COMP4650/6490 Document Analysis

Sequence-to-Sequence & Attention

ANU School of Computing

Administrative matters

- Assignment 2
  - Due: 5pm Tuesday 19 September
  - Extension application:
    24 hours before due date + supporting documents
- Assignment 3
  - Will be released later this week
- Labs
  - No lab in this week
  - Lab3 solution (practical) will be released once all A2 late submissions are collected

Contents

- Motivation for attention
- Sequence-to-sequence models
- Attention in sequence-to-sequence models

Recap

- Sequential Structure - Recurrent Neural Network

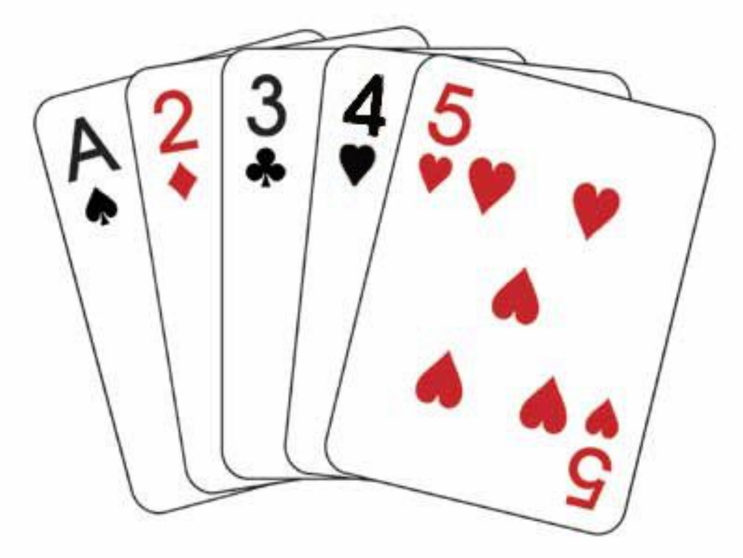
Unknown implicit structure

Unknown structure

- We infer the structure by attention

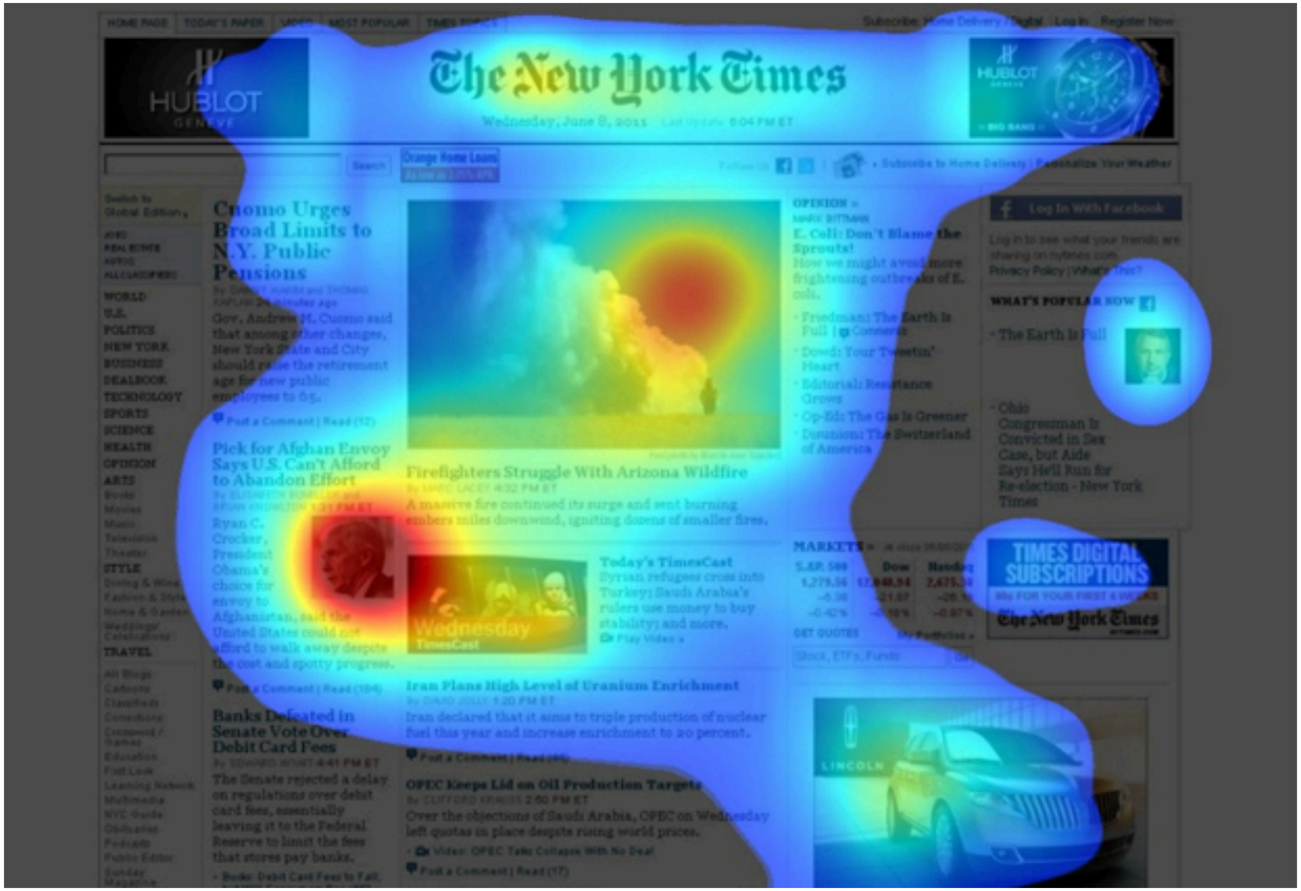
Attention: A neural network layer that learns to select out relevant parts of the input.

Motivation for attention



What is the sum of all **black** cards?

