Attention

DLAI – MARTA R. COSTA-JUSSÀ

SLIDES ADAPTED FROM GRAHAM NEUBIG'S LECTURES

What advancements excite you most in the field?

I am very excited by the recently introduced attention models, due to their simplicity and due to the fact that they work so well. Although these models are new, I have no doubt that they are here to stay, and that they will play a very important role in the future of deep learning.

ILYA SUTSKEVER, RESEARCH DIRECTOR AND COFUNDER OF OPENAI

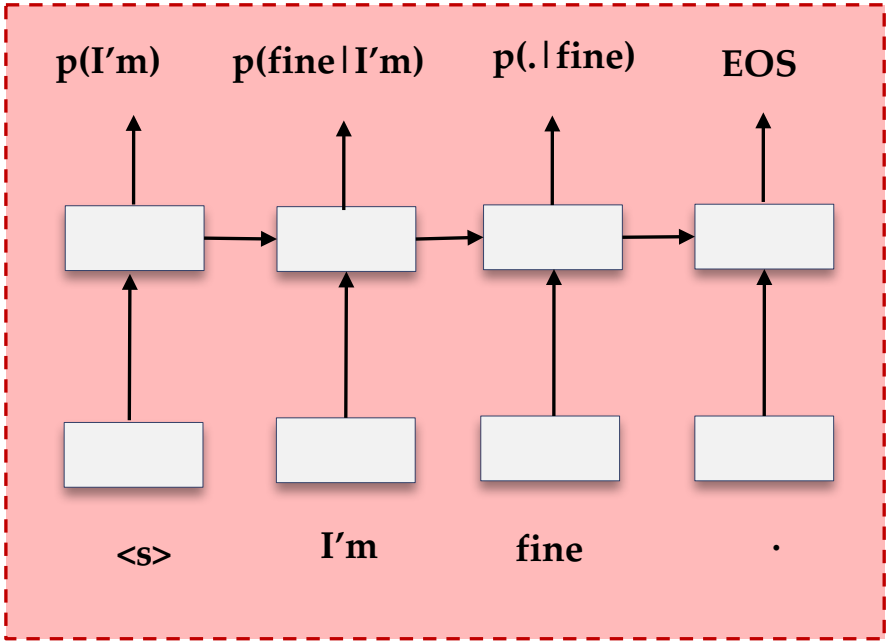
Outline

1. Sequence modeling & Sequence-to-sequence models [WRAP-UP FROM PREVIOUS RNN's SESSION]
2. Attention-based mechanism
3. Attention varieties
4. Attention Improvements
5. Applications
6. "Attention is all you need"
7. Summary

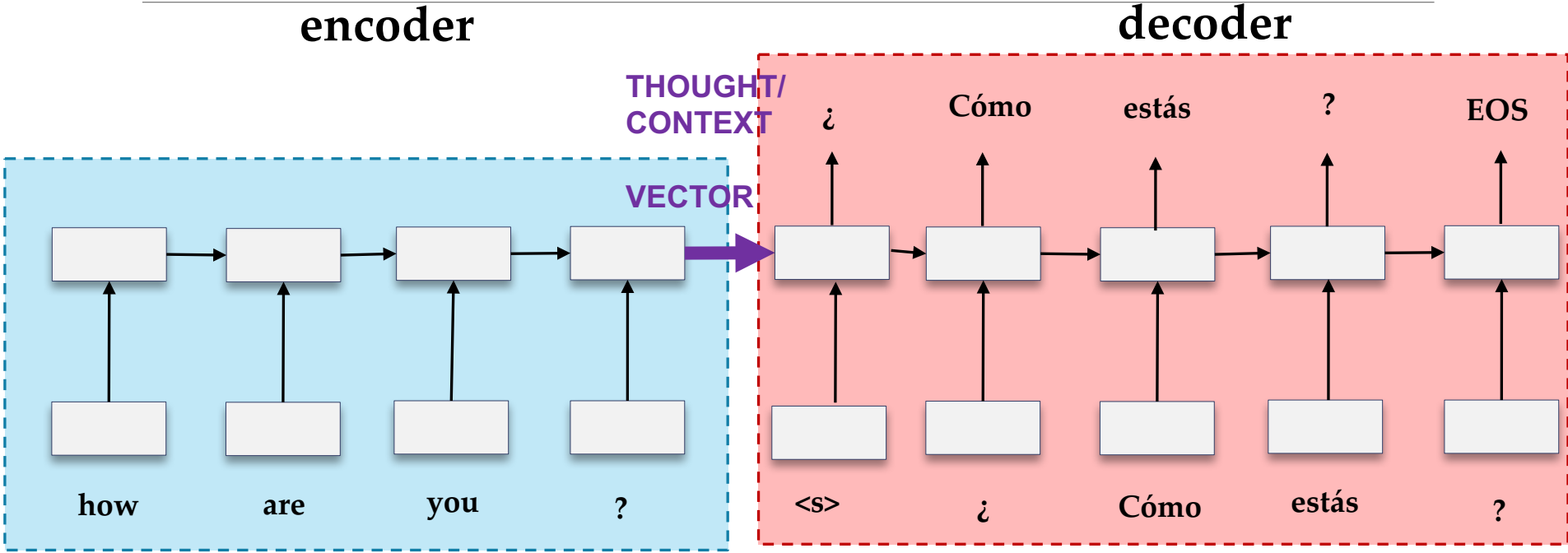
Sequence modeling

Model the probability of sequences of words

From previous lecture… we model sequences ith RNNs



Sequence-to-sequence models

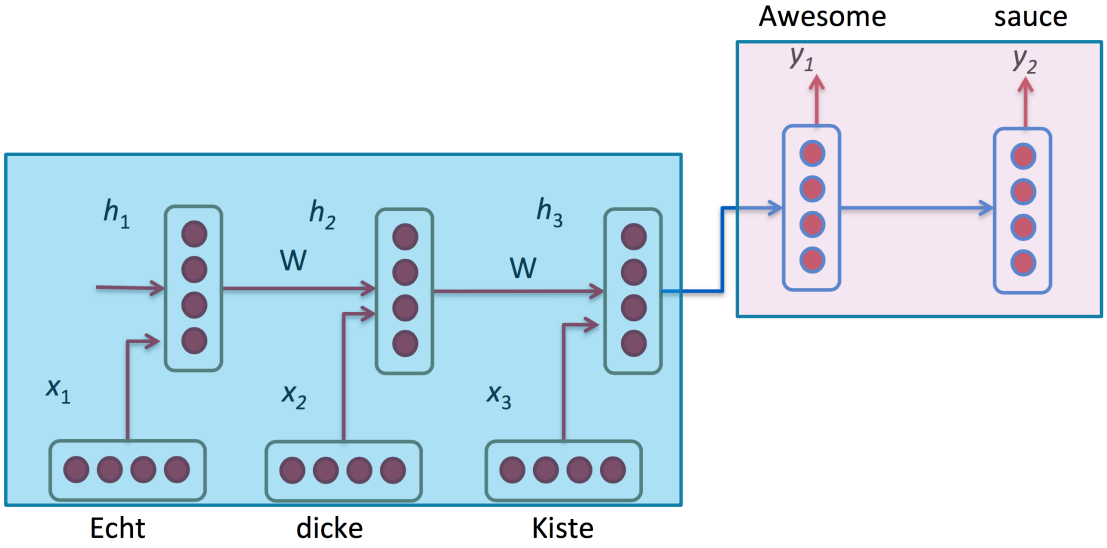


Any problem with these models?

