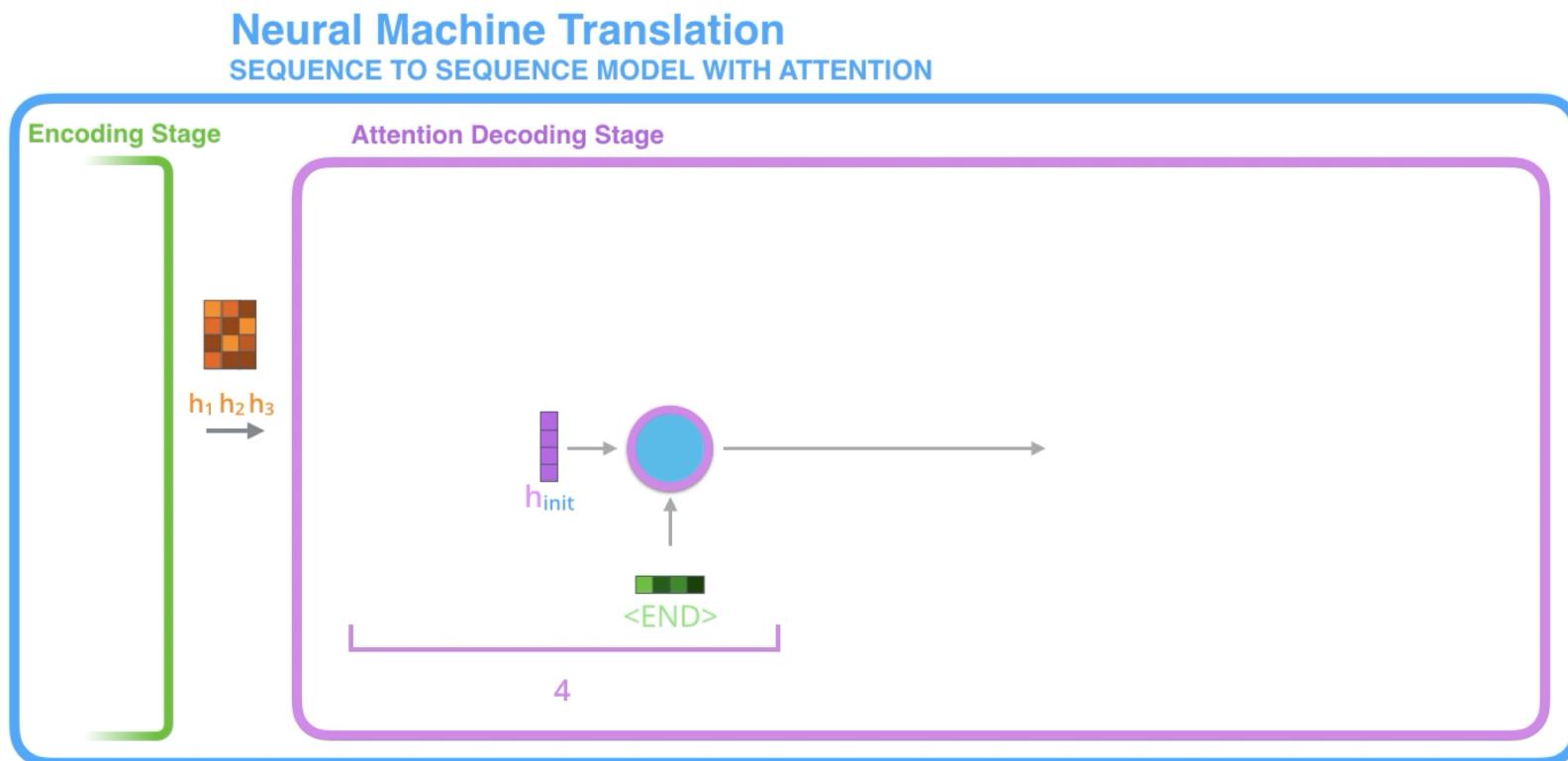
Attention Based

- First, extract the features from image
- We have a $8*8*2048$ size feature map, the last layer has $8*8$ pixel locations which corresponds to certain portion in image
- That means we have 64 pixel locations
- The model will then learn an attention over these locations



