DLAI – Marta R. Costa-jussà

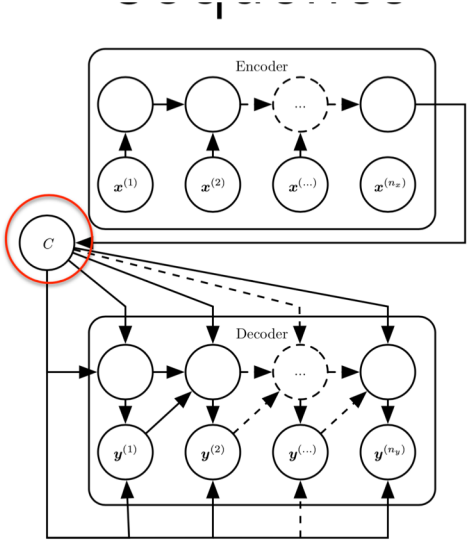
Attention

Outline

- Introduction
- Additive Attention
- Multiplicative Attention
- Improvements
- Applications

Introduction

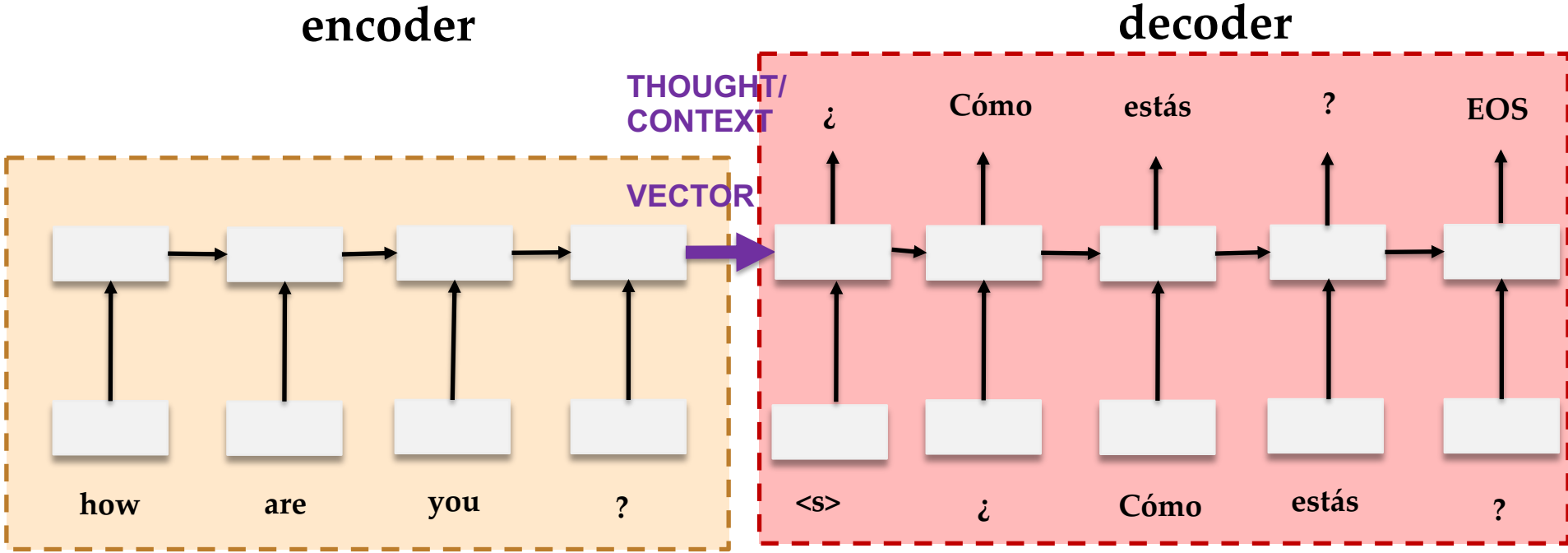
Sequence to sequence



HOT!!!

To learn the context Variable C which represents a semantic summary of the input sequence, and for later decoder RNNN

Sequence-to-sequence



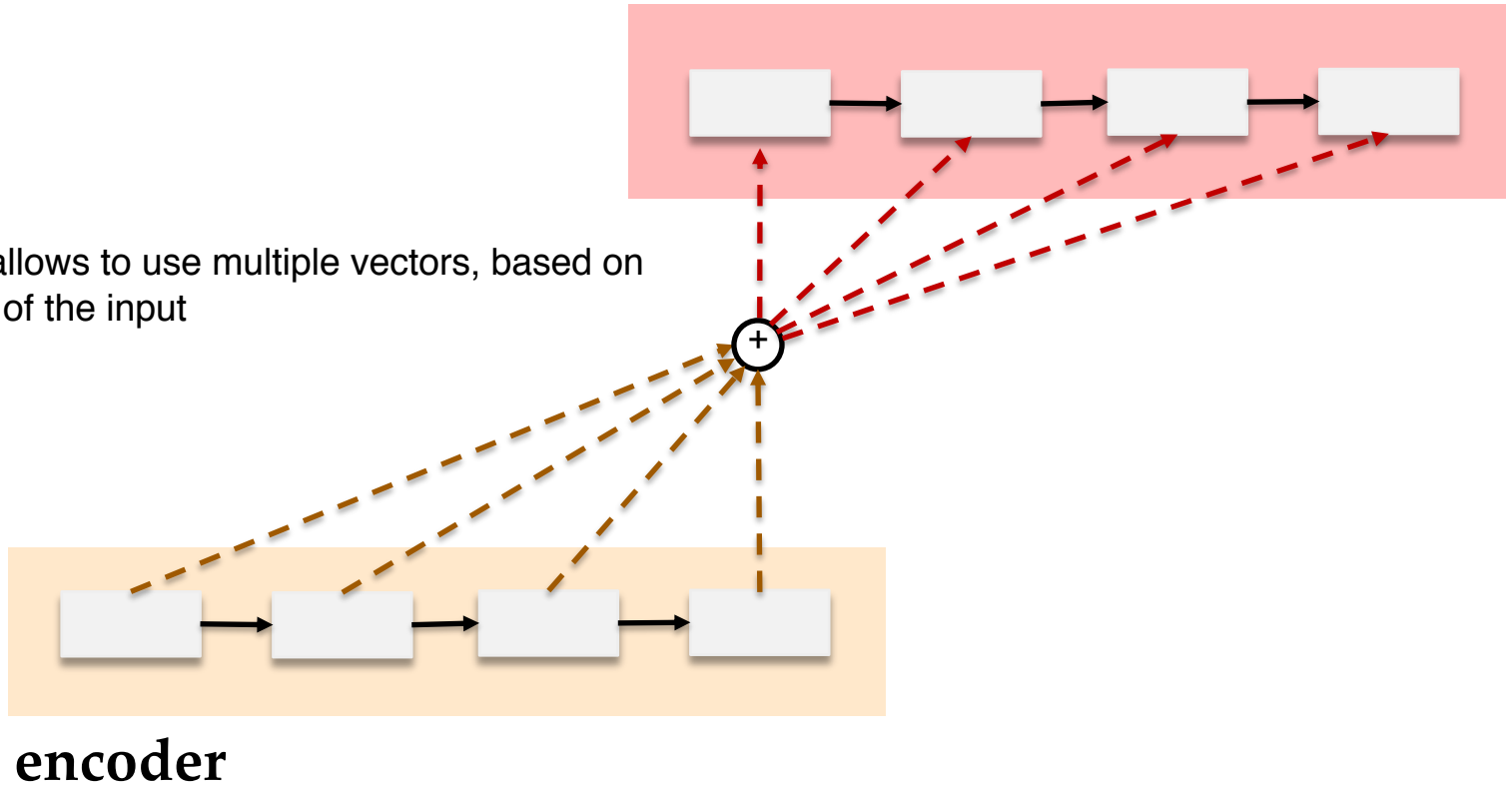
Any problem with these models?