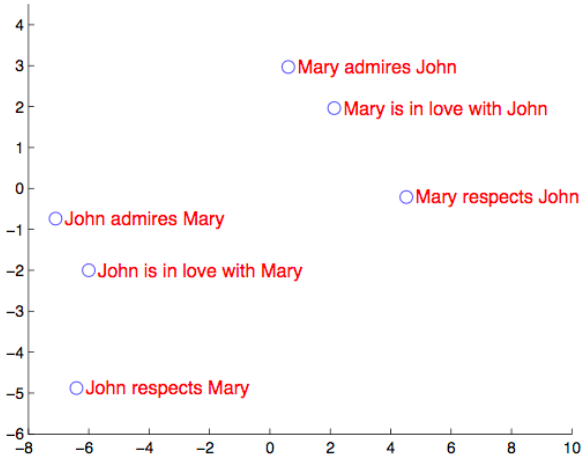
"You can't cram the meaning of a whole %&!$ing sentence into a single $&!*ing vector!"
— Ray Mooney

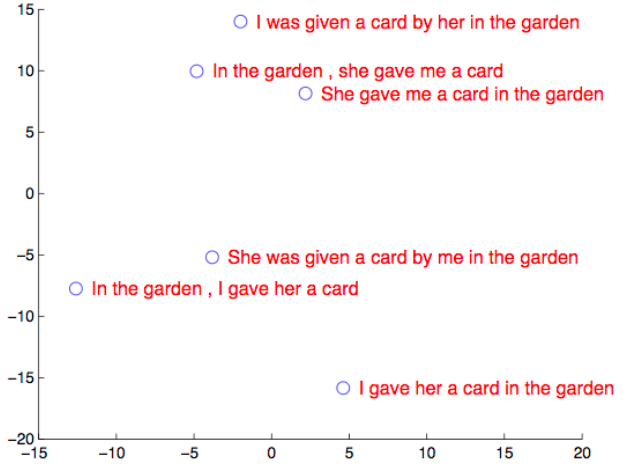
2. Attention-based mechanism

Motivation in the case of MT

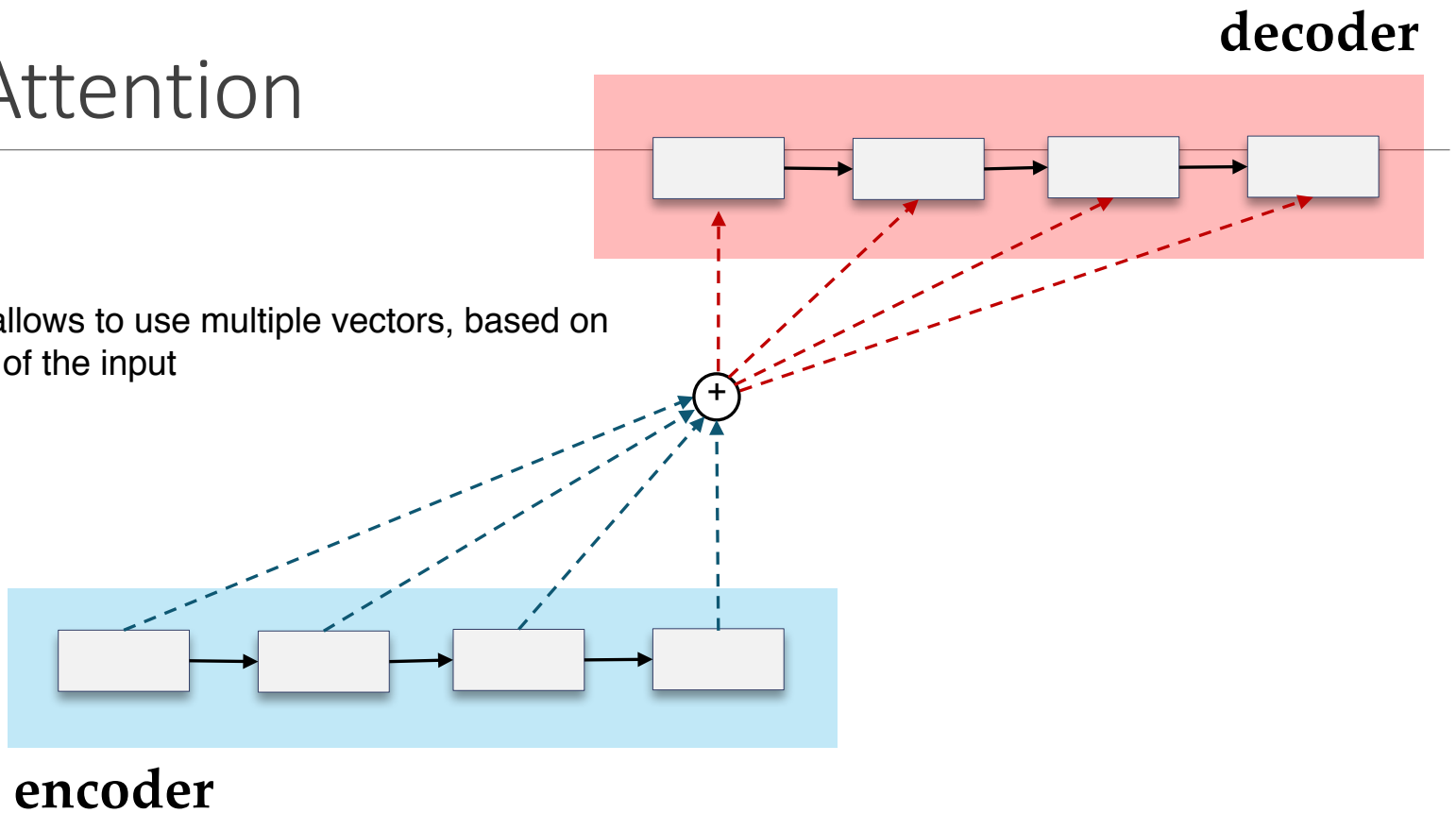


Motivation in the case of MT





Attention



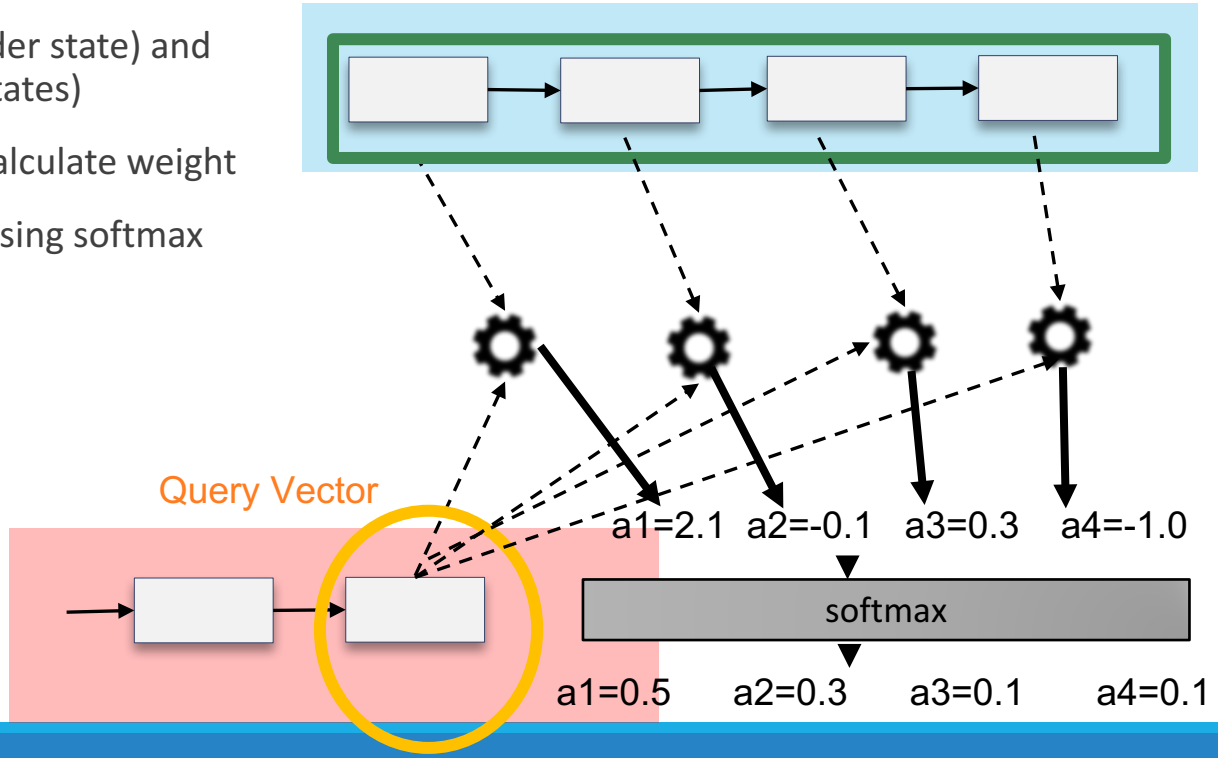
Attention allows to use multiple vectors, based on the length of the input

Attention Key Ideas

- Encode each word in the input and output sentence into a vector
- When decoding, perform a linear combination of these vectors, weighted by “attention weights”
- Use this combination in picking the next word

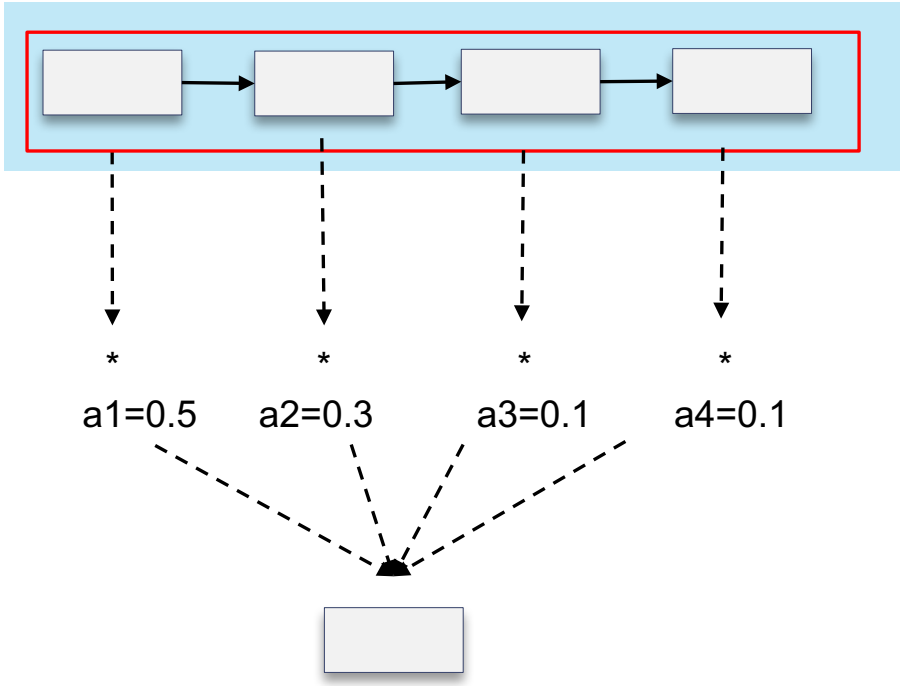
Attention computation I

- Use “query” vector (decoder state) and “key” vectors (all encoder states)
- For each query-key pair, calculate weight
- Normalize to add to one using softmax



Attention computation II

- Combine together value vectors (usually encoder states, like key vectors) by taking the weighted sum

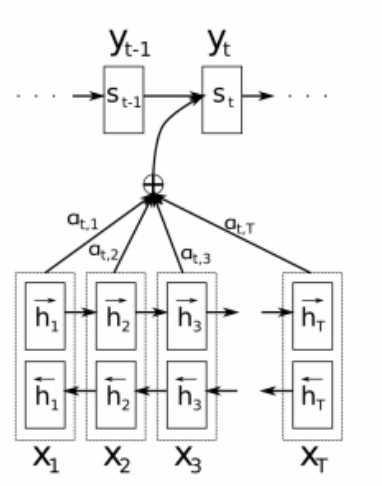