Motivation for attention

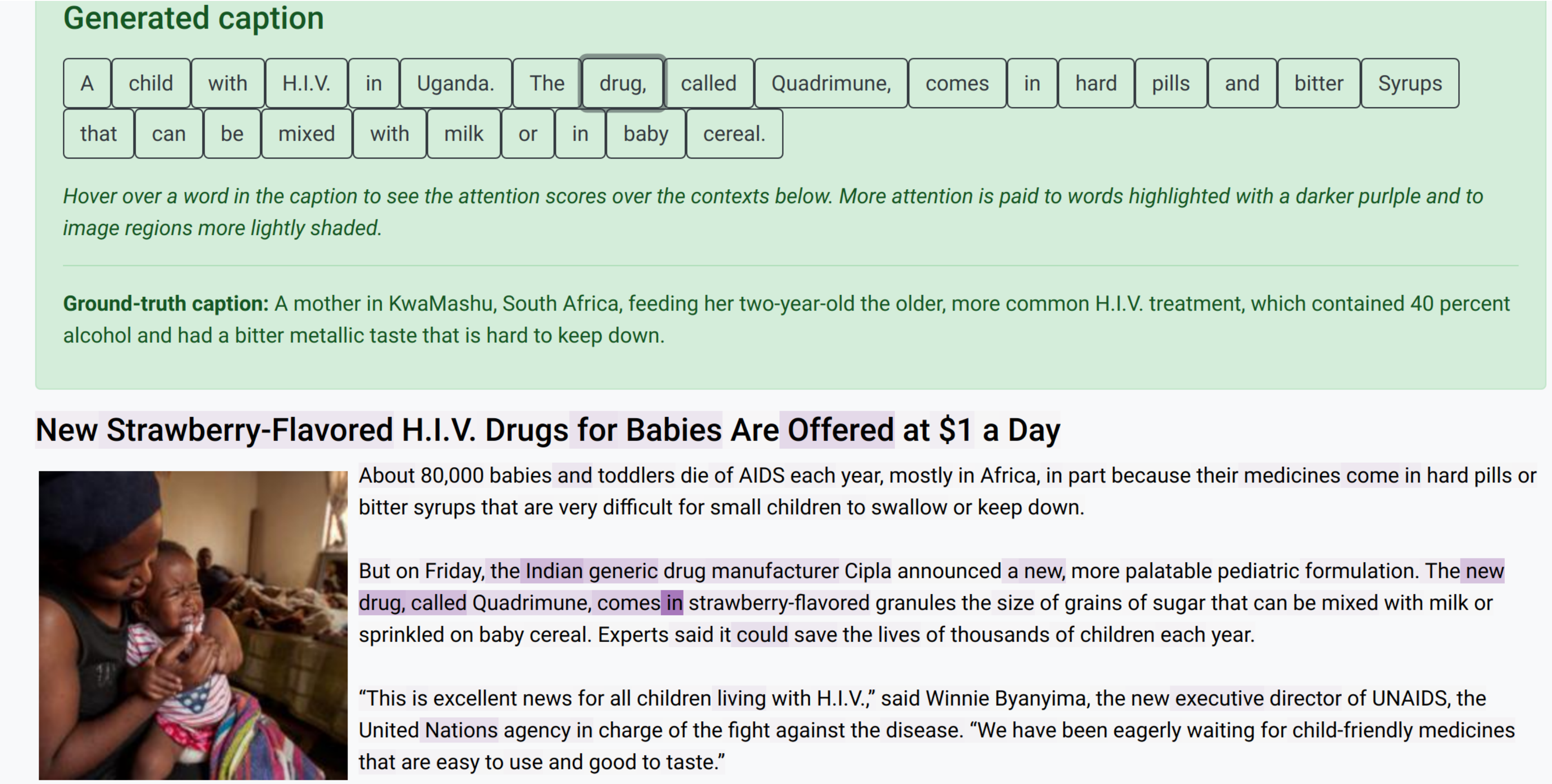

The New York Times
Wednesday, June 8, 2011
HUBLOT
Cuomo Urges Broad Limits to N.Y. Public Pensions
Pick for Afghan Envoy Says U.S. Can't Afford to Abandon Effort
Banks Defeated in Senate Vote Over Debit Card Fees
Firefighters Struggle With Arizona Wildfire
Today's TimesCast
Wednesday
Iran Plans High Level of Uranium Enrichment
OPEC Keeps Lid on Oil Production Targets
OPINION »
E. Coli: Don't Blame the Sprouts!
MARKETS »
Log In With Facebook
WHAT'S POPULAR NOW
TIMES DIGITAL SUBSCRIPTIONS
LINCOLN

Motivation for attention

Example from the Stanford Question Answering Dataset (SQuAD 2.0)

Question: What city in Victoria is called the sporting capital of Australia?

Victoria's total gross state product (GSP) is ranked second in Australia, although Victoria is ranked fourth in terms of GSP per capita because of its limited mining activity. Culturally, Melbourne is home to a number of museums, art galleries and theatres and is also described as the "sporting capital of Australia".



Generated caption

A child with H.I.V. in Uganda. The drug, called Quadrimune, comes in hard pills and bitter Syrups that can be mixed with milk or in baby cereal.

Hover over a word in the caption to see the attention scores over the contexts below. More attention is paid to words highlighted with a darker purlple and to image regions more lightly shaded.

Ground-truth caption: A mother in KwaMashu, South Africa, feeding her two-year-old the older, more common H.I.V. treatment, which contained 40 percent alcohol and had a bitter metallic taste that is hard to keep down.

New Strawberry-Flavored H.I.V. Drugs for Babies Are Offered at $1 a Day

About 80,000 babies and toddlers die of AIDS each year, mostly in Africa, in part because their medicines come in hard pills or bitter syrups that are very difficult for small children to swallow or keep down.

But on Friday, the Indian generic drug manufacturer Cipla announced a new, more palatable pediatric formulation. The new drug, called Quadrimune, comes in strawberry-flavored granules the size of grains of sugar that can be mixed with milk or sprinkled on baby cereal. Experts said it could save the lives of thousands of children each year.

“This is excellent news for all children living with H.I.V.,” said Winnie Byanyima, the new executive director of UNAIDS, the United Nations agency in charge of the fight against the disease. “We have been eagerly waiting for child-friendly medicines that are easy to use and good to taste.”

Tran, Alasdair, Alexander Mathews, and Lexing Xie. "Transform and tell: Entity-aware news image captioning." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.

