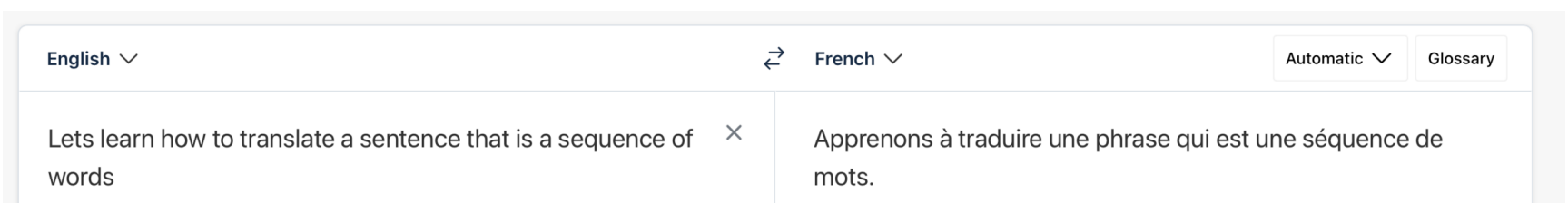
Applied Machine Learning

Neural Networks for Sequences

Oumar Kaba

Deep Neural Networks

- Neural Networks for **Tabular Data**
  - MLP
- Neural Networks for **Images**
  - CNN
- Neural Networks for **Sequences**
  - input is a sequence, the output is a sequence, or both are sequences
    - *e.g. machine translation, speech recognition, text classification, image captioning*



Learning objectives

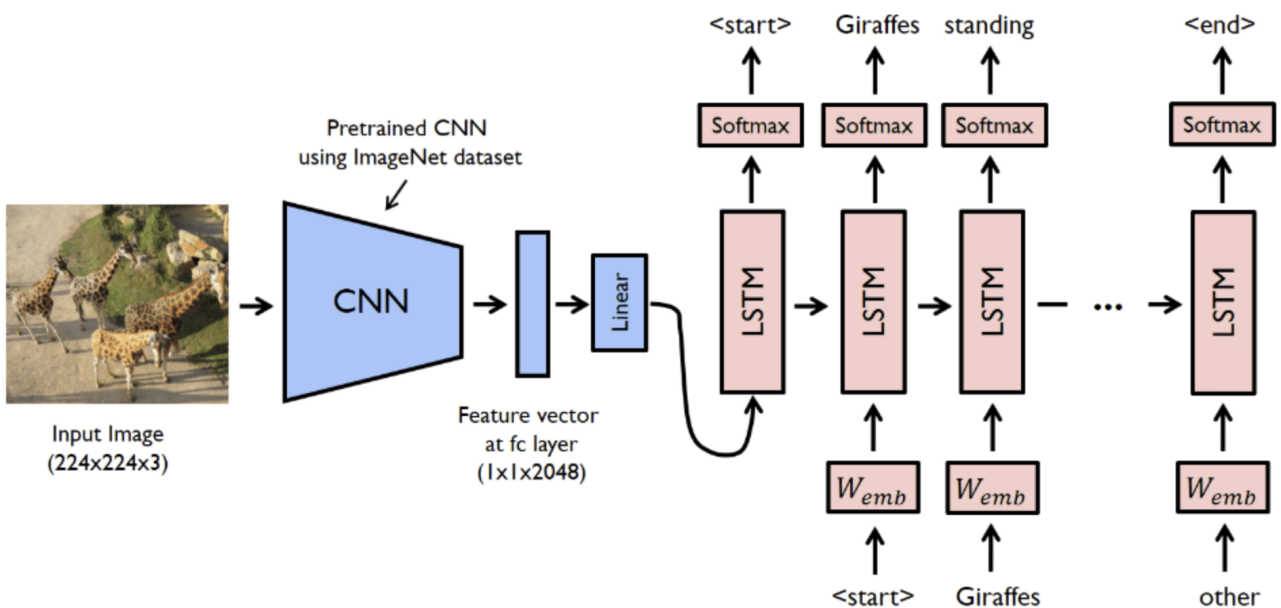
- Recurrent neural networks (RNNs)
  - 3 different models for different input/output
  - training with back propagation through time
- understand the attention mechanisms
- The architecture of transformer