Attention Based

- The rest is similar to the neural machine translation task



Outline

- Encoder-Decoder model
- Attention-based
- Assignment

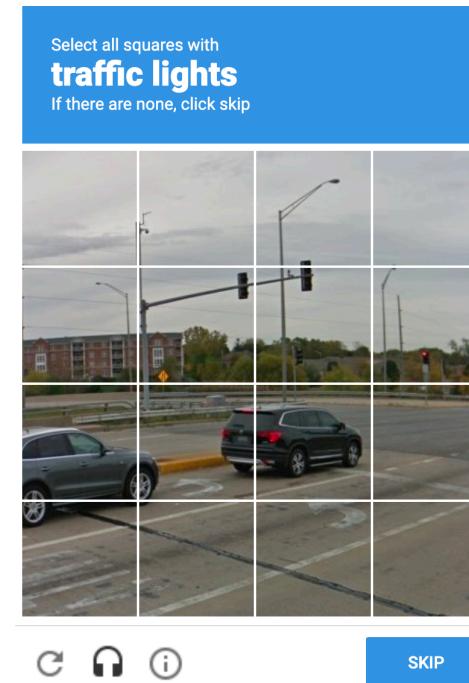
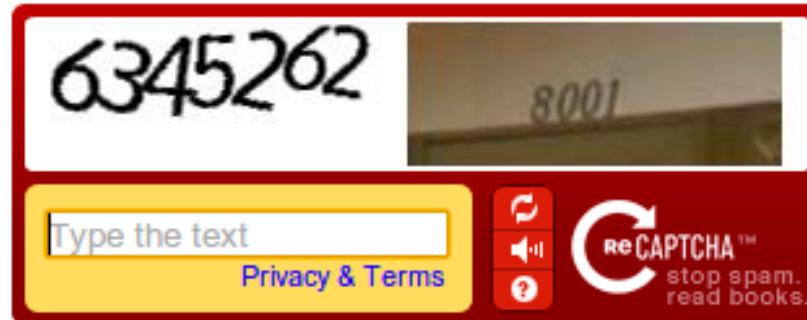
Assignment

- CAPTCHA
 - An acronym for “Completely Automated Public Turing test to tell Computers and Humans Apart”
 - A type of challenge–response test used in computing to determine whether or not the user is human
 - Prevents spam attacks and protects websites from bots

$$7 \times \begin{matrix} 5 \\ \text{[Colorful 5]} \end{matrix} = \boxed{} \quad \text{↻}$$

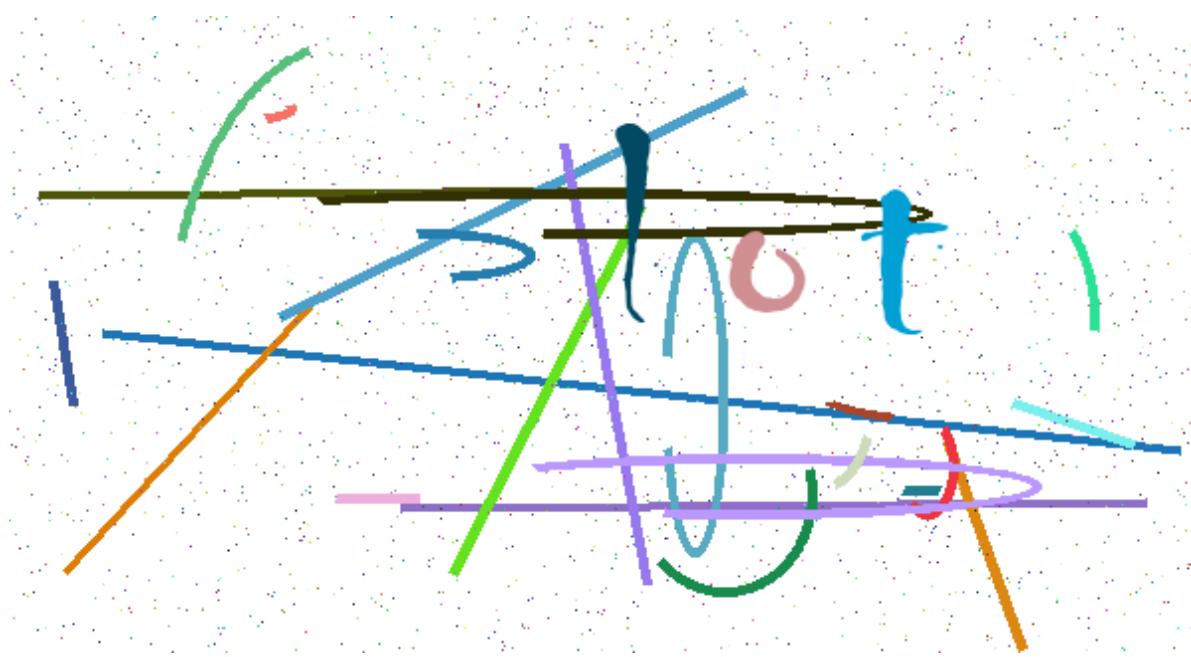
Assignment

- reCAPTCHA
 - Establish that a computer user is human
 - Assist in the digitization of books or improve machine learning



Assignment

- We are going to train a captcha recognizer in this lab
- Dataset
 - 140,000 CAPTCHAs



Assignment

- Requirement
 - Use any model architectures you want
 - Design your own model architecture
 - The first 100,000 as training data, the next 20,000 as validation data, and the rest as testing data
 - Only if the whole word matches exactly does it count as correct
 - Predict the answer to the testing data and write them in a file
 - Testing accuracy should be at least 90%
- Please submit your code file and the answer file

Assignment

- Requirement
 - Use any model architectures you want
 - Design your own model architecture
 - The first 100,000 as training data, the next 20,000 as validation data, the rest as testing data
 - Only if the whole word matches exactly does it count as correct
 - Predict the answer to the testing data and write them in a file
 - Testing accuracy should be at least 90%
- Please submit your code file and the answer file

a0	thus
a1	www
a2	tied
a3	ids
a4	jam
a5	zoo
a6	apple
a7	big
a8	lot
a9	above
a10	ooo