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

Additive attention

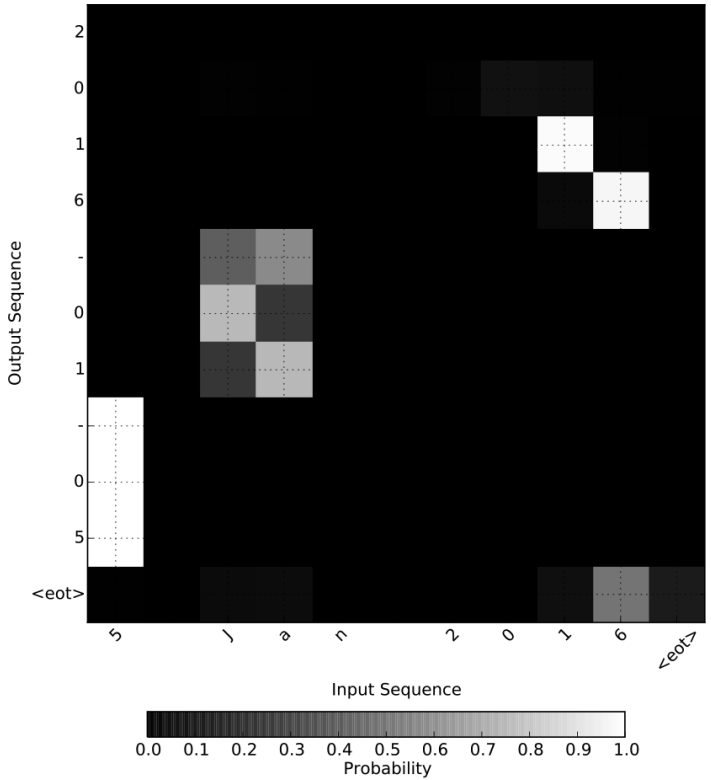
Attention

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



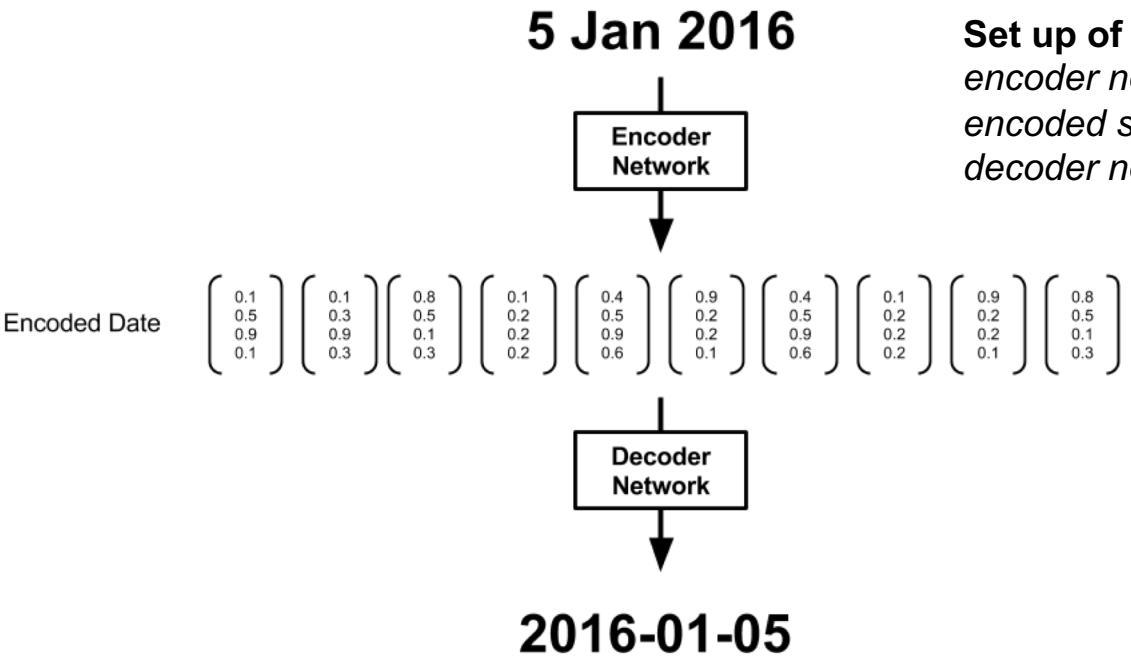
Attention Integration & Visualization Step-by-step

[from Medium]



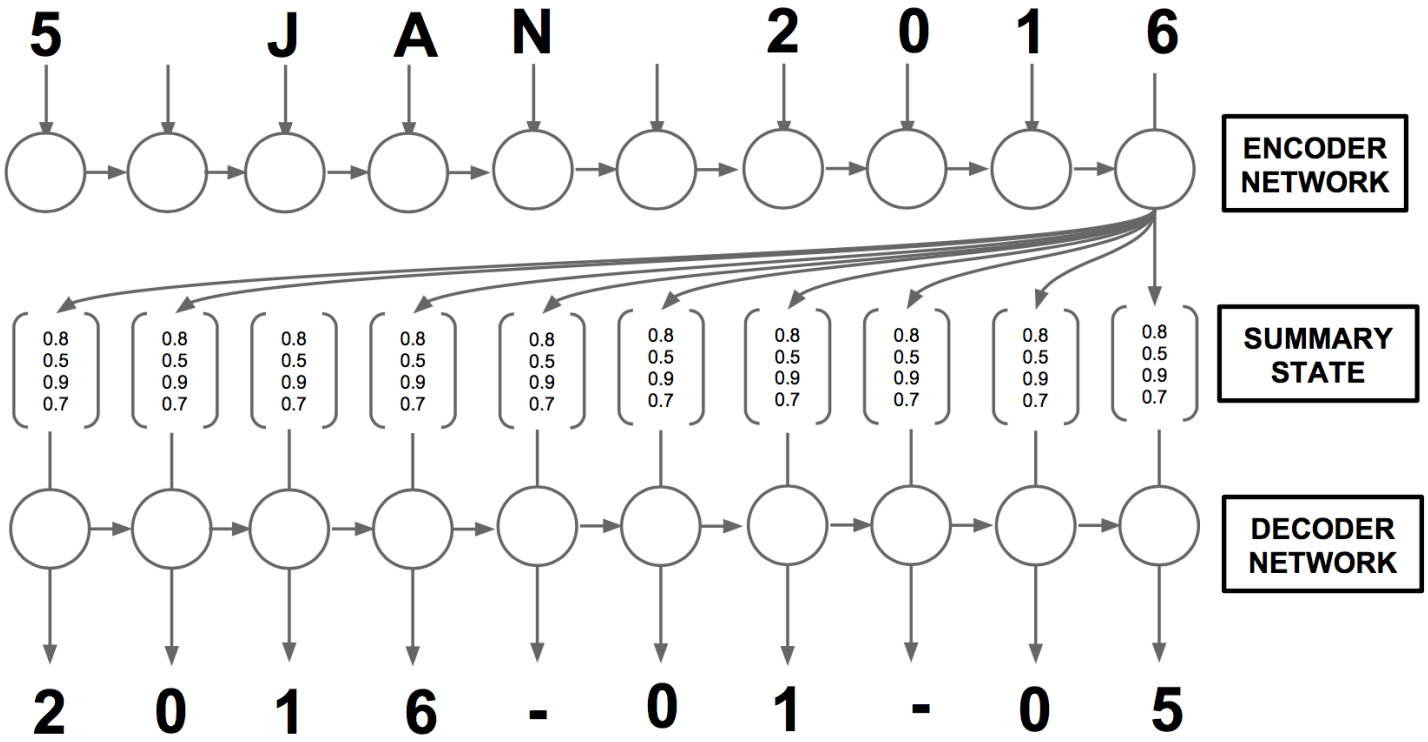
Attention map for the freeform date “5 Jan 2016”. We can see that the neural network used “16” to decide that the year was 2016, “Ja” to decide that the month was 01 and the first bit of the date to decide the day of the month.