Recurrent neural networks (RNNs)

- Vec2Seq (sequence generation)
- Seq2Vec (sequence classification)
- Seq2Seq (sequence translation)

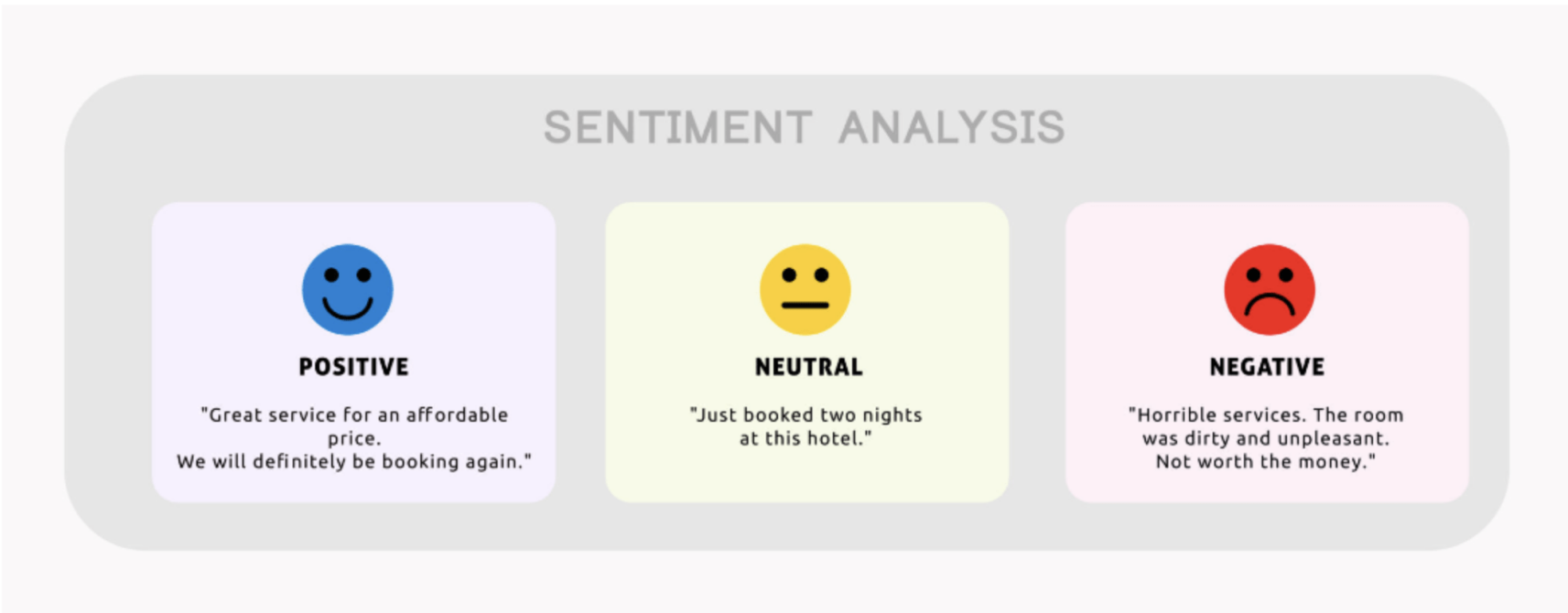
Vec2Seq (sequence generation)

Example : Caption generation



Seq2Vec (sequence classification)

