Multi-modal Neural Machine Translation

What is it and why bother?

Iacer Calixto

December 13, 2016

ADAPT Centre
School of Computing
Dublin City University
iacer.calixto@adaptcentre.ie

Outline

Introduction

Definitions

- **Machine Translation (MT)**: the task in which we wish to **learn a model** to **translate text from one natural language** (e.g., English) **into another** (e.g., Brazilian Portuguese).
- Related NLP tasks:
    - **speech translation**,
    - **text simplification**,
    - **text summarisation**,
    - **question answering**,
    - etc.

Definitions

- **Machine Translation (MT)**: the task in which we wish to **learn a model** to **translate text from one natural language** (e.g., English) **into another** (e.g., Brazilian Portuguese).
- Related NLP tasks:
  - **speech translation**,
  - **text simplification**,
  - **text summarisation**,
  - **question answering**,
  - etc.

Definitions

- **Machine Translation (MT)**: the task in which we wish to **learn a model** to **translate text from one natural language** (e.g., English) **into another** (e.g., Brazilian Portuguese).
- Related NLP tasks:
  - **speech translation**,
  - **text simplification**,
  - **text summarisation**,
  - **question answering**,
  - etc.

Definitions

- **Machine Translation (MT)**: the task in which we wish to **learn a model** to **translate text from one natural language** (e.g., English) **into another** (e.g., Brazilian Portuguese).
- Related NLP tasks:
  - **speech translation**,
  - **text simplification**,
  - **text summarisation**,
  - **question answering**,
  - etc.

Definitions

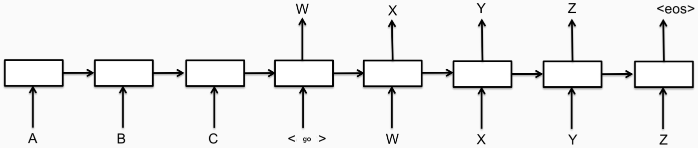
- **Machine Translation (MT)**: the task in which we wish to **learn a model** to **translate text from one natural language** (e.g., English) **into another** (e.g., Brazilian Portuguese).
- Related NLP tasks:
  - **speech translation**,
  - **text simplification**,
  - **text summarisation**,
  - **question answering**,
  - etc.

Neural MT Architectures

Sequence-to-sequence (encoder–decoder)

Cho et al. (2014); Sutskever et al. (2014)

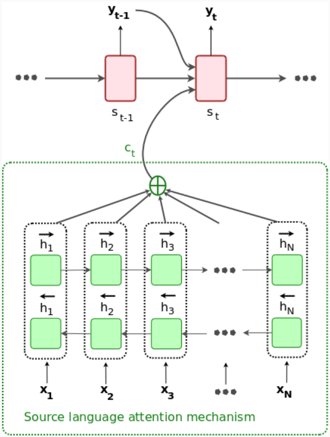
- **encoder RNN** maps source sequence $X = (x_1, x_2, \cdots, x_N)$ into a **fixed-length vector $\boldsymbol{x}$**.
- **decoder RNN** unravels target sequence $Y = (y_1, y_2, \cdots, y_M)$ from $\boldsymbol{x}$.



https://www.tensorflow.org/versions/r0.9/tutorials/seq2seq/index.html

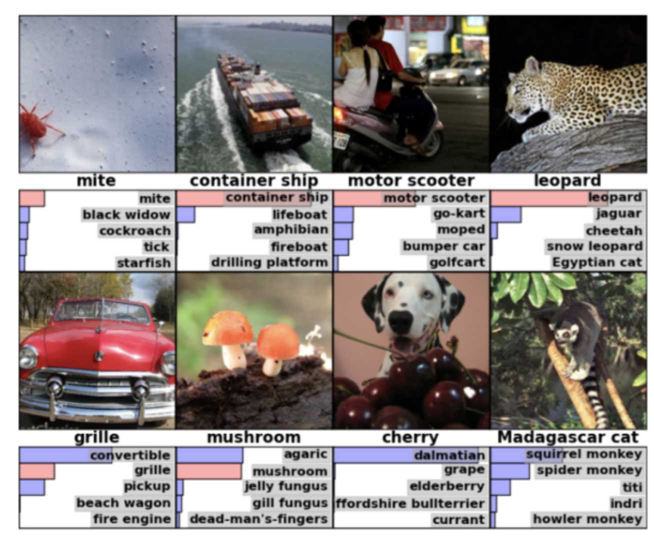
Attentional sequence-to-sequence

Attention mechanism removes the main bottleneck (fixed-size vector $\boldsymbol{x}$) and allows for **searching** for the best source words when generating each target word. — Bahdanau et al. (2015)