Attention Score Functions

q is the query and k is the key

			Reference
Multi-layer Perceptron	$a(q,k) = \tanh(\mathcal{W}_1[q,k])$	Flexible, often very good with large data	Bahdanau et al., 2015
Bilinear	$a(q,k) = q^T\mathcal{W}k$		Luong et al 2015
Dot Product	$a(q,k) = q^Tk$	No parameters! But requires sizes to be the same	Luong et al. 2015
Scaled Dot Product	$a(q,k) = \frac{q^Tk}{\sqrt{\lvert k\rvert}}$	Scale by size of the vector	Vaswani et al. 2017

Attention Integration

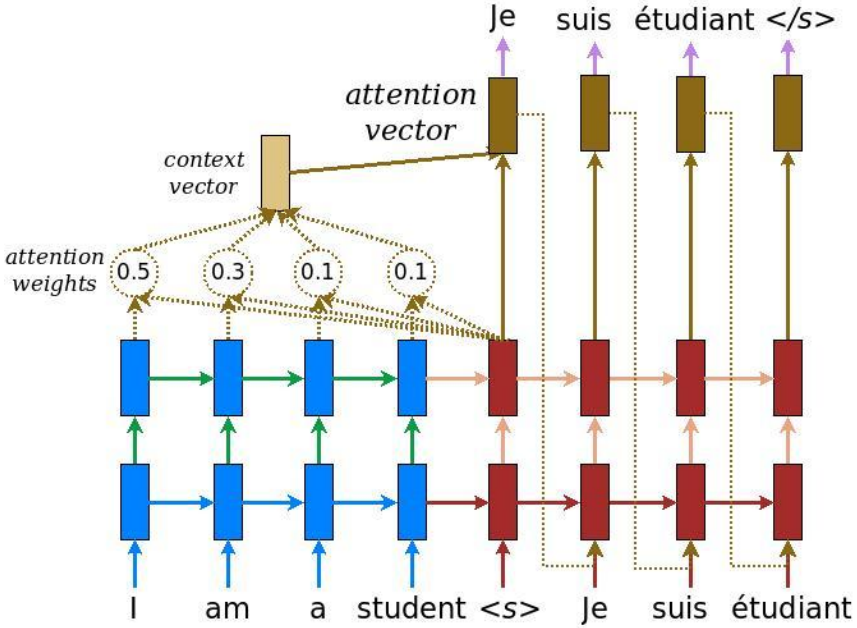


$$\alpha_{ts} = \frac{\exp\left(\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s)\right)}{\sum_{s'=1}^{S} \exp\left(\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_{s'})\right)} \quad \text{[Attention weights]} \quad (1)$$

$$\boldsymbol{c}_t = \sum_s \alpha_{ts} \bar{\boldsymbol{h}}_s \quad \text{[Context vector]} \quad (2)$$

$$\boldsymbol{a}_t = f(\boldsymbol{c}_t, \boldsymbol{h}_t) = \tanh(\boldsymbol{W}_{\boldsymbol{c}}[\boldsymbol{c}_t; \boldsymbol{h}_t]) \quad \text{[Attention vector]} \quad (3)$$