Example : Sentiment classification



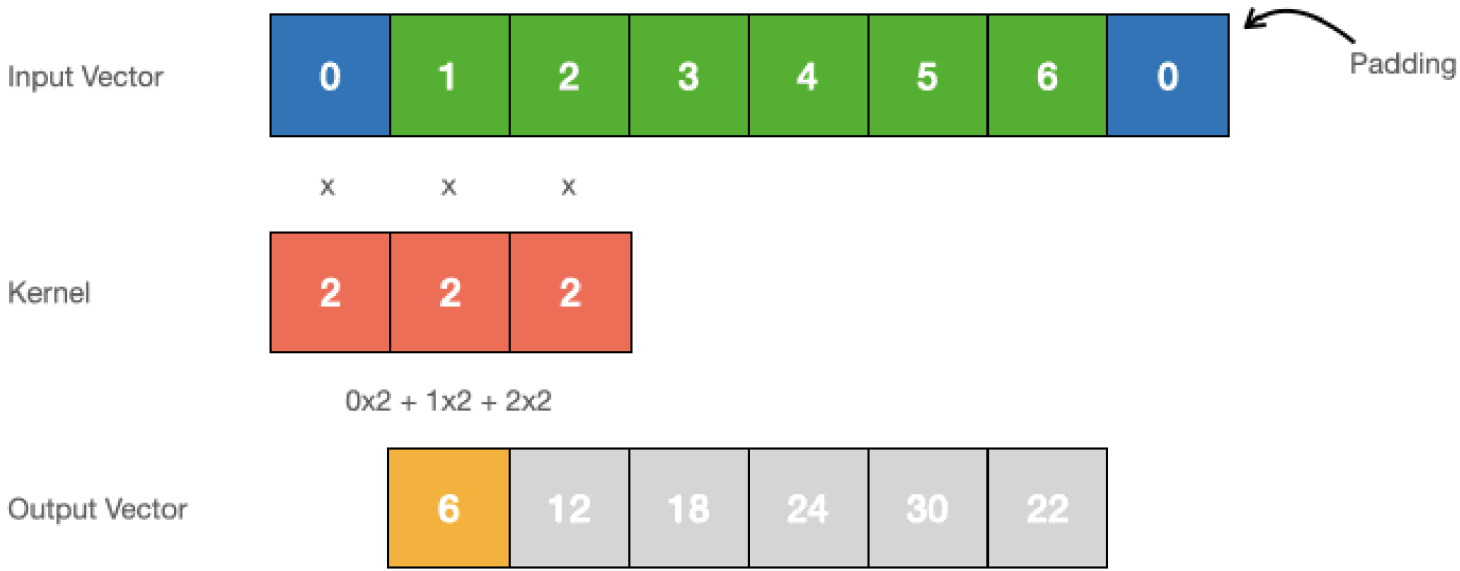
Seq2Seq (sequence translation)

Examples :

- Language translation (e.g. English to French)
- Language modelling (beginning of text to complete text)
- Protein folding (Sequence of amino acids to sequence of angles)
- Time series (e.g. stock price prediction, weather prediction)

Limitation of convolutions

We saw that convolutions can be used to process sequences



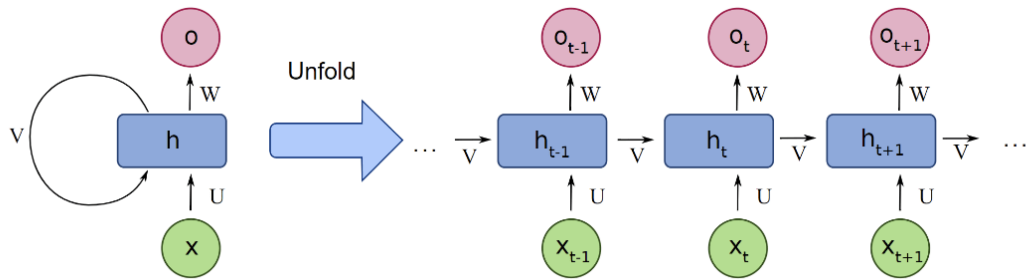
Cannot generate sequences of variable length

Generating sequences of variable length

The method we will look at works in the following way:

- Have a state that keeps track of past sequence information
- Have a special token <EOS> that indicates the end of sequence
- We can also have a <SOS> token that indicates the start of sequence