Encoder-decoder architecture set up

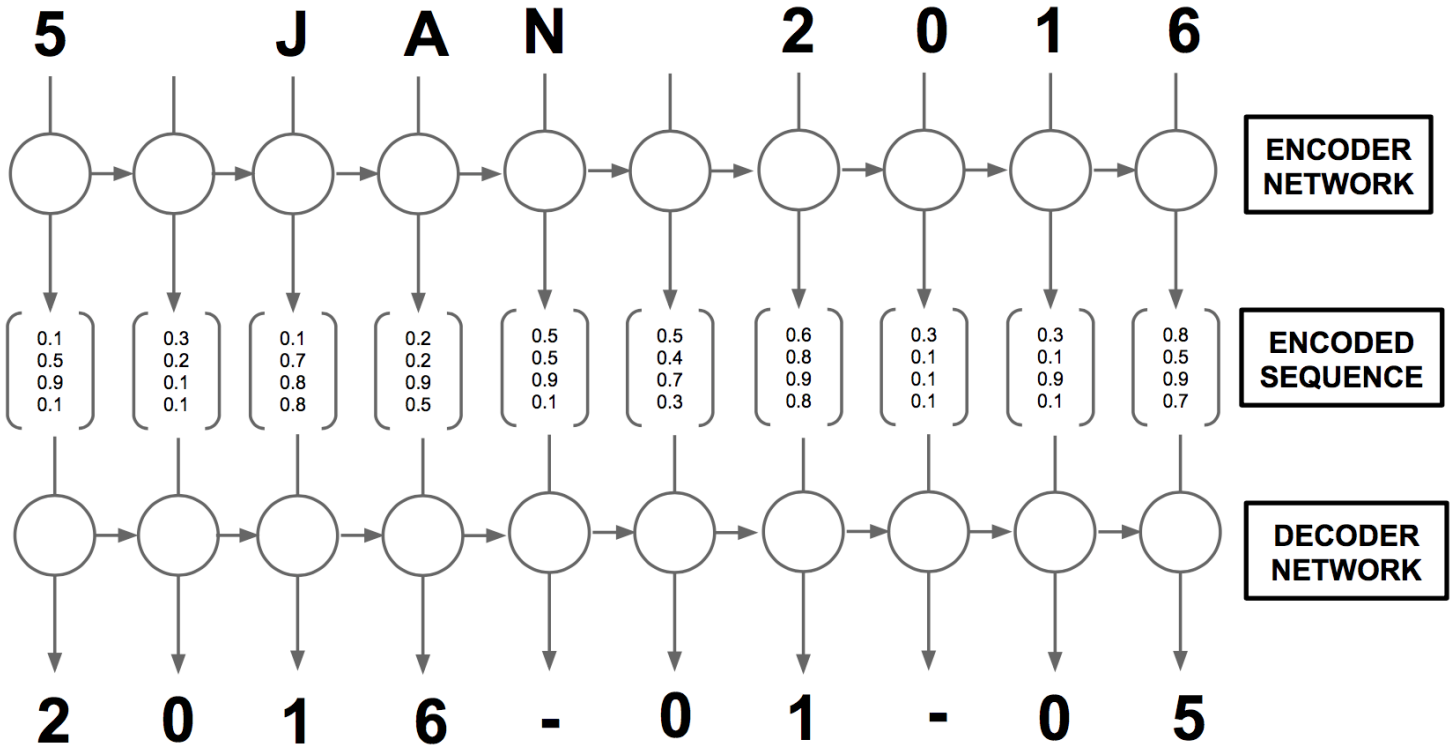


Set up of the encoder-decoder architecture. *The encoder network processes the input sequence into an encoded sequence which is subsequently used by the decoder network to produce the output.*

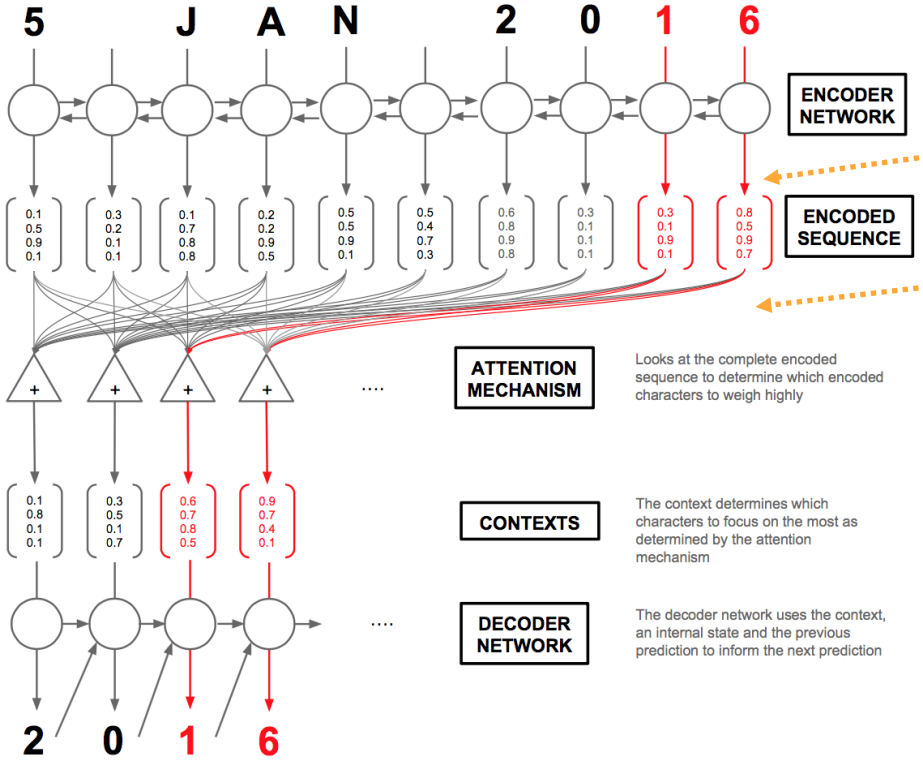
Use of a summary state in the encoder-decoder



