Image Captioning with Neural Networks

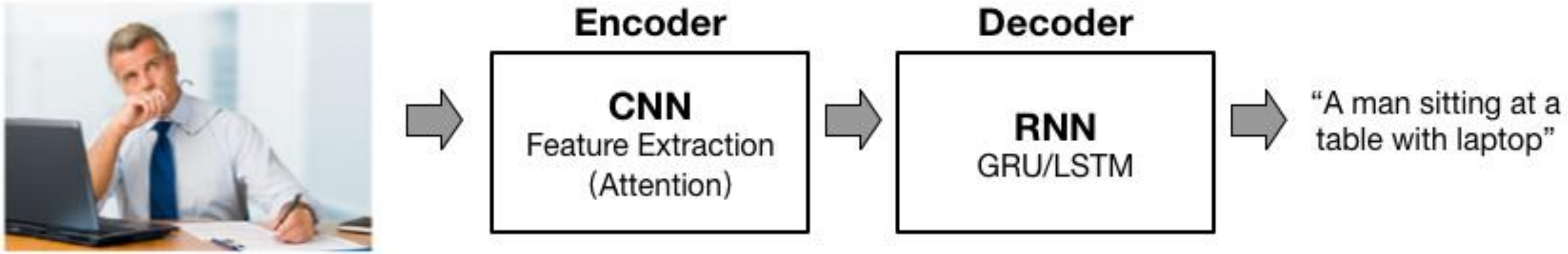
G03 Tamago
Jie Chi   Yirui Huang

April 2018

Introduction

- Image captioning
  - Generate texts from images
  - Image retrieval, visual assistants, etc.

- Encoder-decoder framework



Encoder

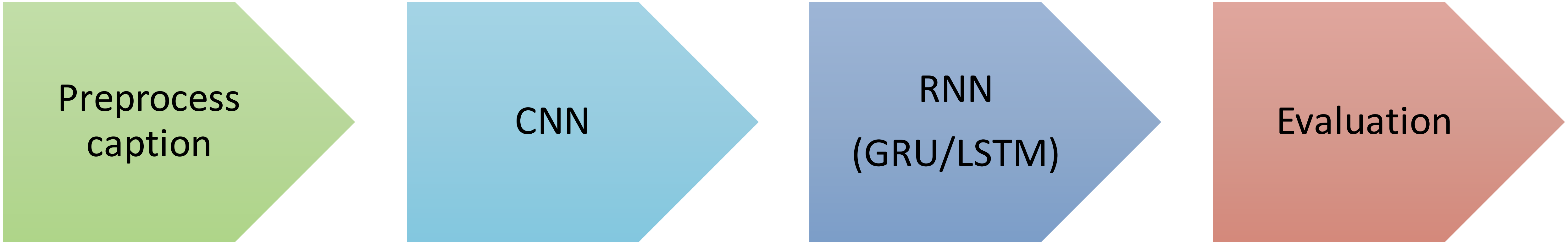
CNN
Feature Extraction
(Attention)

Decoder

RNN
GRU/LSTM

"A man sitting at a table with laptop"

Methodology



Preprocess caption

CNN

RNN

(GRU/LSTM)

Evaluation

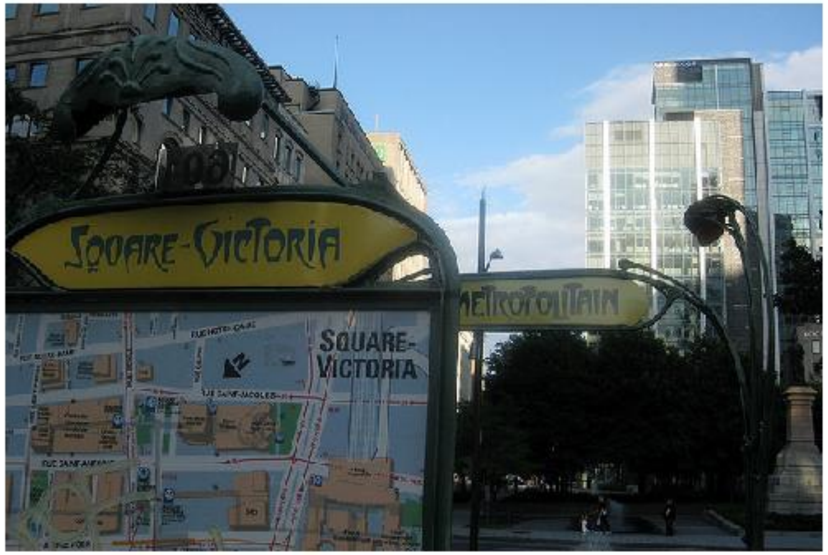
- Dataset: MSCOCO
- Tokenize the caption
- Glove300 word embedding

- Inception V3
- Standard model (1*2048)
- Attention model (64*2048)