Recurrent neural network (RNN)



$$x_t \in \mathbb{R}^d, h_t \in \mathbb{R}^m, o_t \in \mathbb{R}^n$$

Matrix shapes: $U : m \times d, V : m \times m, W : n \times m$

RNN layer : $h_t = f\left(V \cdot h_{t-1} + U \cdot x_t + b_h\right)$

$$o_t = g\left(W \cdot h + b_o\right)$$

f is an activation function (tanh or ReLU)

Recurrent neural network (RNN)

$$h_t = \text{ReLU}\left(V \cdot h_{t-1} + U \cdot x_t + b_h\right)$$
$$o_t = \left(W \cdot h_t + b_o\right), \qquad x_t \in \mathbb{R}^2, h_t \in \mathbb{R}^3, o_t \in \mathbb{R}^1$$

$$U = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 0 \end{bmatrix}, V = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, W = \begin{bmatrix} 1 & 0 & -1 \end{bmatrix}, b_h = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, b_o = 0$$

$$h_{t-1} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, x_t = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

What is h_t ?

Recurrent neural network (RNN)

$$h_t = \text{ReLU}\left(V \cdot h_{t-1} + U \cdot x_t + b_h\right)$$
$$o_t = \left(W \cdot h_t + b_o\right), \qquad x_t \in \mathbb{R}^2, h_t \in \mathbb{R}^3, o_t \in \mathbb{R}^1$$

$$U = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 0 \end{bmatrix}, V = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, W = \begin{bmatrix} 1 & 0 & -1 \end{bmatrix}, b_h = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, b_o = 0$$

$$h_{t-1} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, x_t = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \text{What is } h_t \text{?} \qquad h_t = \text{ReLU}\left(V \cdot h_{t-1} + U \cdot x_t\right)$$

Recurrent neural network (RNN)

$$h_t = \text{ReLU}\left(V \cdot h_{t-1} + U \cdot x_t + b_h\right)$$
$$o_t = \left(W \cdot h_t + b_o\right), \qquad x_t \in \mathbb{R}^2, h_t \in \mathbb{R}^3, o_t \in \mathbb{R}^1$$

$$U = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 0 \end{bmatrix}, V = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, W = \begin{bmatrix} 1 & 0 & -1 \end{bmatrix}, b_h = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, b_o = 0$$

$$h_{t-1} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, x_t = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \text{What is } h_t\text{?} \qquad h_t = \text{ReLU}\left(\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + U \cdot x_t\right)$$

Recurrent neural network (RNN)

$$h_t = \text{ReLU}\left(V \cdot h_{t-1} + U \cdot x_t + b_h\right)$$
$$o_t = \left(W \cdot h_t + b_o\right), \qquad x_t \in \mathbb{R}^2, h_t \in \mathbb{R}^3, o_t \in \mathbb{R}^1$$

$$U = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 0 \end{bmatrix}, V = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, W = \begin{bmatrix} 1 & 0 & -1 \end{bmatrix}, b_h = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, b_o = 0$$

$$h_{t-1} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, x_t = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \text{What is } h_t\text{?} \qquad h_t = \text{ReLU}\left(\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}\right)$$

Recurrent neural network (RNN)

$$h_t = \text{ReLU}\left(V \cdot h_{t-1} + U \cdot x_t + b_h\right)$$
$$o_t = \left(W \cdot h_t + b_o\right), \qquad x_t \in \mathbb{R}^2, h_t \in \mathbb{R}^3, o_t \in \mathbb{R}^1$$

$$U = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 0 \end{bmatrix}, V = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, W = \begin{bmatrix} 1 & 0 & -1 \end{bmatrix}, b_h = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, b_o = 0$$

$$h_{t-1} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, x_t = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \text{What is } o_t \text{?} \qquad o_t = W \cdot h_t = \begin{bmatrix} 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = 1$$

Recurrent neural networks (RNNs)