Attention Integration



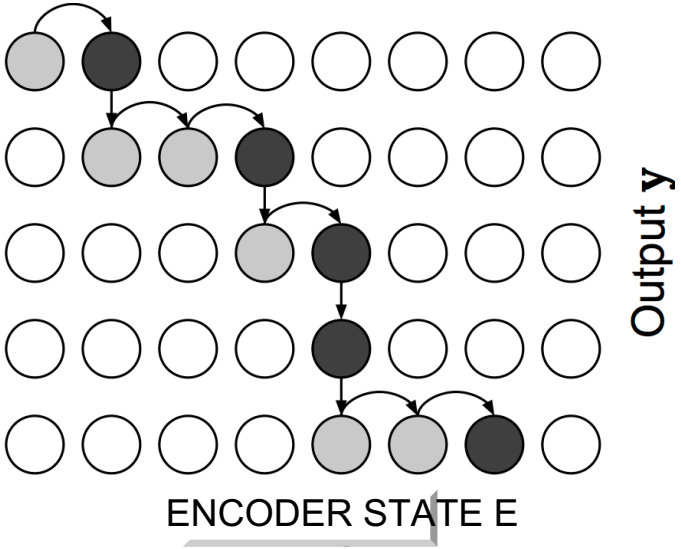
3. Attention Varieties

Hard Attention

* Instead of a soft interpolation, make a zero-one decision about where to attend (Xu et al. 2015)

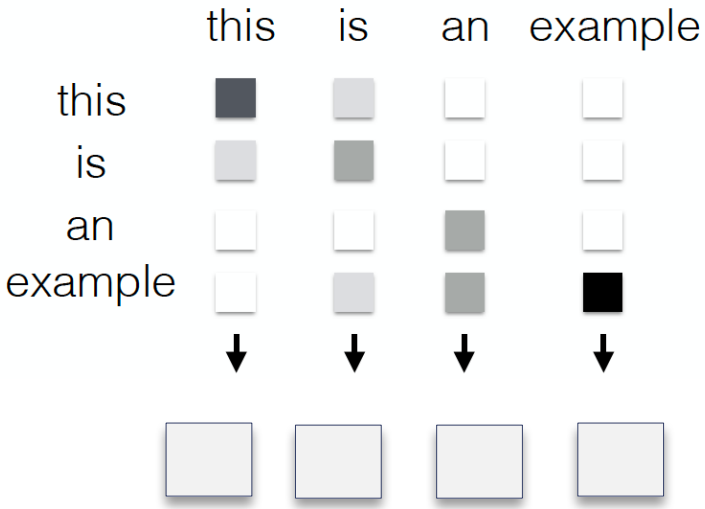
Monotonic Attention

This approach "softly" prevents the model from assigning attention probability before where it attended at a previous timestep by taking into account the attention at the previous timestep.



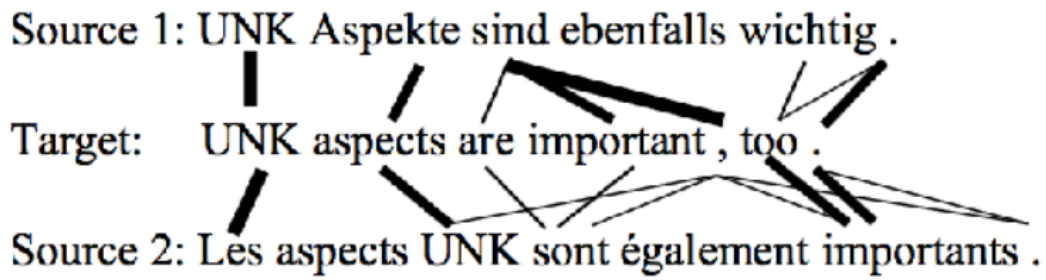
Intra-Attention / Self- Attention

Each element in the sentence attends to other elements from the SAME sentence → context sensitive encodings!

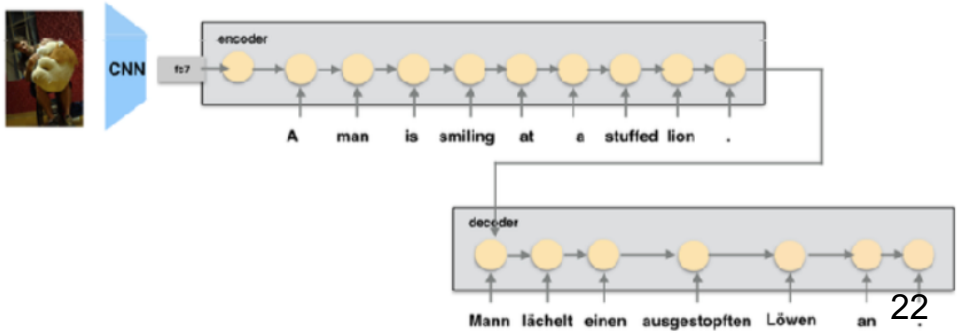


Multiple Sources

Attend to multiple sentences (Zoph et al., 2015)



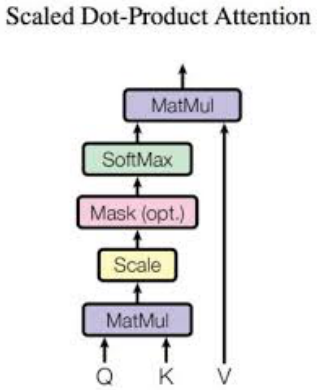
Attend to a sentence and an image (Huang et al. 2016)



Multi-headed Attention I

Multiple attention “heads” focus on different parts of the sentence

$$a(q,k) = \frac{q^T k}{\sqrt{|k|}}$$

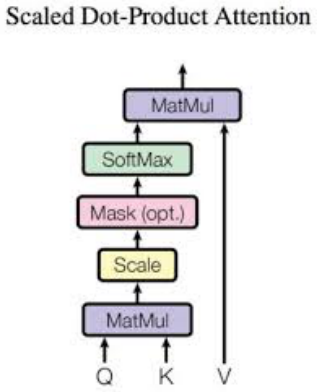


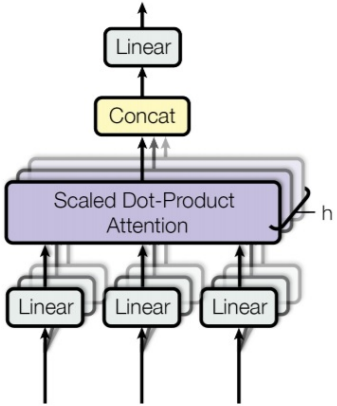
Multi-headed Attention II

Multiple attention “heads” focus on different parts of the sentence

E.g. Multiple independently learned heads (Vaswani et al. 2017)

$$a(q,k) = \frac{q^T k}{\sqrt{|k|}}$$





4.Improvements in Attention

IN THE CONTEXT OF MT

Coverage

Problem: Neural models tends to drop or repeat content

In MT,

1.Over-translation: some words are unnecessarily translated for multiple times;

2. Under-translation: some words are mistakenly untranslated.

SRC: Señor Presidente, abre la sesión.