Use of the complete encoded sequence in the decoder



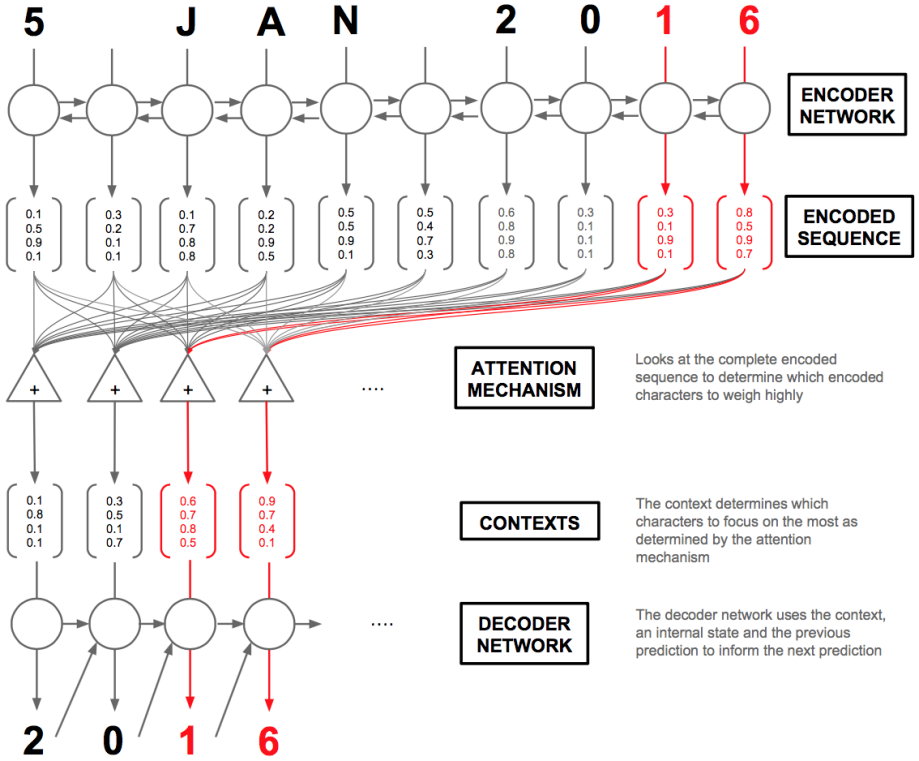
Overview of the attention mechanism



$$e_{j,t} = V_a \cdot \tanh(W_a s_{t-1} + U_a h_j)$$

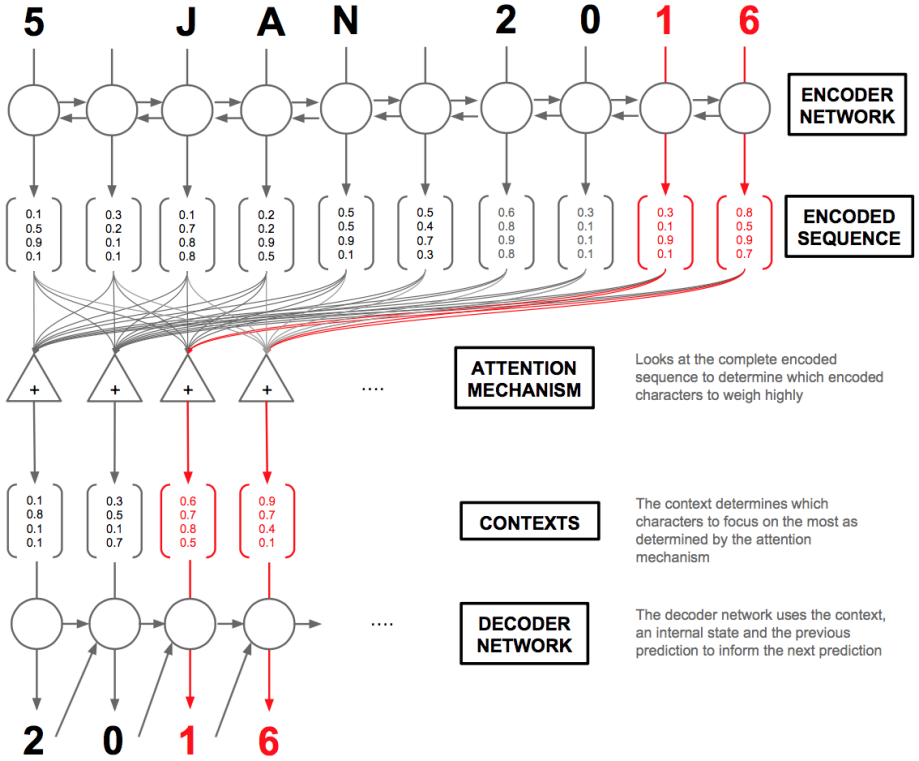
$$\alpha_{j,t} = \frac{\exp(e_j)}{\sum_{k=1}^{T} \exp(e_k)}$$