- Vec2Seq (sequence generation)
  - output, $y_{1:T}$ is generated one token at a time
  - at each step we sample y_t from the hidden state h_t and then feed it back to the model to get h_{t+1}

arbitrary-length sequence of vectors

$$f_\theta : \mathbb{R}^D \rightarrow \mathbb{R}^{N_\infty C}$$

D: input vector size

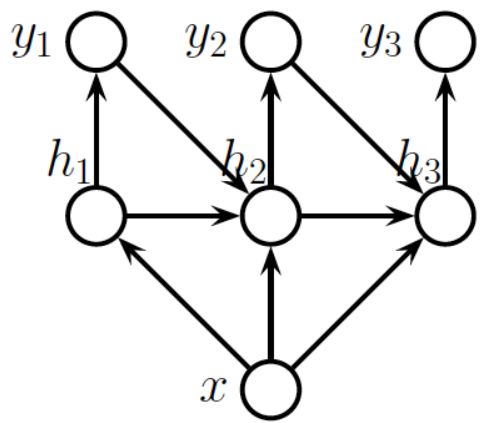
N_∞: *arbitrary-length sequence of vectors of length C*

C: each output vector size

conditional generative model:

$$p(y_{1:T}|x) = \sum_{h_{1:T}} p(y_{1:T}, h_{1:T}|x) = \sum_{h_{1:T}} \prod_{t=1}^{T} p(y_t|h_t)p(h_t|h_{t-1}, y_{t-1}, x)$$

with the initial hidden state$p(h_1|h0, y0, x) = p(h_1|x)$



Recurrent neural networks (RNNs)

- Vec2Seq (sequence generation) $f_\theta : \mathbb{R}^D \rightarrow \mathbb{R}^{TC}$

conditional generative model:

$$p(y_{1:T}|x) = \sum_{h_{1:T}} p(y_{1:T}, h_{1:T}|x) = \sum_{h_{1:T}} \prod_{t=1}^{T} p(y_t|h_t) p(h_t|h_{t-1}, y_{t-1}, x)$$

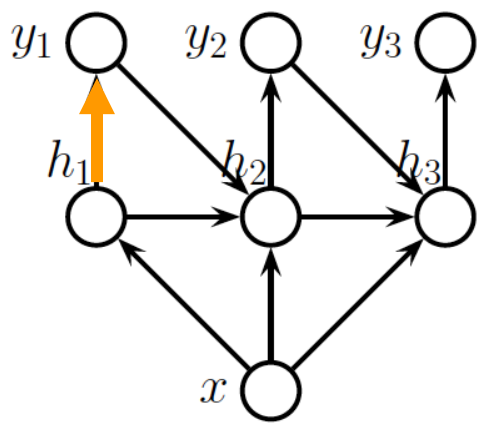
hidden-to-output weights

- real-valued output: $\hat{y}_t = W_{hy} h_t$

$$p(y_t|h_t) = \mathcal{N}(y_t|\hat{y}_t, \mathbf{I})$$

- categorical output: $\hat{y}_t = \mathrm{softmax}(W_{hy} h_t)$

$$p(y_t|h_t) = \mathrm{Categorical}(y_t|\hat{y}_t)$$



** per usual dropping bias terms for simplicity

Recurrent neural networks (RNNs)

- Vec2Seq (sequence generation)

$$f_\theta : \mathbb{R}^D \rightarrow \mathbb{R}^{TC}$$

conditional generative model:

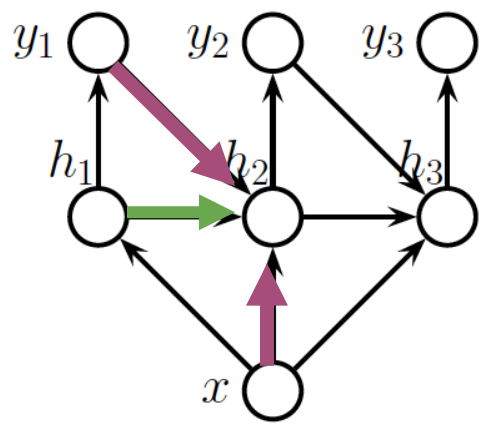
$$p(y_{1:T}|x) = \sum_{h_{1:T}} p(y_{1:T}, h_{1:T}|x) = \sum_{h_{1:T}} \prod_{t=1}^{T} p(y_t|h_t) p(h_t|h_{t-1}, y_{t-1}, x)$$

hidden state:

$$p(h_t|h_{t-1}, y_{t-1}, x) = \mathbb{I}(h_t = f(h_{t-1}, y_{t-1}, x))$$

input-to-hidden weights hidden-to-hidden weights

$$h_t = \varphi(W_{xh}[x; y_{t-1}] + W_{hh} h_{t-1})$$



Recurrent neural networks (RNNs)

- Vec2Seq (sequence generation)

$$f_\theta : \mathbb{R}^D \rightarrow \mathbb{R}^{TC}$$

model

hidden-to-output weights

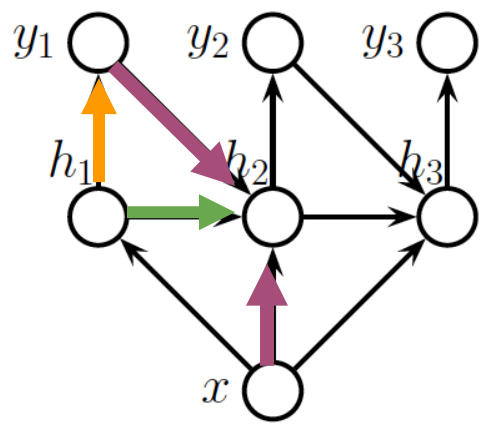
$$\hat{y}_t = g(W_{hy} h_t)$$

input-to-hidden weights hidden-to-hidden weights

$$h_t = \varphi(W_{xh}[x; y_{t-1}] + W_{hh} h_{t-1})$$

RNNs are powerful

- In theory can have unbounded memory and are as powerful as a Turing machine
- In practice, memory size is determined by the size of the latent space and strength of the parameters



Recurrent neural networks (RNNs)

- Vec2Seq (sequence generation)

conditional generative model:

$$p(y_{1:T}|x) = \sum_{h_{1:T}} p(y_{1:T}, h_{1:T}|x) = \sum_{h_{1:T}} \prod_{t=1}^{T} p(y_t|h_t) p(h_t|h_{t-1}, y_{t-1}, x)$$

language modelling: generating sequences unconditionally (by setting $x = \emptyset$) which is learning joint probability distributions over sequences of discrete tokens, i.e., $p(y_1, \ldots, y_T)$

Example:

character level RNN trained on the book The Time Machine by H. G. Wells (32,000 words and 170k character)

Output when given prefix "the":

in his hand was a glitteringmetallic framework scarcely larger than a small clock and verydelicately made there was ivory in it and the latter than s bettyre tat howhong s ie time thave ler simk you a dimensions le ghat dionthat shall travel indifferently in any direction of space and timeas the driver determinesfilby contented himself with laughterbut i have experimental verification said the time travellerit would be remarkably convenient for the histo

See the code here, read more here

Recurrent neural networks (RNNs)