Contents

- Motivation for attention
- Sequence-to-sequence models
- Attention in sequence-to-sequence models

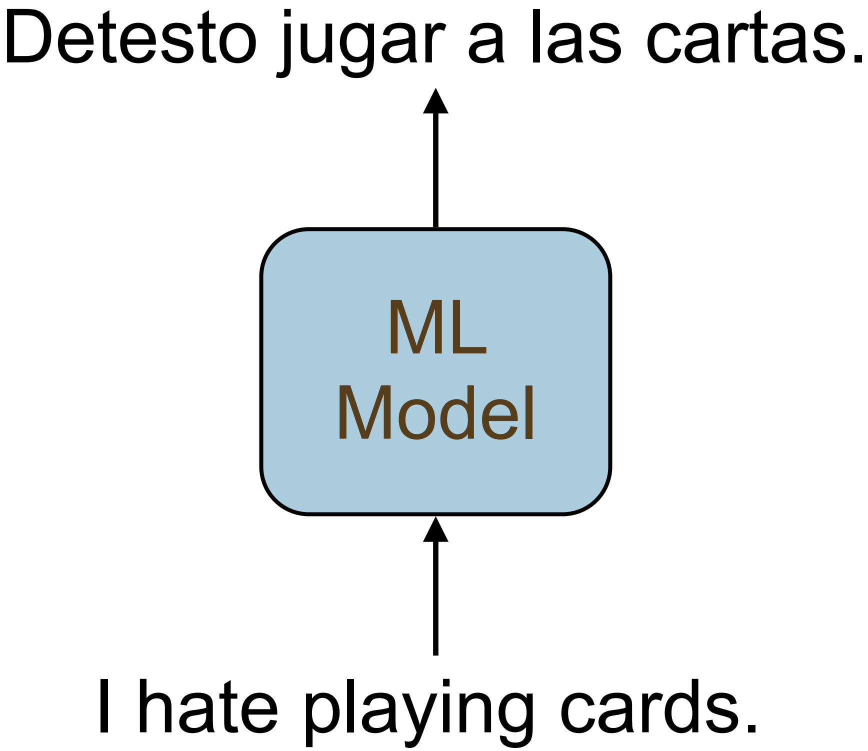
Origin of RNN attention

- The attention mechanism was first proposed in natural language translation models.
- These models receive as input a sentence in one language and output the corresponding sentence in the target language.
  - For example: English to Spanish

Translation

- We’ve seen how RNNs can process sequence inputs.
- For translation, the output is also a sequence (and not necessarily the same length as the input!)
- How would you design a model that can output different length sequences?

Sequence to sequence

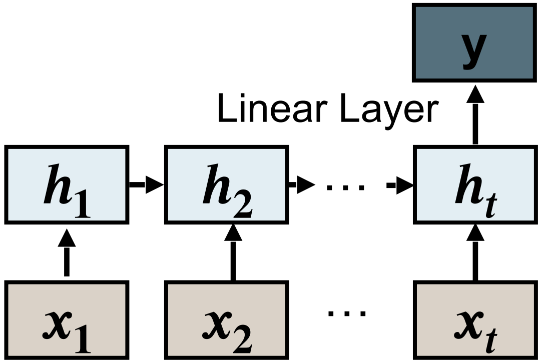


Detesto jugar a las cartas.

ML Model

I hate playing cards.

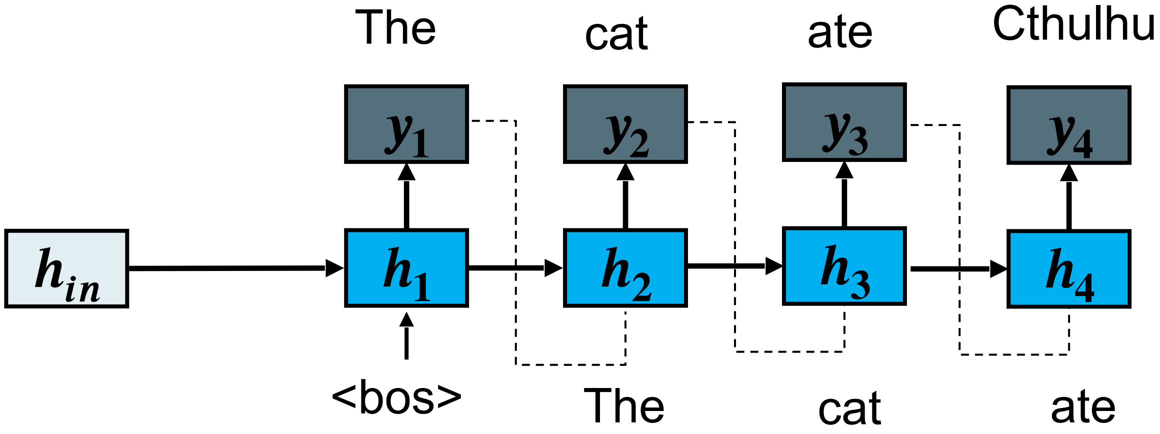
First part: RNN on the input sequence



We don’t have to use the outputs at each step we can ignore them and use the last one.