TRG: Mr President Mr President Mr President.

Solution: Model how many times words have been covered e.g. maintaining a coverage vector to keep track of the attention history (Tu et al., 2016)

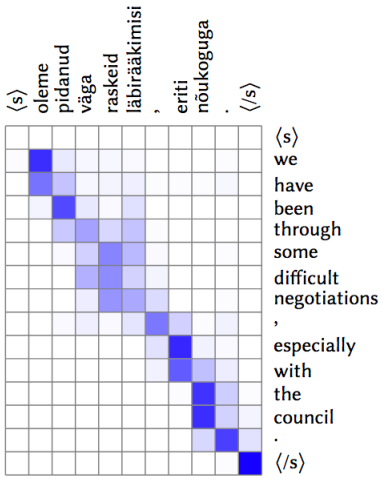
Modeling Coverage for Neural Machine Translation

Zhaopeng Tu[†] Zhengdong Lu[†] Yang Liu[‡] Xiaohua Liu[†] Hang Li[†]

Incorporating Markov Properties

Intuition: Attention from last time tends to be correlated with attention this time

Approach: Add information about the last attention when making the next decision



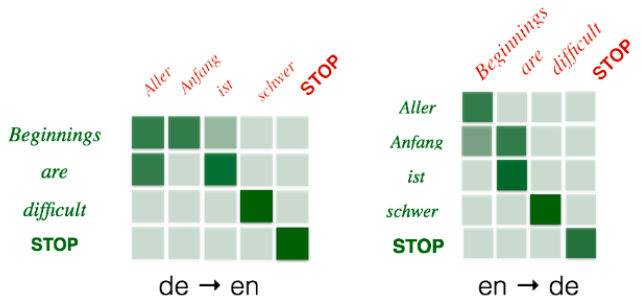
Incorporating Structural Alignment Biases into an Attentional Neural Translation Model

Trevor Cohn and **Cong Duy Vu Hoang** and **Ekaterina Vymolova**

Bidirectional Training

-Background: Established that for latent variable translation models the alignments improve if both directional models are combined (koehn et al, 2005)

-Approach: joint training of two directional models



Incorporating Structural Alignment Biases into an Attentional Neural Translation Model

Trevor Cohn and **Cong Duy Vu Hoang** and **Ekaterina Vymolova**

Supervised Training

Sometimes we can get “gold standard” alignments a –priori

- Manual alignments
- Pre-trained with strong alignment model

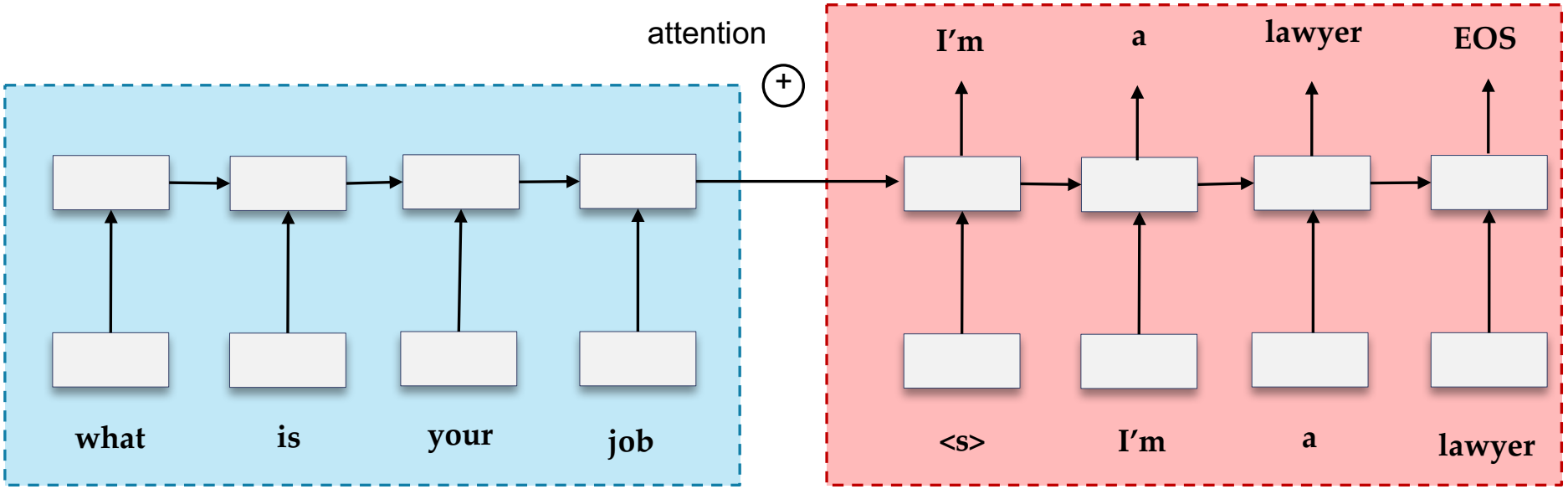
Train the model to match these strong alignments

5. Applications

Chatbots

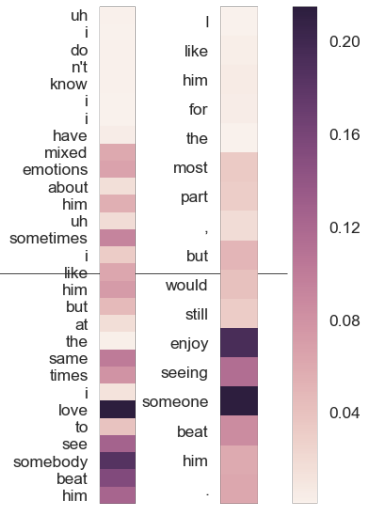
Human: what is your job
Enc-dec: i'm a lawyer
Human: what do you do ?
Enc-dec: i'm a doctor .

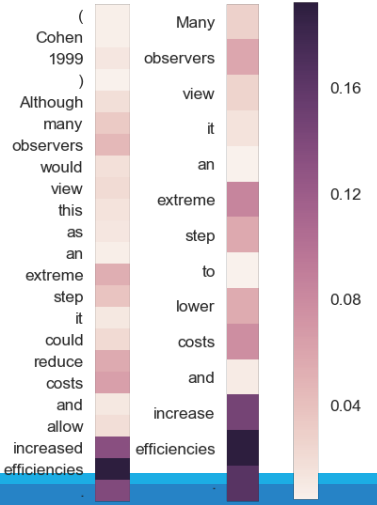
a computer program that conducts a conversation



Natural Language Inference

Caption	A person in a black wetsuit is surfing a small wave.
Entailment	A person is surfing a wave.
Contradiction	A woman is trying to sleep on her bed.
Neutral	A person surfing a wave in Hawaii.





Character-level Intra Attention Network for Natural Language Inference

Han Yang and **Marta R. Costa-jussà** and **José A. R. Fonollosa**

Other NLP Tasks

***Text summarization**:* process of shortening a text document with software to create a summary with the major points of the original document.

Question Answering: automatically producing an answer to a question given a corresponding document.