Overview of the attention mechanism



$$c_t = \sum_{k=1}^{T} \alpha_{k,t} h_k$$

Overview of the attention mechanism



$$r_t = \sigma(W_r y_{t-1} + U_r s_{t-1} + C_r c_t)$$

$$z_t = \sigma(W_z y_{t-1} + U_z s_{t-1} + C_z c_t)$$

$$\hat{s}_t = \tanh(W_p y_{t-1} + U_p [r_t \circ s_{t-1}] + C_p c_t)$$

$$s_t = (1 - z_t) \circ s_{t-1} + z_t \circ \hat{s}_t$$

$$y_t = \sigma(W_o y_{t-1} + U_o s_t + C_o c_t)$$

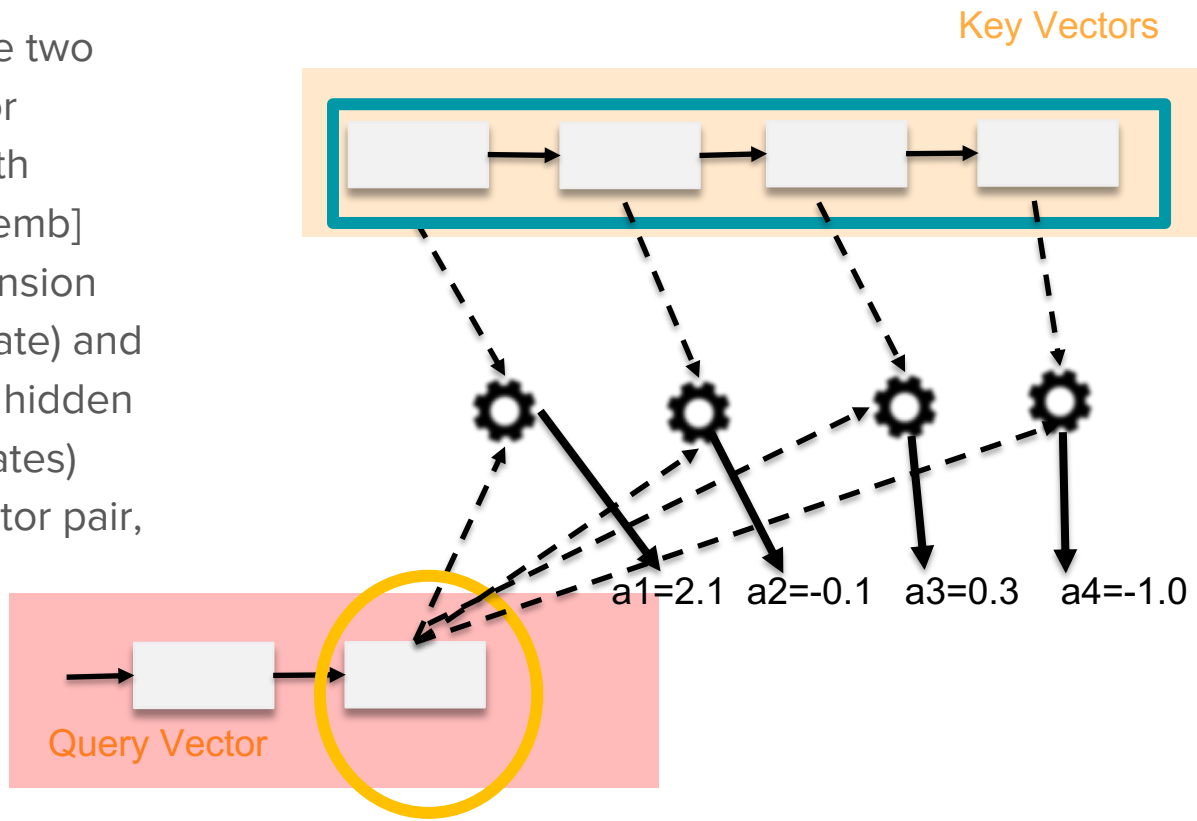
Multiplicative attention

Attention Key Ideas

- Encode each word in the 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: first step

- ”query” and “key” encode two vectors either the same or different [dimension length sentence x hidden units/emb]
- Use “query vector” [dimension hidden units] (decoder state) and “key vectors” [dimension hidden units/emb](all encoder states)
- For each “query-key” vector pair, calculate weight (value)



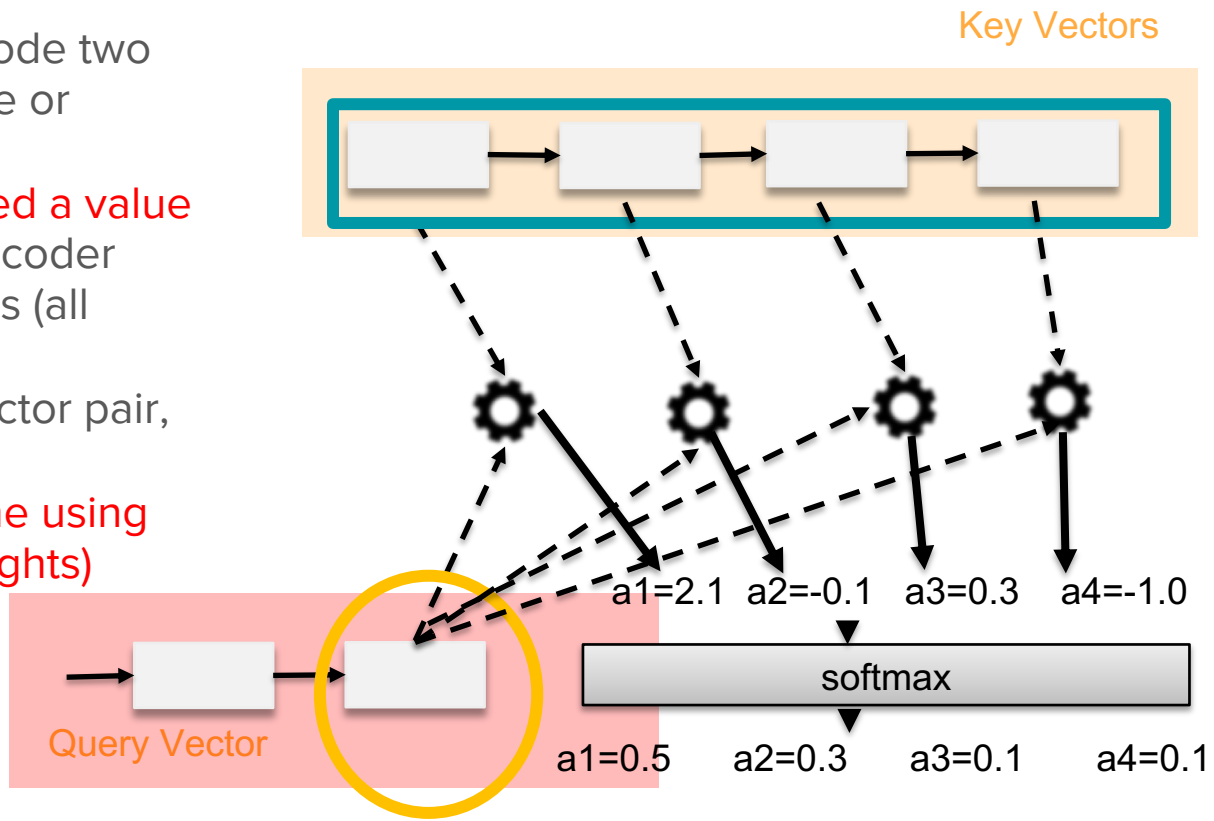
Attention Score Functions: second step

- q is the query and k is the key

			Reference
Multi-layer Perceptron	$a(q,k) = w_3^t \tanh(\mathcal{W}_1 q + \mathcal{W}_2 k)$	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	$a(q,k) = q^T k$	No parameters! But requires sizes to be the same	Luong et al. 2015
Scaled Dot Product	$a(q,k) = \frac{q^T k}{\sqrt{\lvert d_k \rvert}}$	Scale by dimension of the key vector	Vaswani et al. 2017