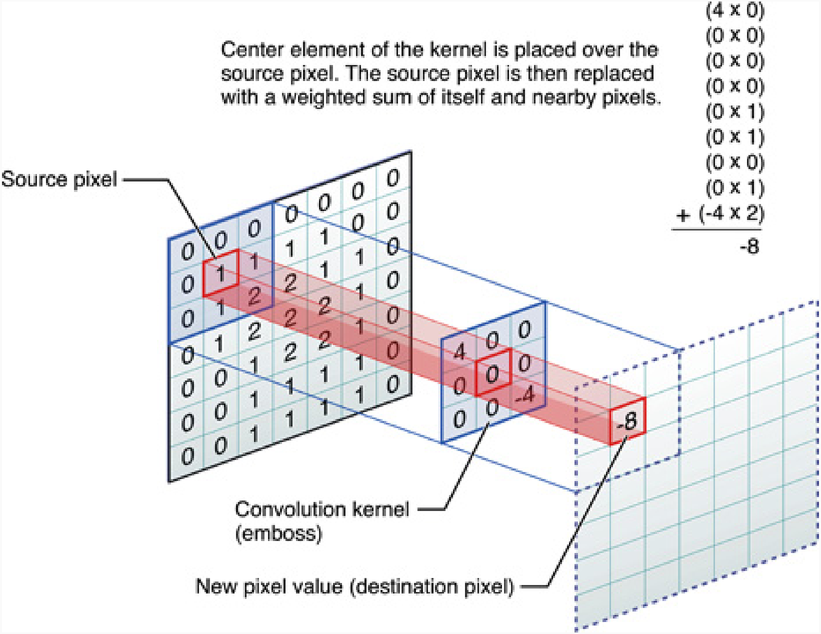


Computer Vision

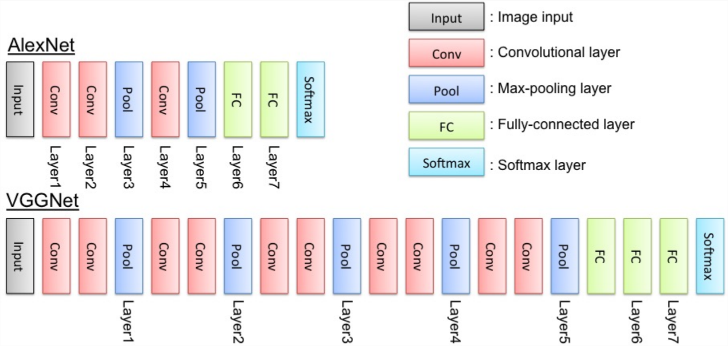
- **Computer Vision**: **how to make machines understand images**.



[Krizhevsky,Sutskever,Hinton2012]



https://developer.apple.com/library/content/documentation/Performance/Conceptual/vImage/Art/kernel_convolution.jpg



http://www.hirokatsukataoka.net/research/cnnfeatureevaluation/cnnarchitecture.jpg

Multi-modal Neural Machine Translation

Use cases?

Few use cases:

- news articles;
- picture captions (Facebook?);
- e-commerce product descriptions;
- etc.

Use cases?

Few use cases:

- news articles;
- picture captions (Facebook?);
- e-commerce product descriptions;
- etc.

Use cases?

Few use cases:

- news articles;
- picture captions (Facebook?);
- e-commerce product descriptions;
- etc.

Use cases?

Few use cases:

- news articles;
- picture captions (Facebook?);
- e-commerce product descriptions;
- etc.

Integrating FC features

- FC are *fully-connected features* that encode the *entire image* in *one single vector*.

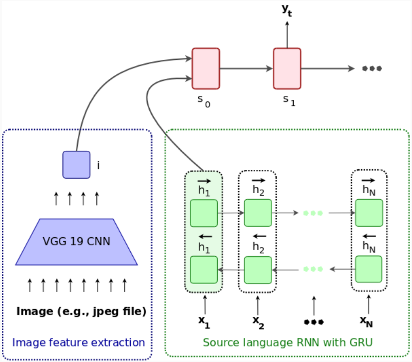


Figure 1: Using image to initialise the decoder hidden state.

Integrating FC features (2)

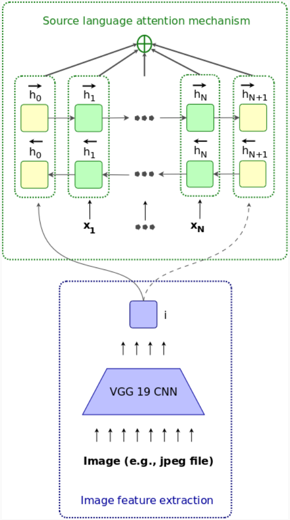


Figure 2: Using projected image as words in the source sentence.