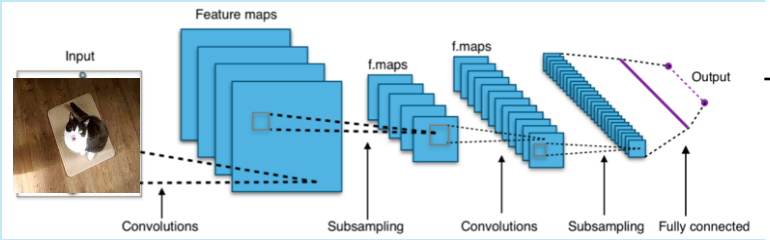
Semantic Parsing: mapping natural language into a logical form that can be executed on a knowledge base and return an answer

Syntactic Parsing: process of analysing a string of symbols, either in natural language or in computer languages, conforming to the rules of a formal grammar

Image captioning I

A cat on the mat

encoder



decoder

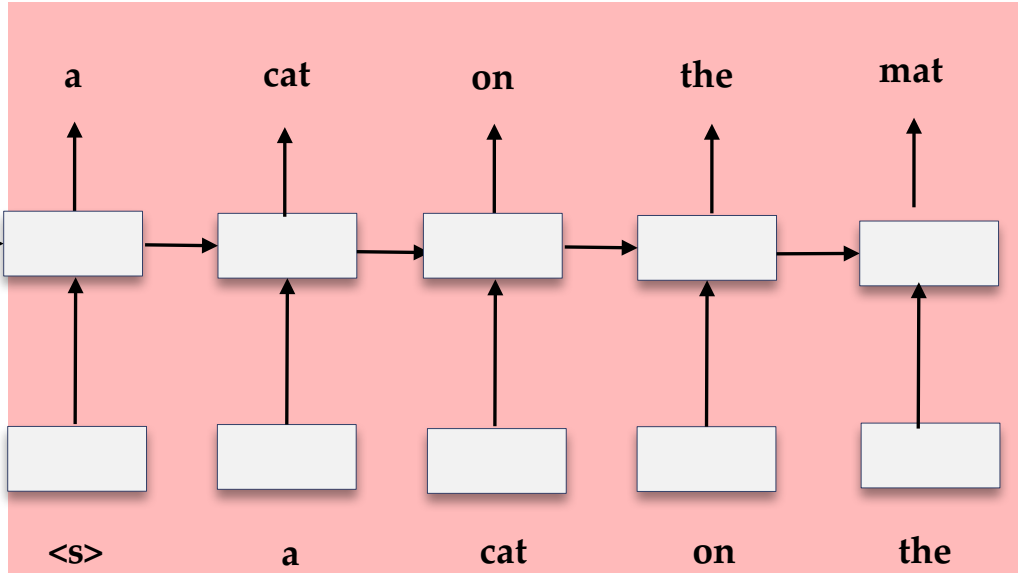
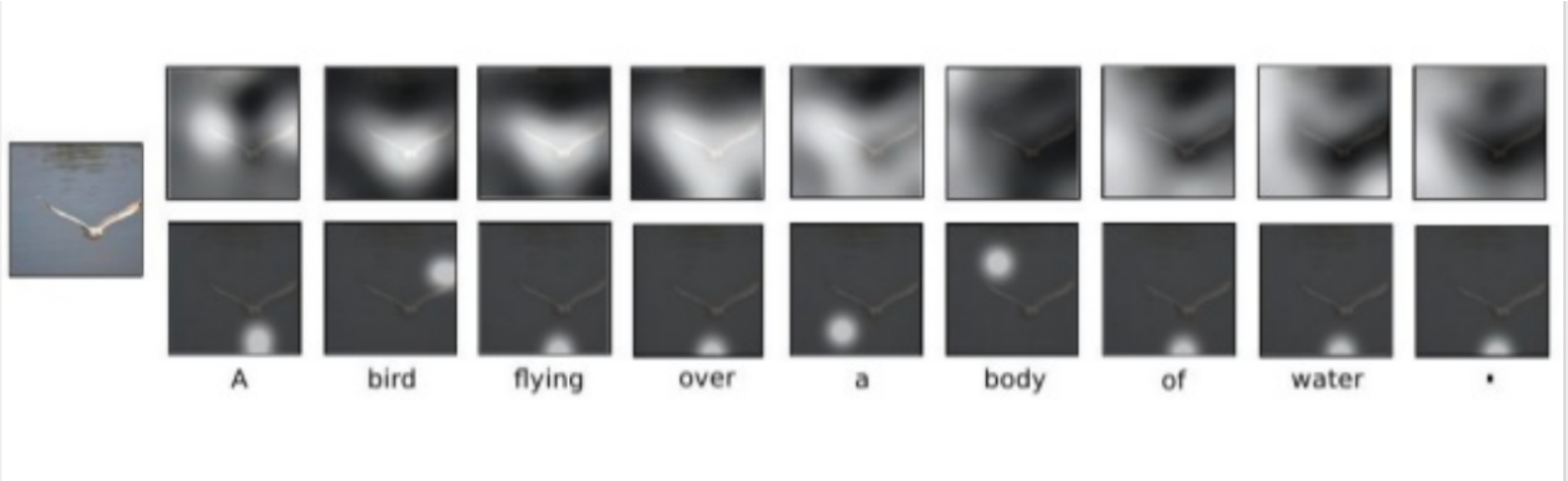


Image Captioning II

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

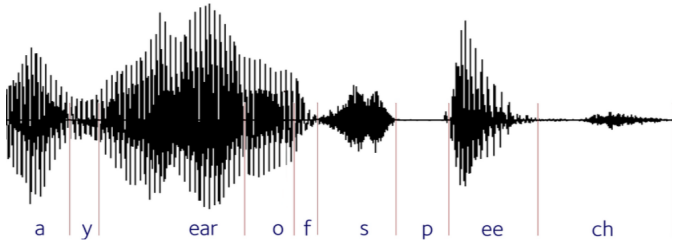


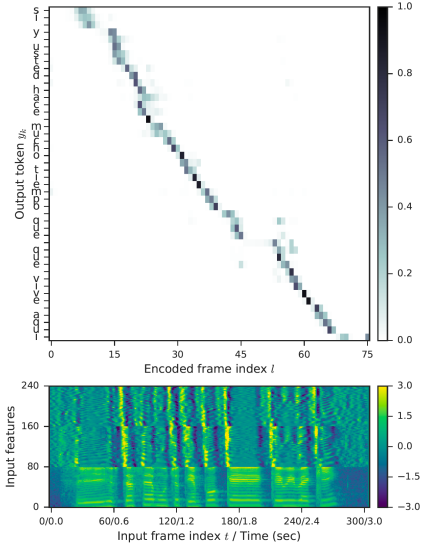
Other Computer Vision Tasks with Attention

Visual Question Answering: given an image and a natural language question about the image, the task is to provide an accurate natural language answer.

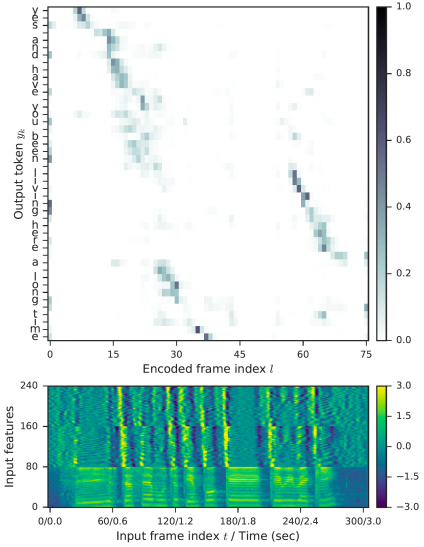
Video Caption Generation: attempts to generate a complete and natural sentence, enriching the single label as in video classification, to capture the most informative dynamics in videos.