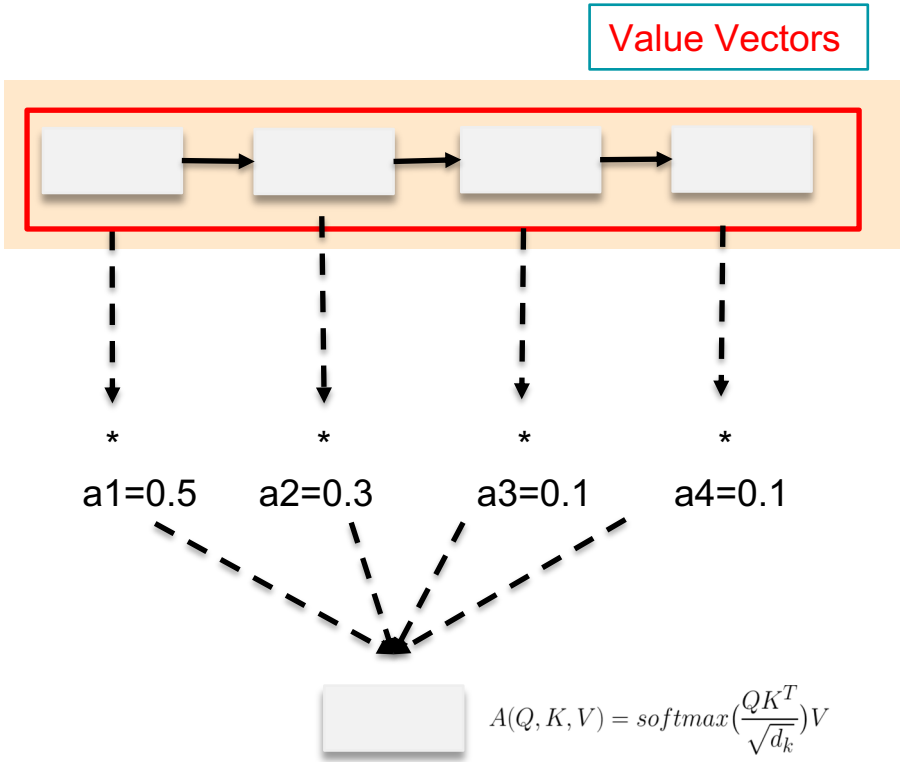
Attention computation: third step

- ”query” and “key” encode two vectors either the same or different
- Each key has associated a value
- Use “query” vector (decoder state) and “key” vectors (all encoder states)
- For each query-key vector pair, calculate weight
- Normalize to add to one using softmax (obtaining weights)



Attention computation: fourth step

- Combine together value vectors (usually encoder states, like key vectors) by taking the weighted sum
- [dimension sentence length x hidden units/emb]
- Use this in any part of your model



Wrap up: Additive vs Multiplicative

Additive

$$a(q,k) = w_3^t \tanh(\mathcal{W}_1 q + \mathcal{W}_2 k)$$

Pro's

performs better for larger dimensions

Multiplicative

$$a(q,k) = q^T \mathcal{W} k$$

Pro's

more efficient (matrix multiplication)

Question

Given the query vector q=[0.3,0.2,0.1], the key vector 1 k1=[0.1,0.3,0.1] and the key vector 2 k2=[0.6,0.4,0.2].

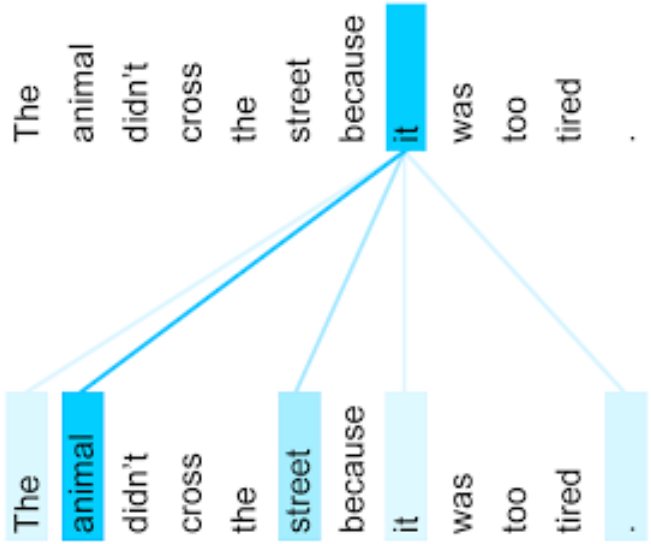
A. What are the attention weights 1&2 computing the dot product?
B. What are the attention weights 1&2 when computing the scaled dot product (dim of Key is 3)?
C. To what key vector are we giving more attention?

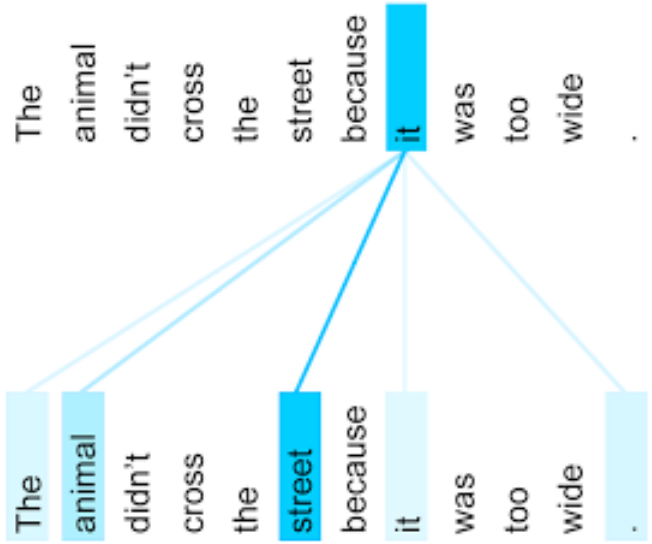
Attention weights

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

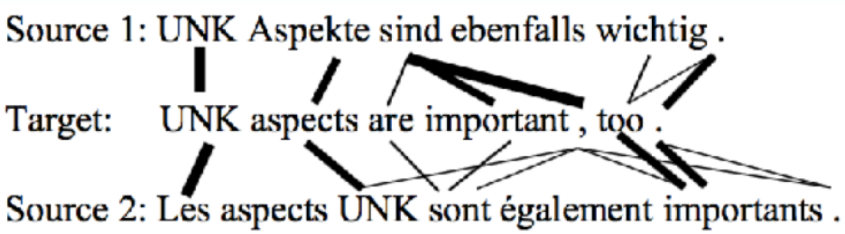




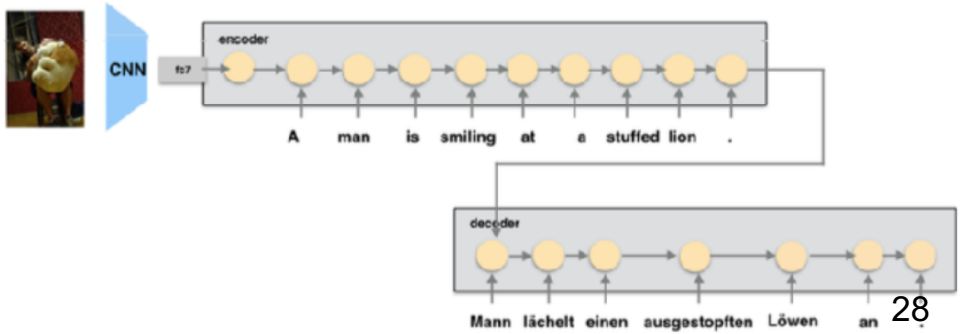
Improvements

Multiple Sources

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



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



Coverage

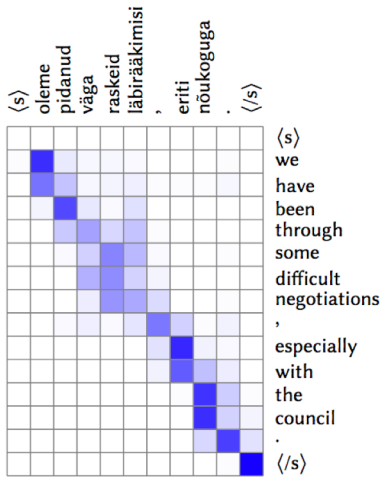
- **Problem:** Neural models tends to drop or repeat content
- In MT,
- 1.Over-translation: some words are unnecessarily translated for multiple times;
- 2. Under-translation: some words are mistakenly untranslated.
- SRC: Señor Presidente, abre la sesión.
- TRG: Mr President Mr President Mr President.
- **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