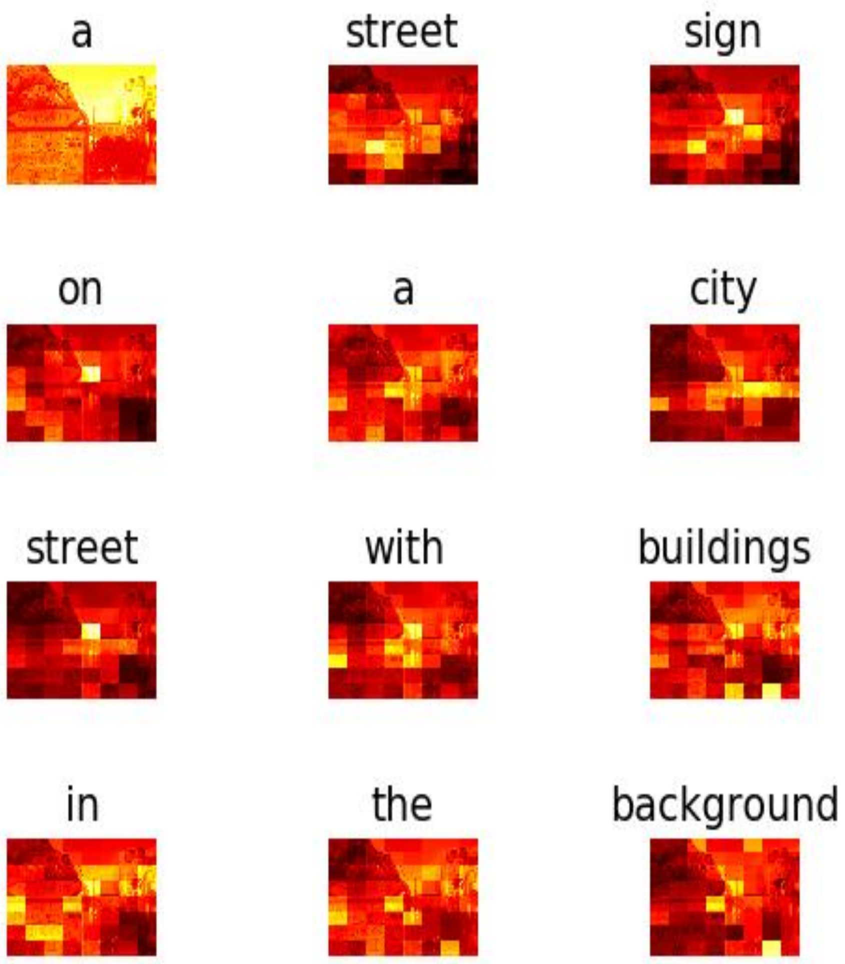
- Standard model: image feature vectors fed only once
- Attention model: image vectors fed at every step