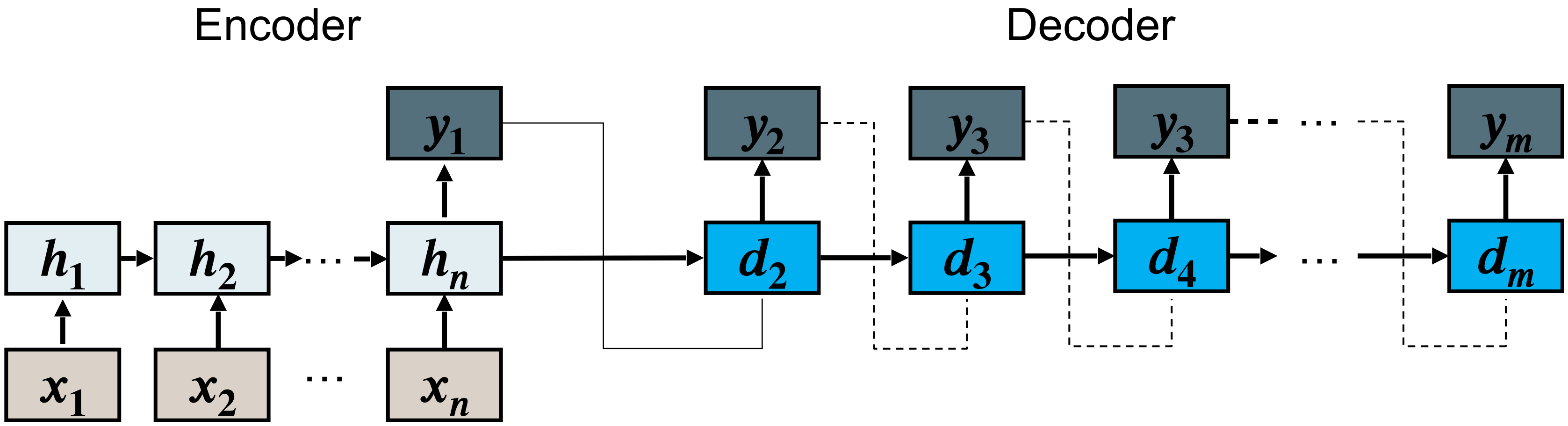
Second part: RNN for language modelling

The RNN takes the word generated in the last step as the input to the next step.

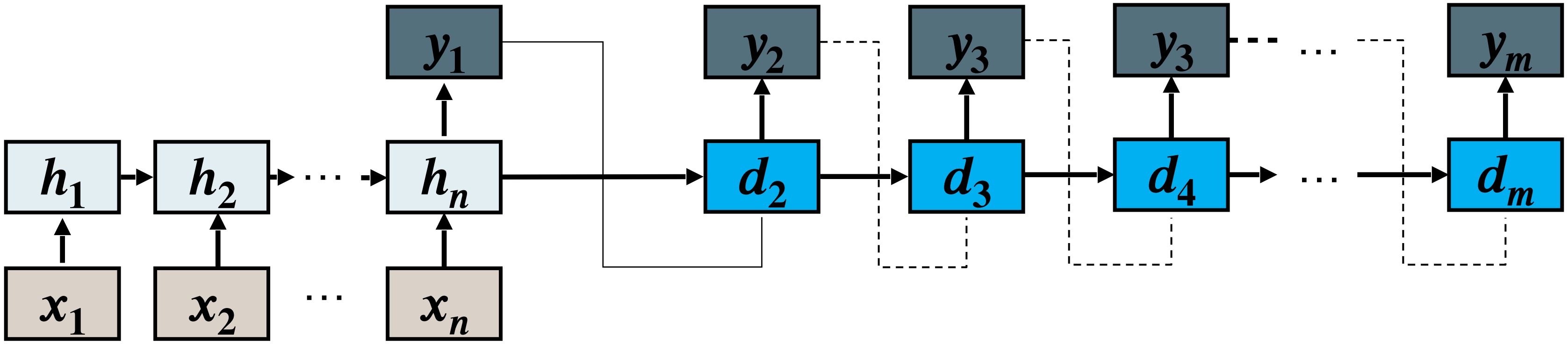


Encoder-Decoder (Seq2Seq)

- We use 2 RNN models
- The first (encoder) maps the input sequence to a vector representation
- The second (decoder) maps all previously generated tokens to the next token
- The output of the encoder is used to initialise the decoder’s history



Encoder-Decoder (Seq2Seq)