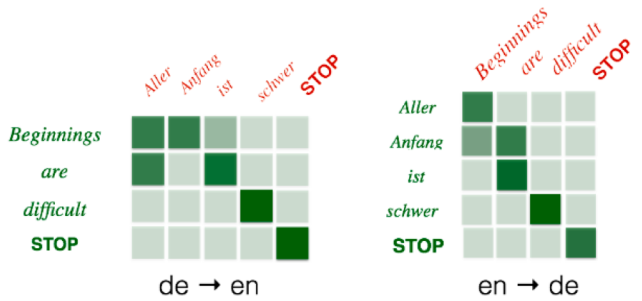
Incorporating Structural Alignment Biases into an Attentional Neural Translation Model

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

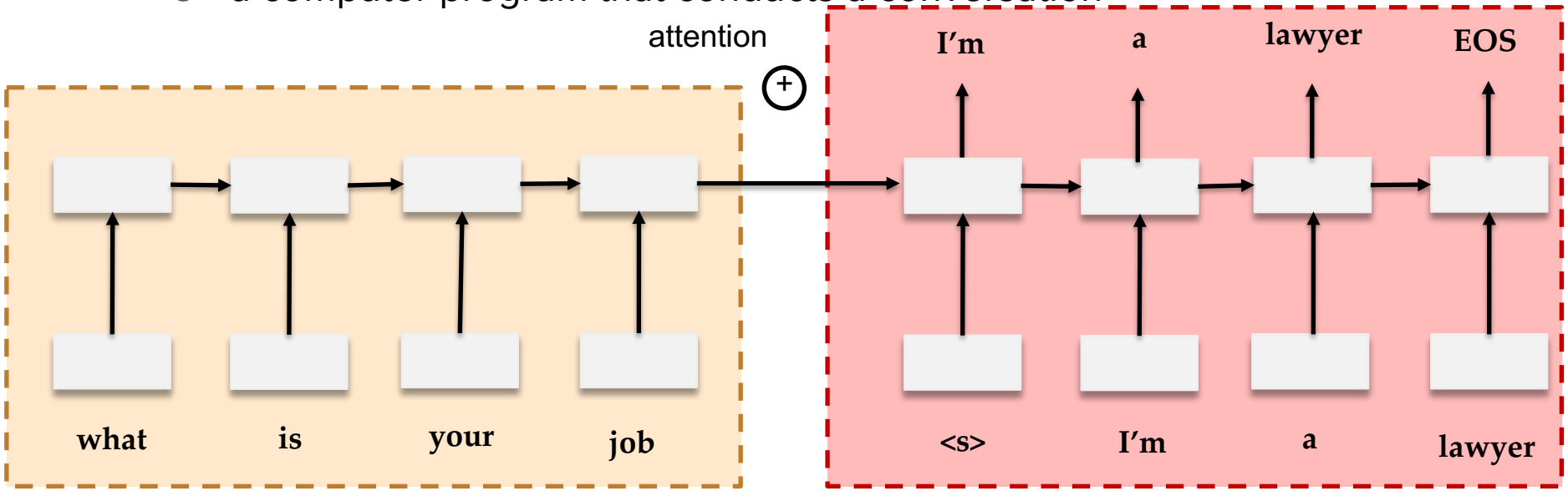
- Sometimes we can get “gold standard” alignments a –priori
  - Manual alignments
  - Pre-trained with strong alignment model
- Train the model to match these strong alignments

APPLICATIONS

Chatbots

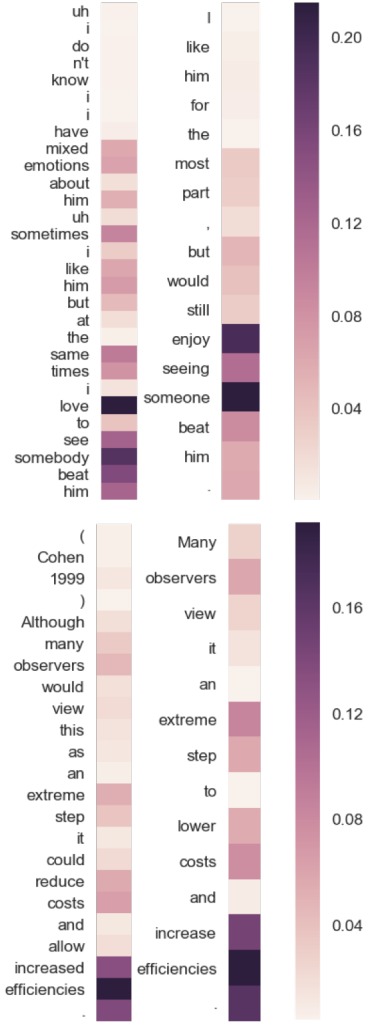
Human: what is your job
Enc-dec: i'm a lawyer
Human: what do you do ?
Enc-dec: i'm a doctor .

- 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.



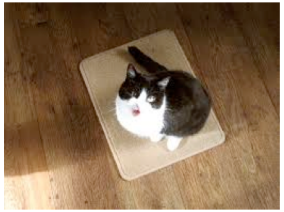
Character-level Intra Attention Network for Natural Language Inference

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

Other NLP Tasks

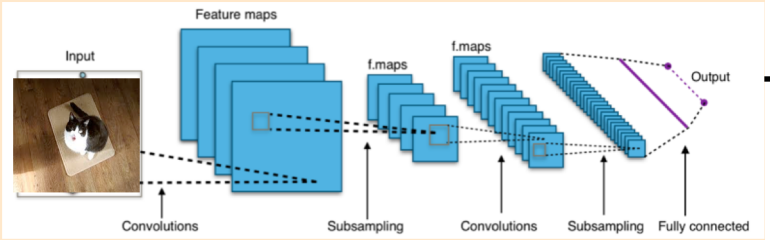
- ***Text summarization**:* process of shortening a text document with software to create a summary with the major points of the original document.
- ***Question Answering:*** automatically producing an answer to a question given a corresponding document.
- ***Semantic Parsing:*** mapping natural language 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