- Metrics: BLEU, CIDEr, METEOR, ROUGE_L
- Early stopping on CIDEr score

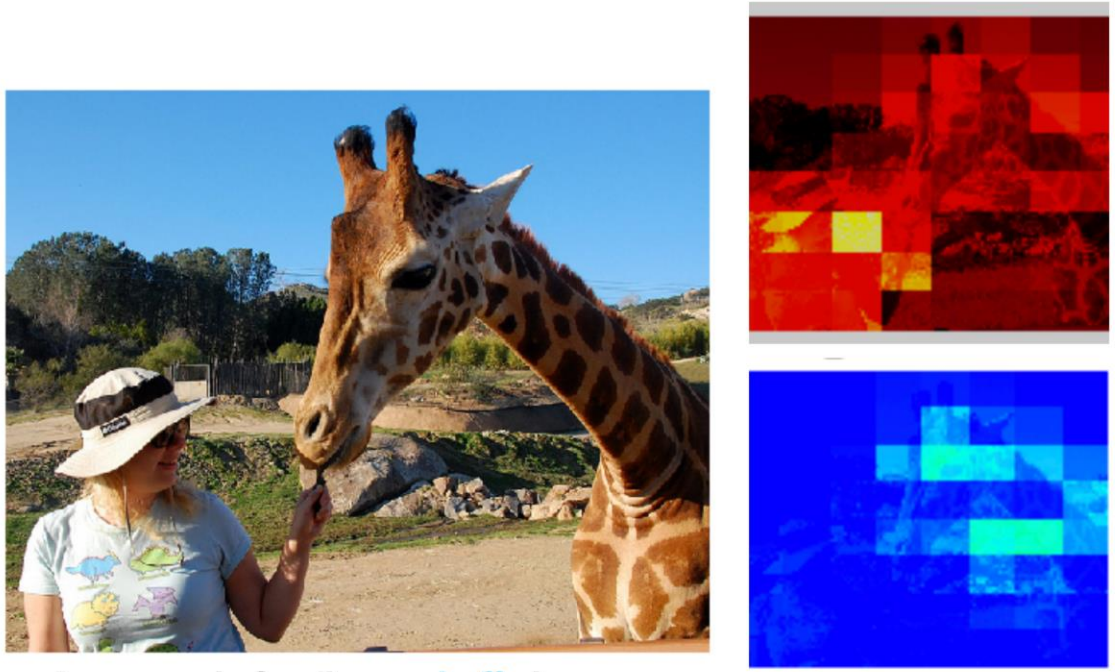
Results – captions



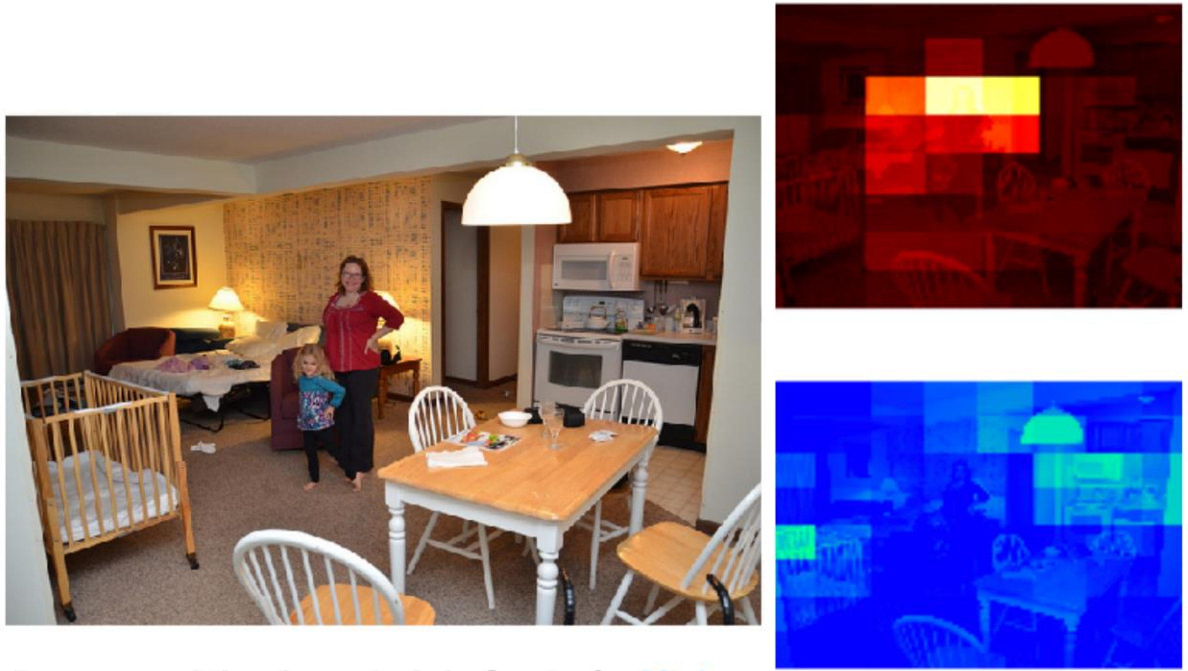
MODEL	GENERATED CAPTIONS
LSTM	a sign that is on a pole on a street
GRU	a sign on a building with a sign on it
Att-LSTM	a street sign on a pole in front of a building
Att-GRU	a street sign on a city street with buildings in the background