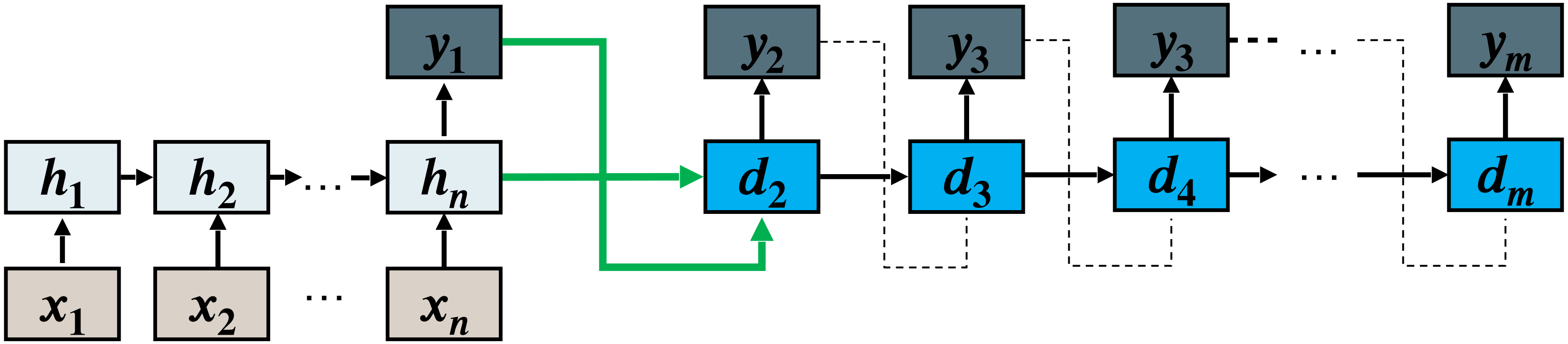


$h_t = \text{Encoder}(h_{t-1}, x_t)$

$d_t = \text{Decoder}(d_{t-1}, y_{t-1})$

Problems with Encoder-Decoders

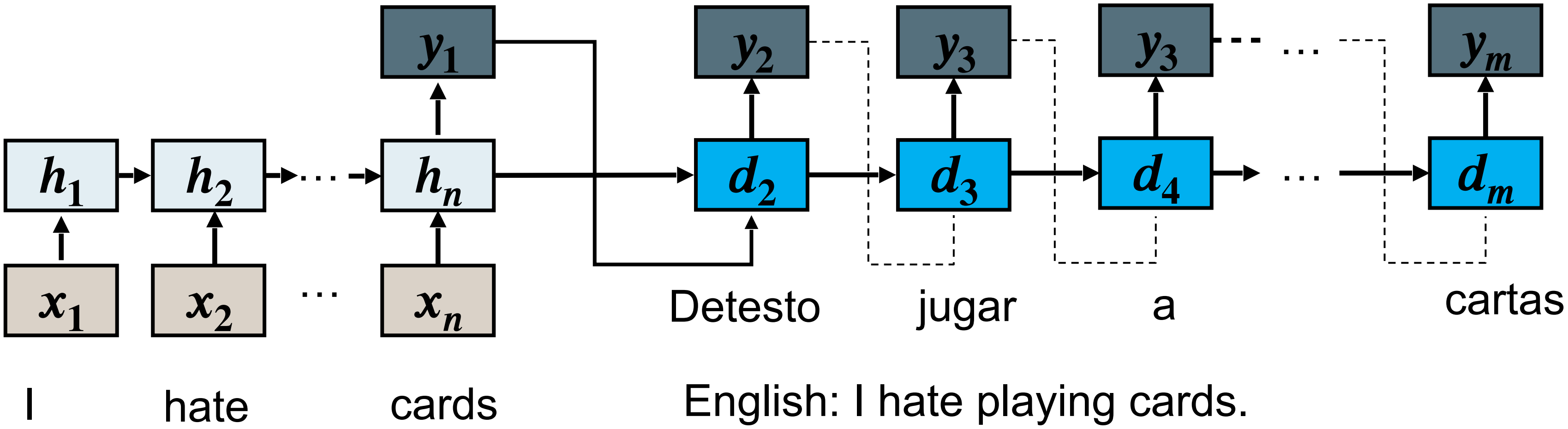
- All of the information in the input is compressed into *one or two vectors* (i.e. bottleneck)
- The decoder must extract information relevant for producing the next output
  - Often, only some parts of the input are relevant



Problems with Encoder-Decoders

There can be many steps between an input term and a corresponding output term, so:

- Forgetting is a problem.
- Training is hard (e.g. vanishing or exploding gradients)

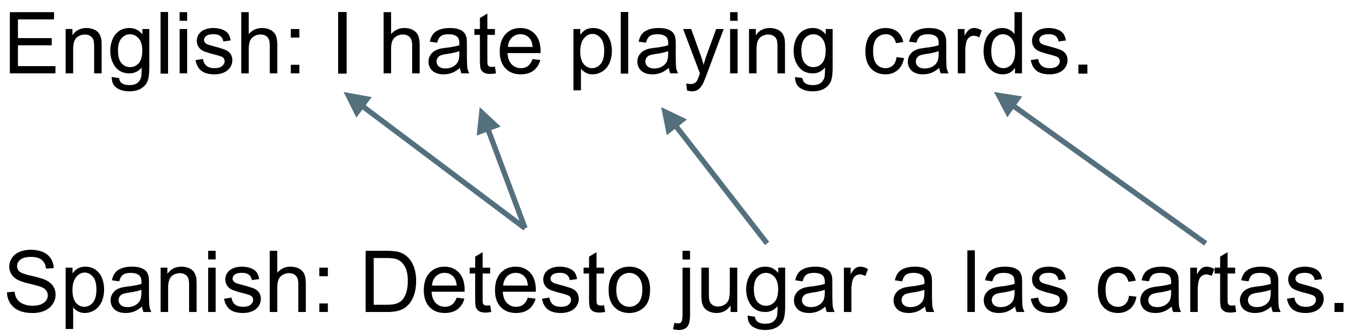


English: I hate playing cards.
Spanish: Detesto jugar a las cartas.

Contents

- Motivation for attention
- Sequence-to-sequence models
- Attention in sequence-to-sequence models

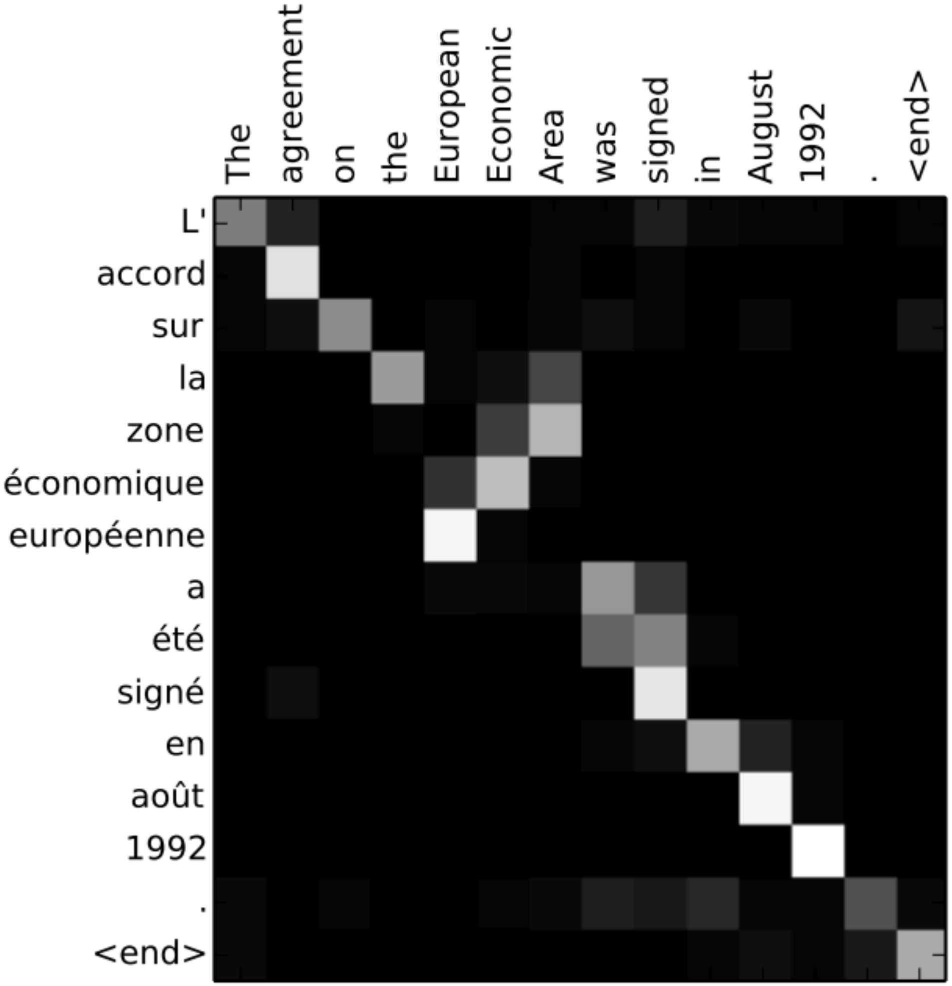
Origin of attention



English: I hate playing cards.