Speech recognition / translation





(a) *Spanish speech recognition decoder attention.*



(b) *Spanish-to-English speech translation decoder attention.*

6. “Attention is all you need”

SLIDES BASED ON HTTPS://RESEARCH.GOOGLEBLOG.COM/2017/08/TRANSFORMER-NOVEL-NEURAL-NETWORK.HTML

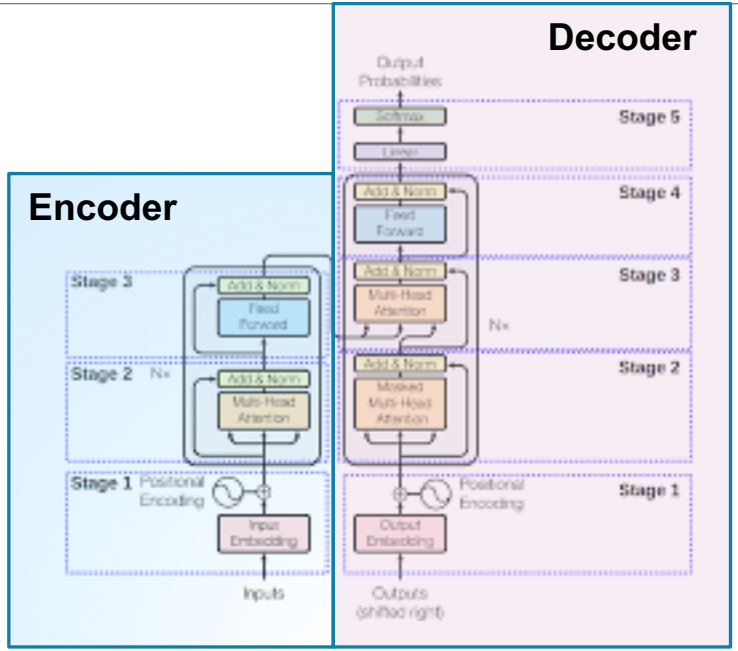
Motivation

Sequential nature of RNNs -→ difficult to take advantage of modern computing devices such as TPUs (Tensor Processing Units)

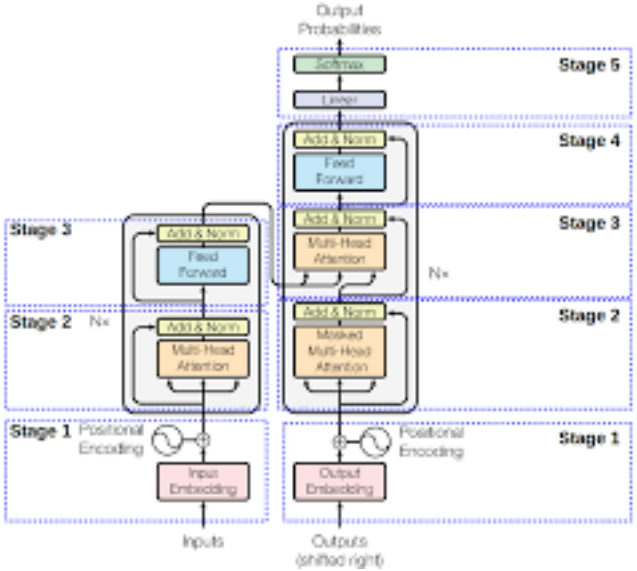
Transformer

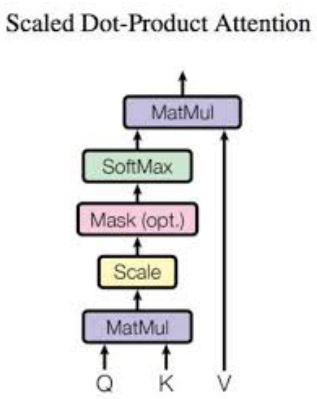
I arrived at the bank after crossing the river