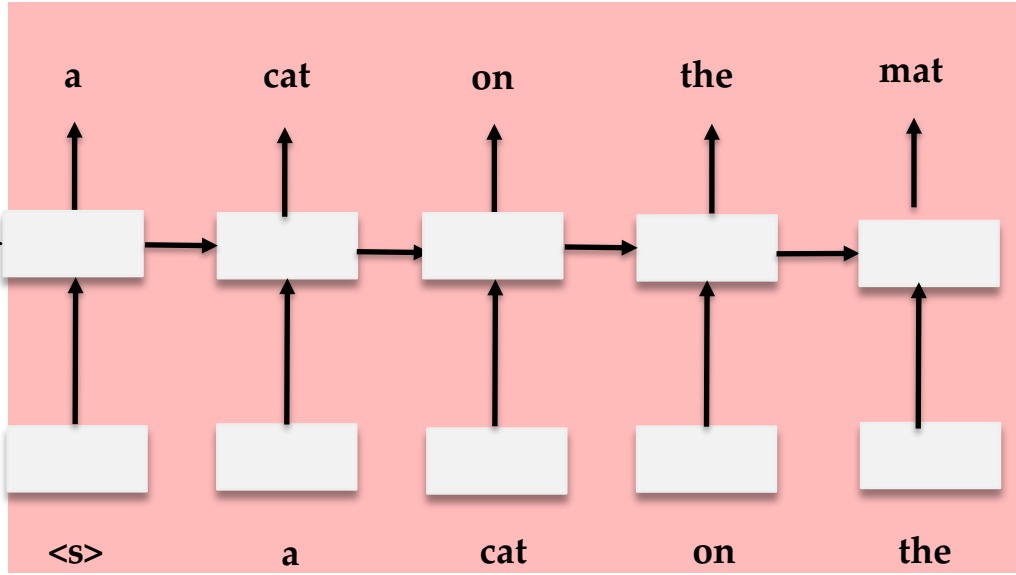
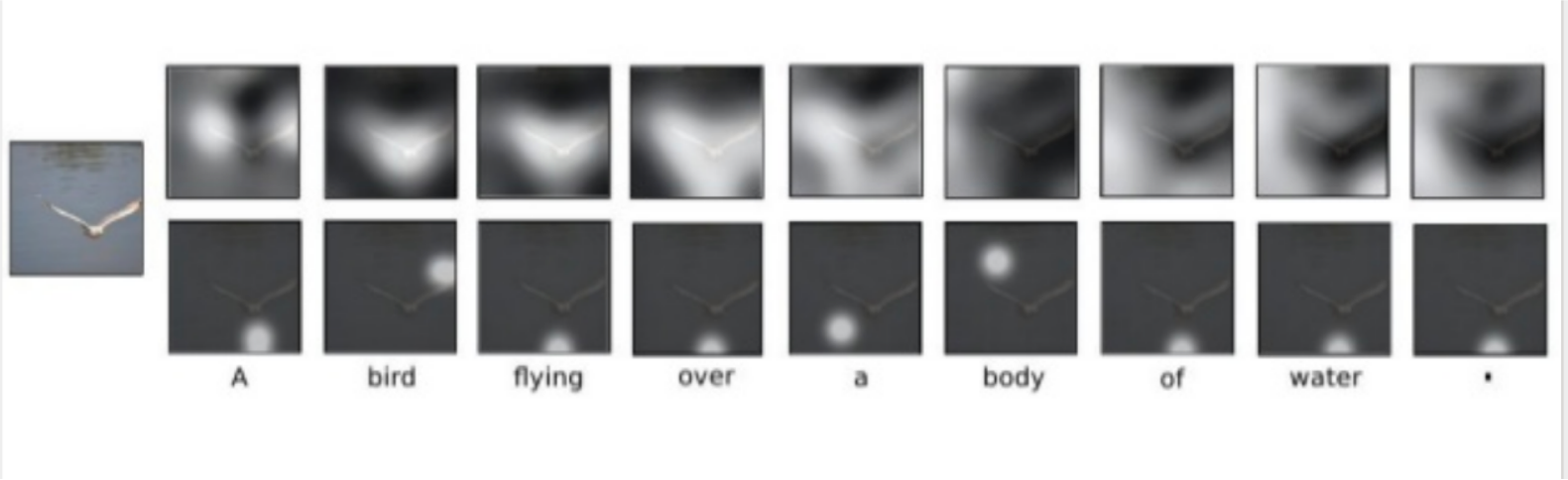


Image Captioning II

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

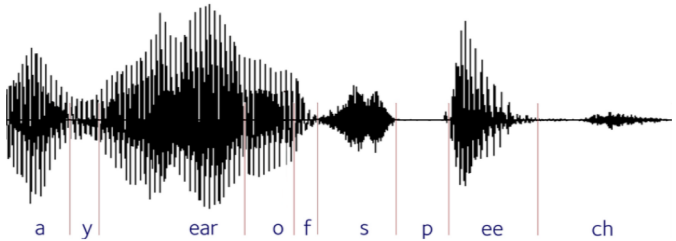
Kelvin Xu KELVIN.XU@UMONTREAL.CA
Jimmy Lei Ba JIMMY@PSI.UTORONTO.CA
Ryan Kiros RKIROS@CS.TORONTO.EDU
Kyunghyun Cho KYUNGHYUN.CHO@UMONTREAL.CA
Aaron Courville AARON.COURVILLE@UMONTREAL.CA
Ruslan Salakhutdinov RSALAKHU@CS.TORONTO.EDU
Richard S. Zemel ZEMEL@CS.TORONTO.EDU
Yoshua Bengio FIND-ME@THE.WEB

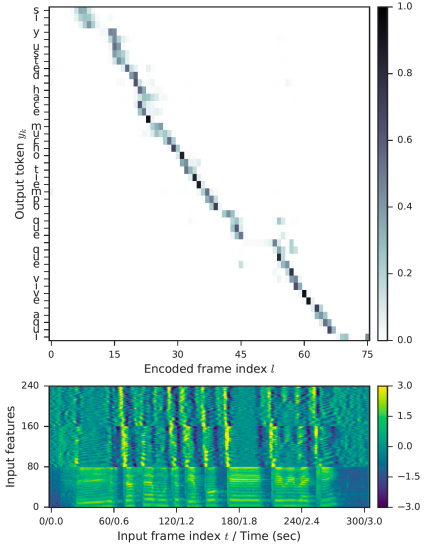


Other Computer Vision Tasks with Attention

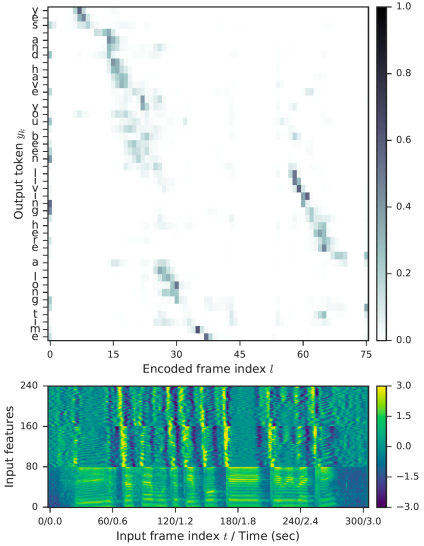
- **Visual Question Answering**: given an image and a natural language question about the image, the task is to provide an accurate natural language answer.
- **Video Caption Generation:** attempts to generate a complete and natural sentence, enriching the single label as in video classification, to capture the most informative dynamics in videos.

Speech recognition / translation





(a) *Spanish speech recognition decoder attention.*



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

Nice Links

ATTENTION

Let’s play