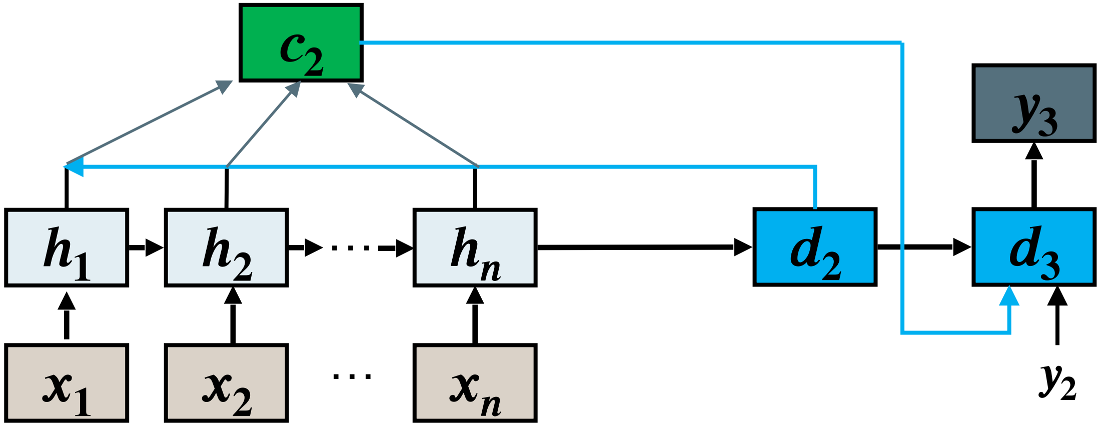
Spanish: Detesto jugar a las cartas.

Neural machine translation



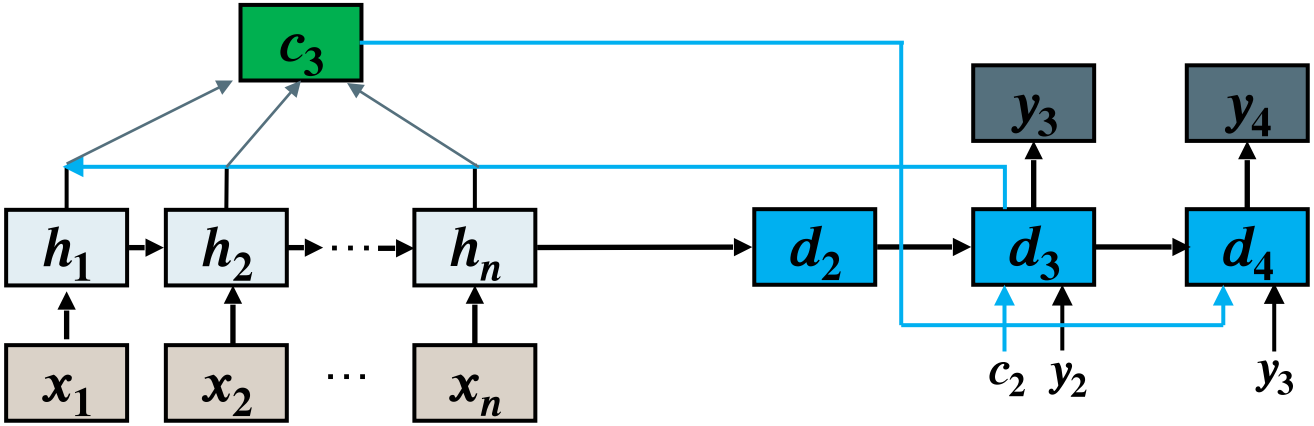
Attention – Key idea

Instead of forcing the decoder to extract the information it needs from h_n, we provide a **context vector** c_j from the encoder computed specifically for the current step.



Attention – Key idea

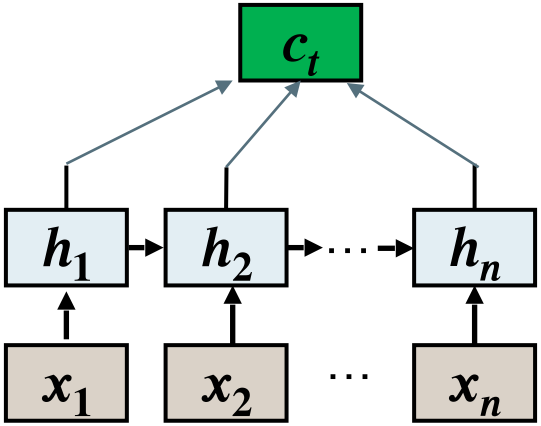
We have a new context vector for every step of the decoding process.



Attention as weighted averaging

The **context vector** is a weighted average of all the encoder outputs.

$$\boldsymbol{c}_j = \sum_{i=1}^{n} a_{j,i}\,\boldsymbol{h}_i$$



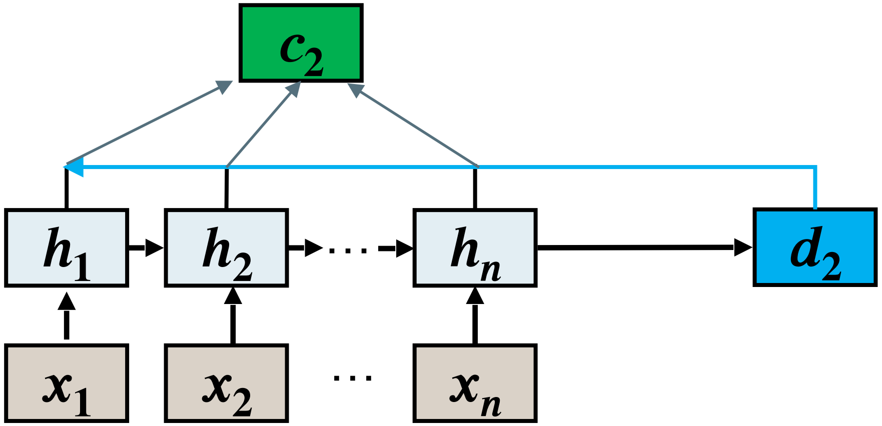
The **attention weights** $a_{j,i}$ should be non-negative and sum to 1.