Integrating FC features (3)

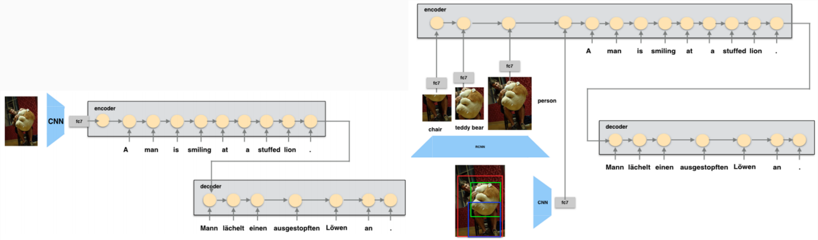


Figure 3: Attention-based Multimodal NMT [Huang et al. (2016)]

Integrating FC features (4)

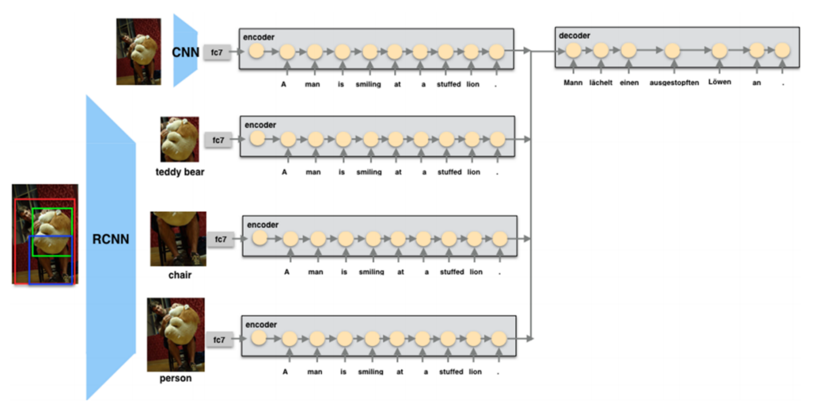


Figure 4: Attention-based Multimodal NMT [Huang et al. (2016)]

Integrating CONV features

- CONV are *convolutional features* that encode different areas (i.e., patches) of the image separately.

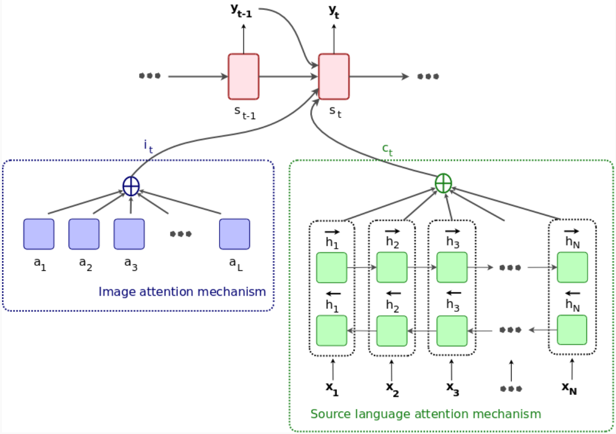


Figure 5: Doubly-attentive decoder with two independent attention mechanisms. [Calixto et al. (2016)]

Some numbers, why not?

	BLEU	METEOR
Text baseline	34.5 (0.7)	51.8 (0.7)
m1:image at tail	34.8 (0.6)	51.6 (0.7)
m1:image at head	35.1 (0.8)	52.2 (0.7)
m2:5 sequential RCNNs	36.2 (0.8)	53.4 (0.6)
m3:5 parallel RCNNs	**36.5** (0.8)	**54.1** (0.7)

Figure 6: BLEU and METEOR scores. [Huang et al. (2016)]

Model	BLEU	METEOR
Doubly-attentive decoder	36.2	53.1

Conclusions

- broad overview of the possibilities;
- few published results;
- what's next?
- possible extrapolations of the original formulation:
    - translating video subtitles;
    - visual question answering;

Conclusions

- broad overview of the possibilities;
- few published results;
- what's next?
- possible extrapolations of the original formulation:
    - translating video subtitles;
    - visual question answering;

Conclusions

- broad overview of the possibilities;
- few published results;
- what's next?
- possible extrapolations of the original formulation:
  - translating video subtitles;
  - visual question answering;

Conclusions

- broad overview of the possibilities;
- few published results;
- what's next?
- possible extrapolations of the original formulation:
  - translating video subtitles;
  - visual question answering;

Conclusions

- broad overview of the possibilities;
- few published results;
- what's next?
- possible extrapolations of the original formulation:
  - translating video subtitles;
  - visual question answering;

References

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations. ICLR 2015*.

Calixto, I., Elliott, D., and Frank, S. (2016). Dcu-uva multimodal mt system report. In *Proceedings of the First Conference on Machine Translation*, pages 634–638, Berlin, Germany. Association for Computational Linguistics.

Cho, K., van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014). On the properties of neural machine translation: Encoder–decoder approaches. *Syntax, Semantics and Structure in Statistical Translation*, page 103.

References II

Huang, P.-Y., Liu, F., Shiang, S.-R., Oh, J., and Dyer, C. (2016). Attention-based multimodal neural machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 639–645, Berlin, Germany. Association for Computational Linguistics.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112.

Questions?

Thank you!