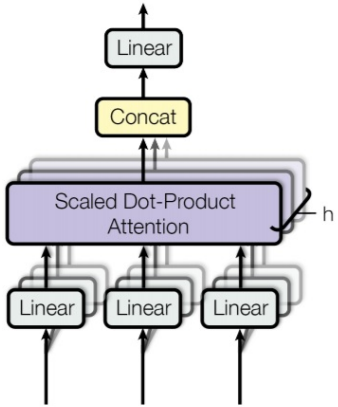
Transformer I



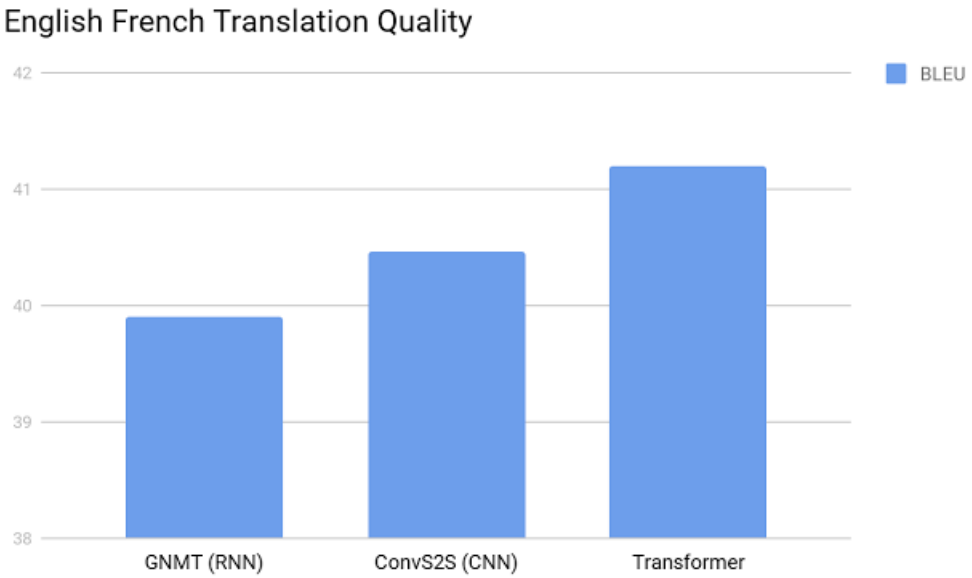
Transformer II







Transformer results

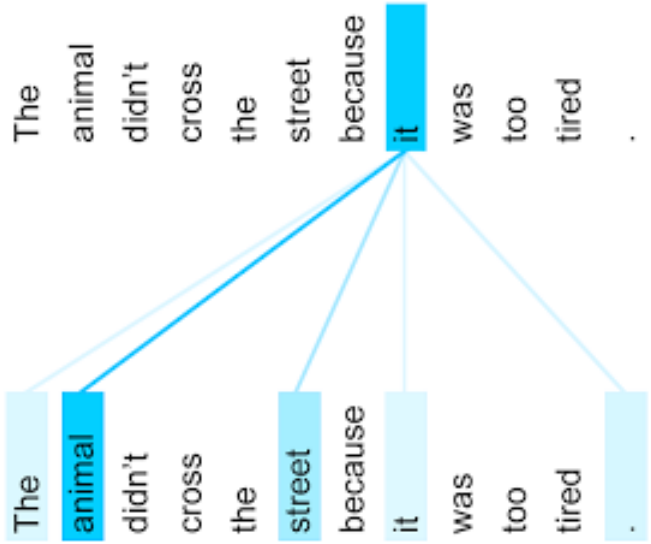


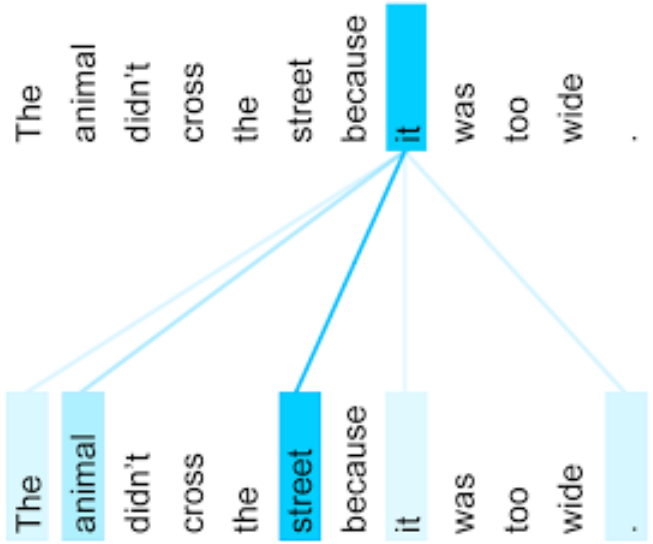
Attention weights

The animal didn't cross the street because it was too tired.
L'animal n'a pas traversé la rue parce qu'il était trop fatigué.

The animal didn't cross the street because it was too wide.
L'animal n'a pas traversé la rue parce qu'elle était trop large.

Attention weights





7. Summary

RNNs and Attention

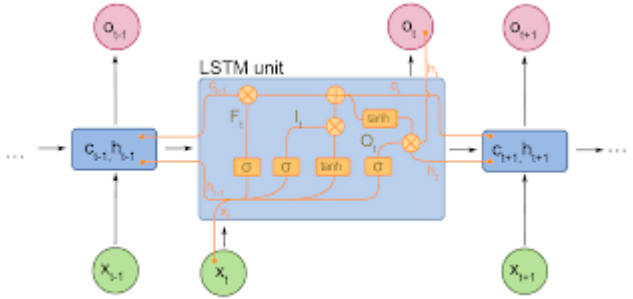
RNNs are used to model sequences

Attention is used to enhance modeling long sequences

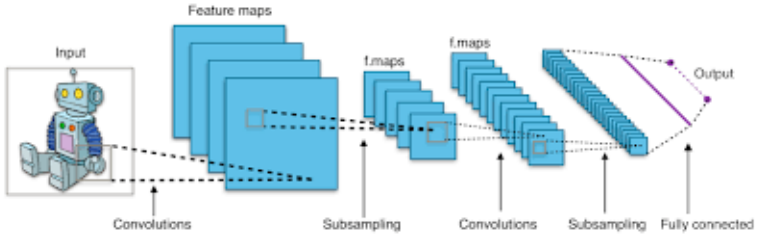
Versatility of these models allows to apply them to **a wide range of applications**

Implementations of Encoder-Decoder

LSTM



CNN

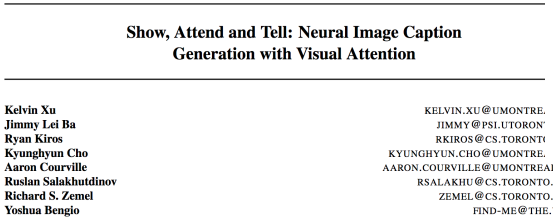


Attention-based mechanisms

Soft vs Hard: soft attention weights all pixels, hard attention crops the image and forces attention only on the kept part.

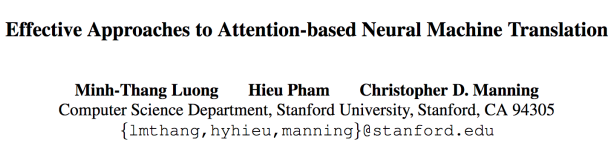
Global vs Local: a global approach which always attends to all source words and a local one that only looks at a subset of source words at a time.

Intra vs External: intra attention is within the encoder's input sentence, external attention is across sentences.



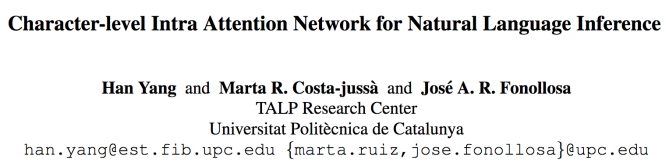
Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

Kelvin Xu KELVIN.XU@UMONTRE.
Jimmy Lei Ba JIMMY@PSI.UTORON
Ryan Kiros RKIROS@CS.TORONT
Kyunghyun Cho KYUNGHYUN.CHO@UMONTRE.
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



Effective Approaches to Attention-based Neural Machine Translation

Minh-Thang Luong Hieu Pham Christopher D. Manning
Computer Science Department, Stanford University, Stanford, CA 94305
{lmthang,hyhieu,manning}@stanford.edu

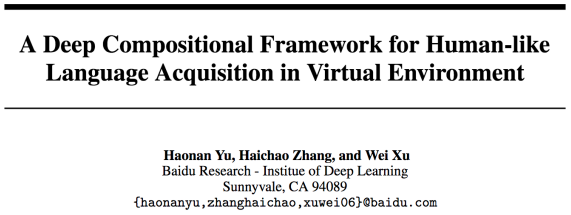


Character-level Intra Attention Network for Natural Language Inference

Han Yang and Marta R. Costa-jussà and José A. R. Fonollosa
TALP Research Center
Universitat Politècnica de Catalunya
han.yang@est.fib.upc.edu {marta.ruiz,jose.fonollosa}@upc.edu

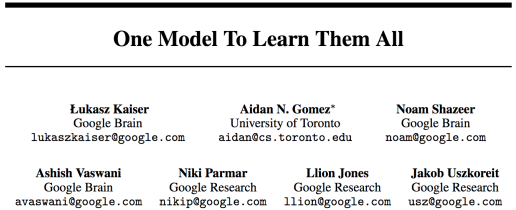
One large encoder-decoder

- Text, speech, image… is all **converging** to a signal paradigm?
- If you know how to build a neural MT system, you may easily learn how to build a speech-to-text recognition system...
- Or you may train them together to achieve **zero-shot** AI.



A Deep Compositional Framework for Human-like Language Acquisition in Virtual Environment

Haonan Yu, Haichao Zhang, and Wei Xu
Baidu Research - Institue of Deep Learning
Sunnyvale, CA 94089
{haonanyu,zhanghaichao,xuwei06}@baidu.com



One Model To Learn Them All

Łukasz Kaiser
Google Brain
lukaszkaiser@google.com

Aidan N. Gomez*
University of Toronto
aidan@cs.toronto.edu

Noam Shazeer
Google Brain
noam@google.com

Ashish Vaswani
Google Brain
avaswani@google.com

Niki Parmar
Google Research
nikip@google.com

Llion Jones
Google Research
llion@google.com

Jakob Uszkoreit
Google Research
usz@google.com

**And other references on this research direction….*

Research going on… Interested?

marta.ruiz@upc.edu

Q&A ?

Quizz

1.Mark all statements that are true

A. Sequence modeling only refers to language applications

B. The attention mechanism can be applied to an encoder-decoder architecture

C. Neural machine translation systems require recurrent neural networks

D. If we want to have a fixed representation (thought vector), we can not apply attention-based mechanisms

2. Given the query vector q=[], the key vector 1 k1=[] and the key vector 2 k2=[].

A. What are the attention weights 1 & 2 computing the dot product?

B. And when computing the scaled dot product?

C. To what key vector are we giving more attention?

D. What is the advantage of computing the scaled dot product?