Attention weights

Measure how important each encoding vector is for the current step of the decoder.

We use a learnable score function.

$$\tilde{a}_{j,i} = \text{score}(\boldsymbol{d}_j, \boldsymbol{h}_i)$$



Attention score function

There are several possible score functions, some of which have learnable parameters.

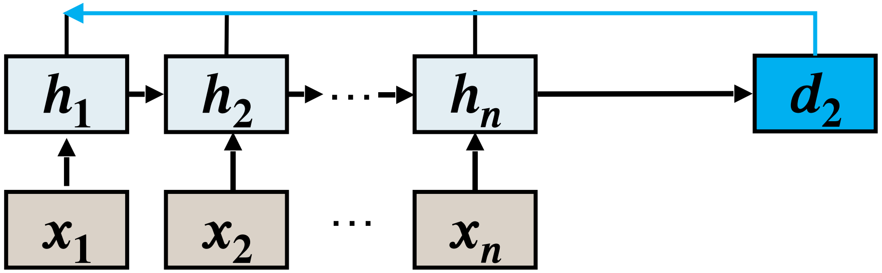
Examples below (where $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^{d\times 1}$):

$$\text{score}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^\top \boldsymbol{y}$$

$$\text{score}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}^\top \boldsymbol{y}}{\sqrt{d}}$$

$$\text{score}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^\top W \boldsymbol{y}$$

$$\text{score}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{v}^\top \tanh(W\boldsymbol{x} + U\boldsymbol{y})$$

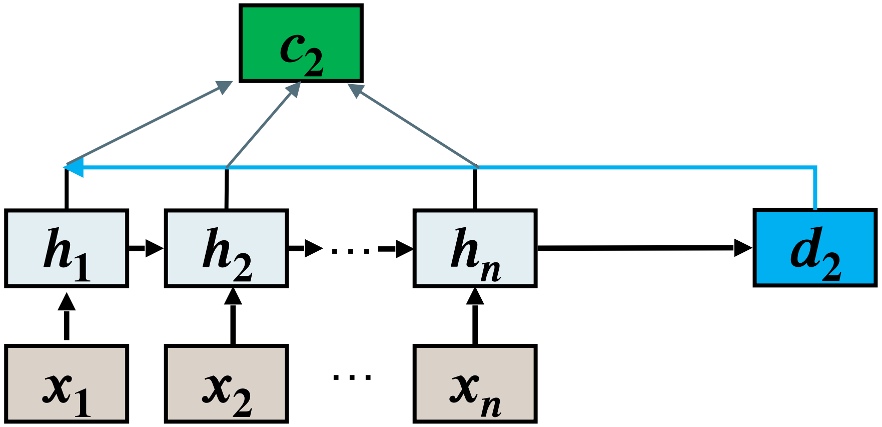


$$\widetilde{a}_{2,i} = \text{score}(\boldsymbol{d}_2, \boldsymbol{h}_i),\ i = 1,\ldots,n$$

$$\boldsymbol{a}_2 = \text{softmax}(\widetilde{\boldsymbol{a}}_2)$$

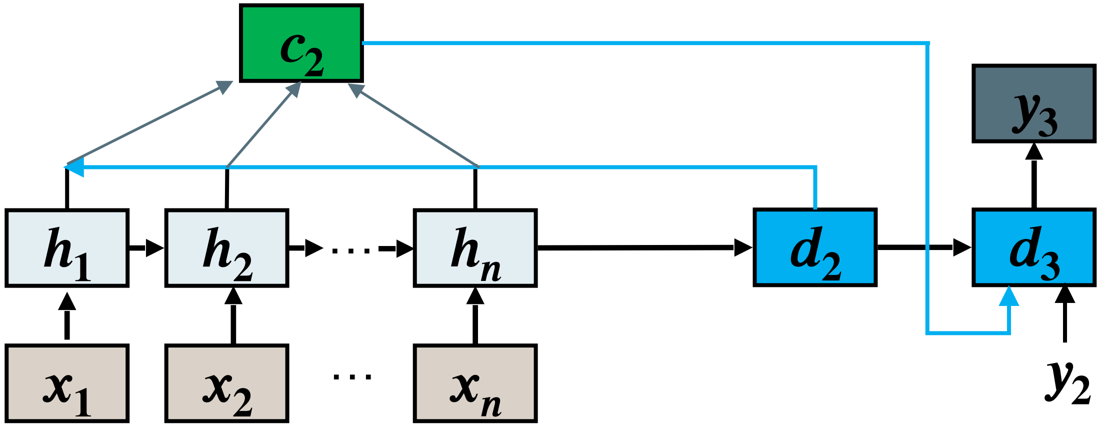
Compare the current state of the decoder $\boldsymbol{d}_j$ with all encoder outputs $\boldsymbol{h}_i$ using a score function.

Calculate the **attention vector** using softmax.



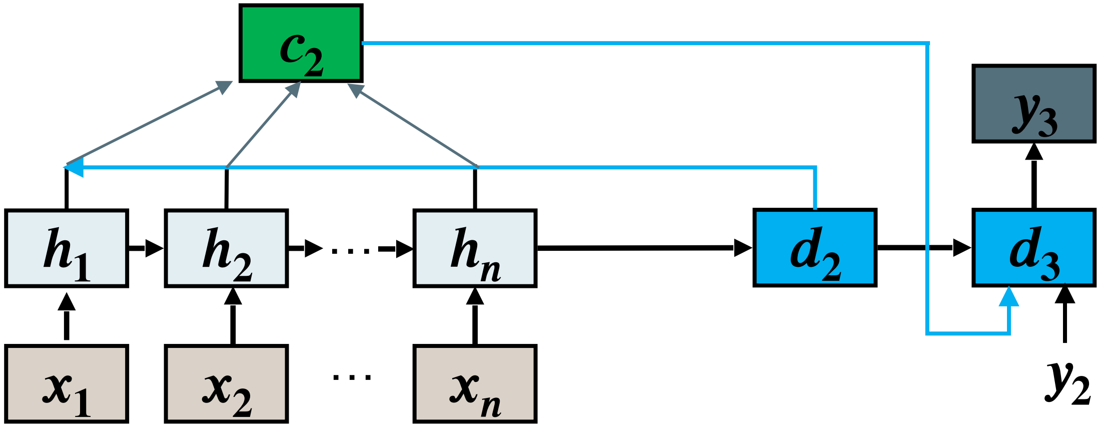
$$c_2 = \sum_{i=1}^{n} a_{2,i} h_i$$

Compute the **context vector** c_j using the attention weights $a_{j,i}$ and the outputs of the encoder h_i.