Results – attention map

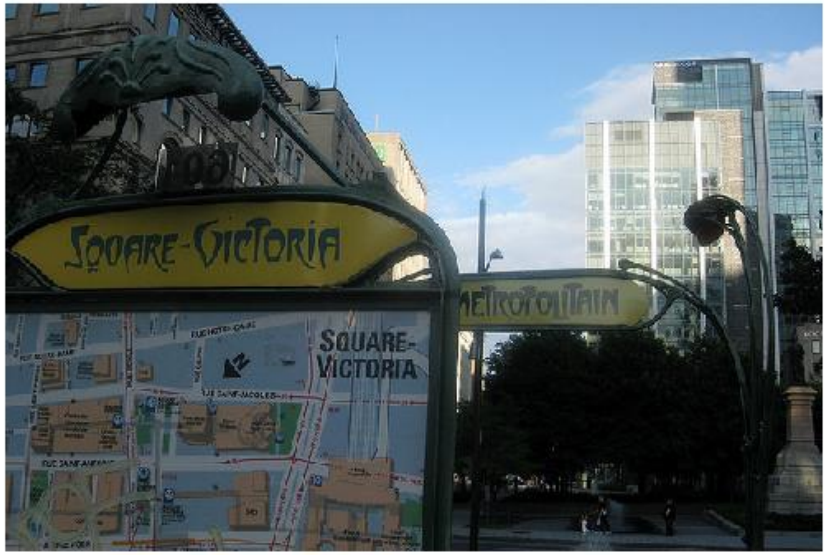


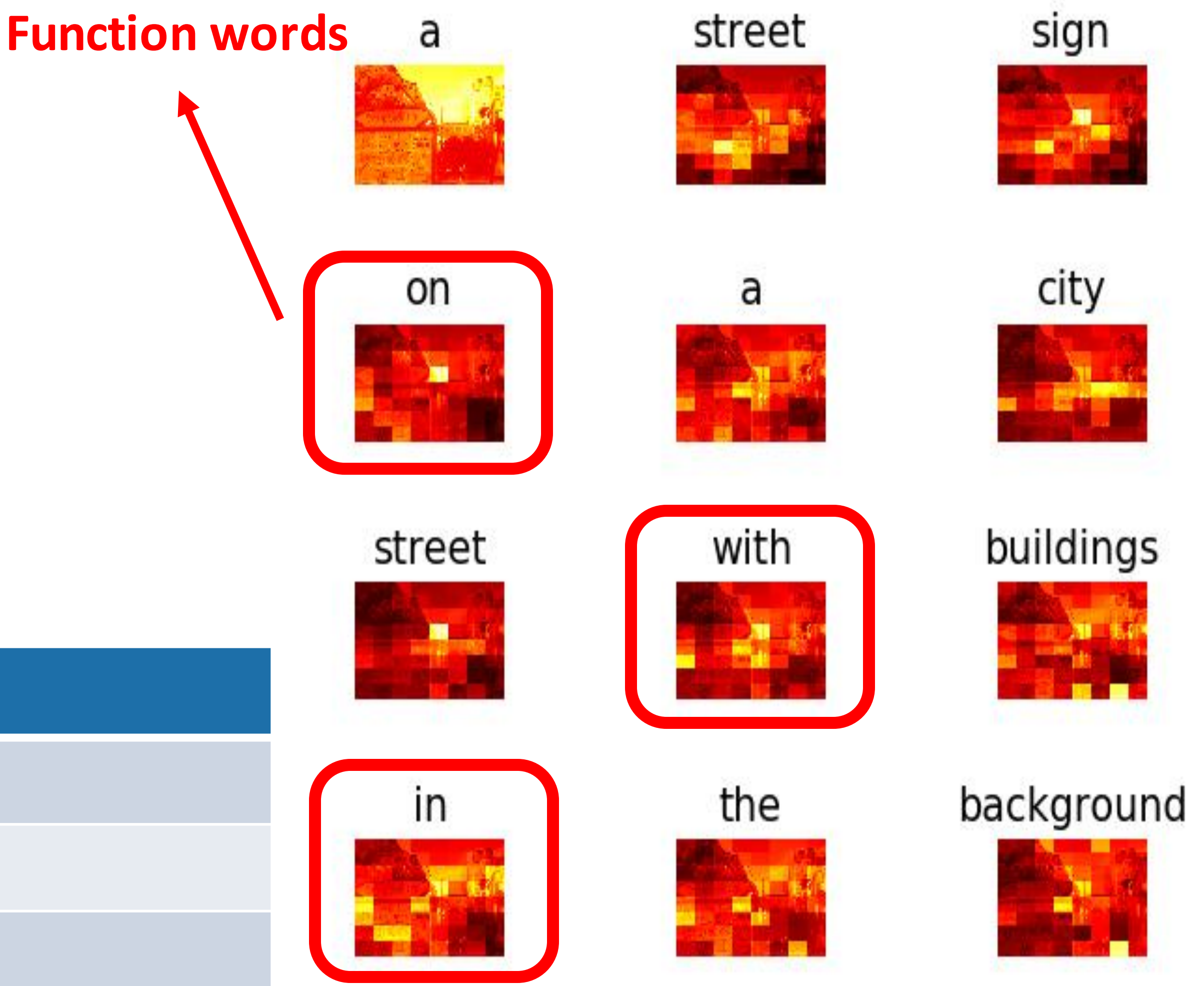
A woman is feeding a giraffe in a zoo



A woman sitting in a chair in front of a kitchen

Results – captions




Function words

MODEL	GENERATED CAPTIONS
LSTM	a sign that is on a pole on a street
GRU	a sign on a building with a sign on it
Att-LSTM	a street sign on a pole in front of a building
Att-GRU	a street sign on a city street with buildings in the background

Conclusion

- **LSTM vs GRU:**
  - **Similar performance**
  - **GRUs converged early**
- **Attention vs Non-attention:**
  - **Attention models have higher scores**
  - **Attends to correct objects as well as function words**
- **Best performance:**

Model	BLEU-4	METEOR	CIDEr	ROUGE_L
Att-LSTM	28.6	24.4	92.2	52.3