$$d_3 = \text{Decoder}(d_2,\ [c_2, y_2])$$

Feed the context vector c_j and last token embedding y_j into the next step of the decoder.



$$\tilde{a}_{j,i} = \text{score}(\boldsymbol{d}_j, \boldsymbol{h}_i),\ i = 1, \ldots, n$$

$$\boldsymbol{a}_j = \text{softmax}(\tilde{\boldsymbol{a}}_j)$$

$$\boldsymbol{c}_j = \sum_{i=1}^{n} a_{j,i}\, \boldsymbol{h}_i$$

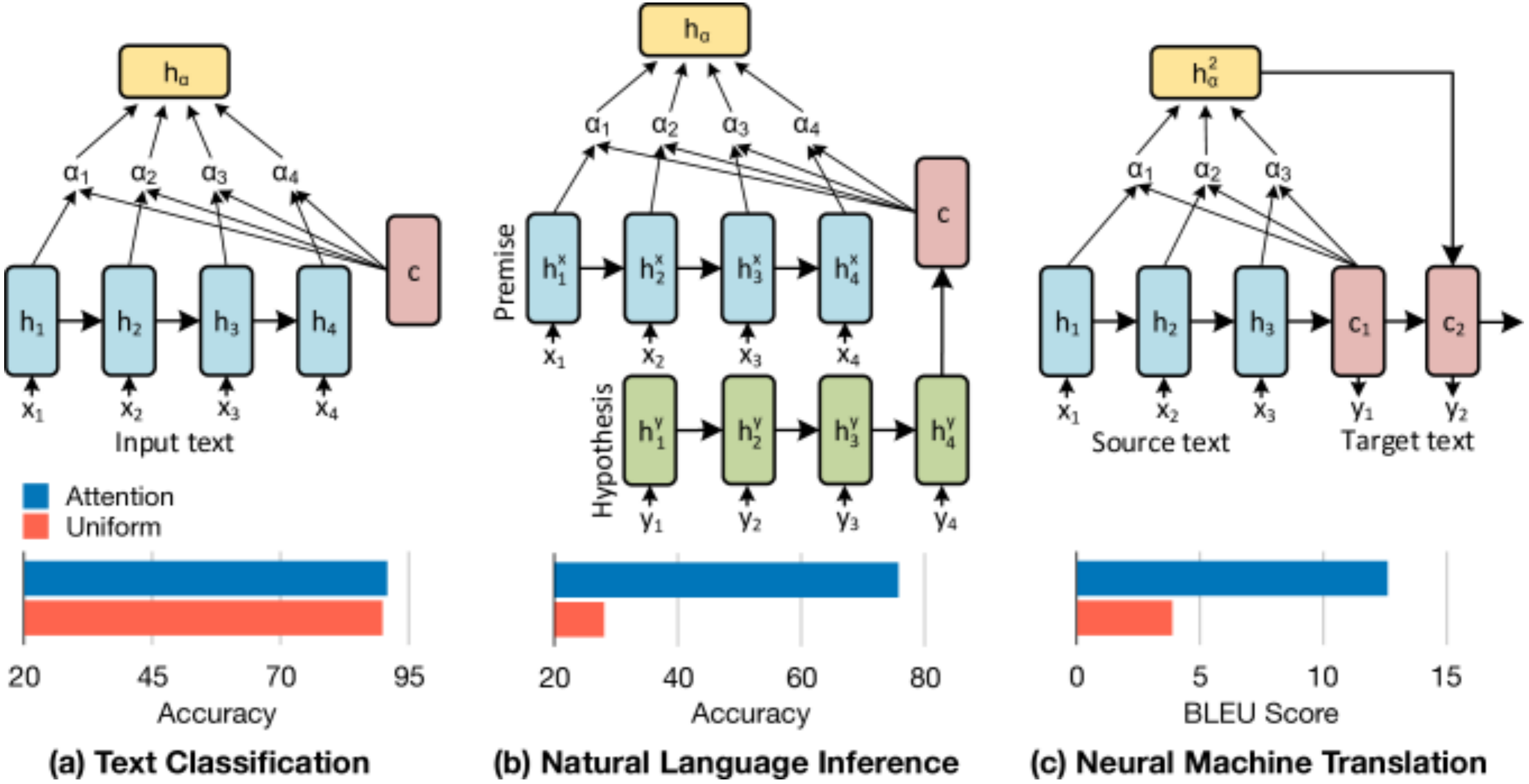
$$\boldsymbol{d}_{j+1} = \text{Decoder}(\boldsymbol{d}_j,\ [\boldsymbol{c}_j, \boldsymbol{y}_j])$$

Simple attention variations

There are many slight modifications on this attention scheme.

- Where the context vector is included in the decoder (input to the RNN Cell or used at the output step)
- Guided attention, only apply attention to some of the encoder.
- Self-attention on the decoder side

Attention for NLP tasks



(a) Text Classification (b) Natural Language Inference (c) Neural Machine Translation

Key-Value attention mechanism

$$\text{Attention}(\boldsymbol{q}, K, V) = \text{softmax}(K\boldsymbol{q})^\top V$$

- $\boldsymbol{q} \in \mathbb{R}^{d \times 1}$ is a query vector
- $K \in \mathbb{R}^{n \times d}$ are key vectors
- $V \in \mathbb{R}^{n \times d_v}$ are value vectors
- More in the Transformers lecture

Summary

- Encoder-Decoders are needed to generate output sequences of varying length from variable length inputs
- The Attention mechanism allows the decoder to focus on the parts of the input relevant to the next decoding step.

Reference

- Sections 9.7, 9.8, Speech and Language Processing