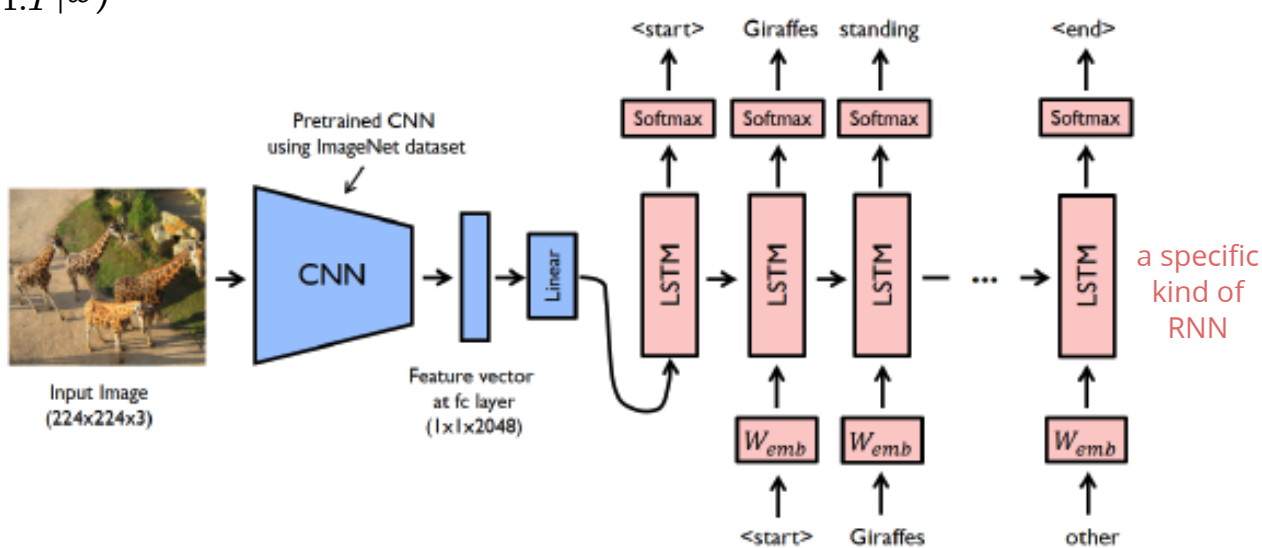
- Vec2Seq (sequence generation)

conditional generative model:

$$p(y_{1:T}|x) = \sum_{h_{1:T}} p(y_{1:T}, h_{1:T}|x)$$

Example:

CNN-RNN model for image captioning when x is embedding by a CNN



See more here

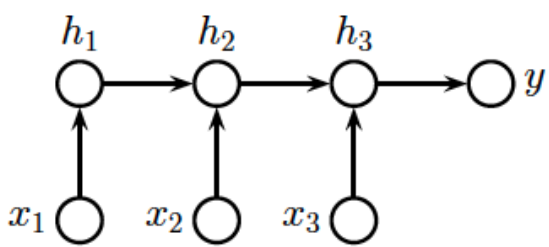
Recurrent neural networks (RNNs)

- Seq2Vec (sequence classification) $f_\theta : \mathbb{R}^{TD} \rightarrow \mathbb{R}^C$
  - predict a single fixed-length output vector given a variable length sequence as input $y \in \{1, \ldots, C\}$

use the final state:

$$\hat{y} = \text{softmax}(W h_T)$$

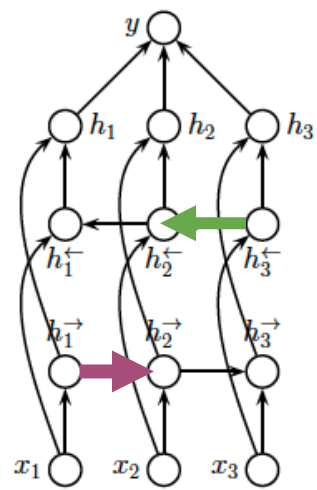
$$p(y|x_{1:T}) = \text{Categorical}(y|\hat{y})$$



Bi-directional RNN:

the hidden states of the RNN depend on the past and future context

gives better results



Recurrent neural networks (RNNs)

- Seq2Vec (sequence classification) $f_\theta : \mathbb{R}^{TD} \rightarrow \mathbb{R}^C$
  - predict a single fixed-length output vector given a variable length sequence as input

$$h_t^{\rightarrow} = \varphi\left(W_{xh}^{\rightarrow} x_t + W_{hh}^{\rightarrow} h_{t-1}^{\rightarrow}\right)$$

$$h_t^{\leftarrow} = \varphi\left(W_{xh}^{\leftarrow} x_t + W_{hh}^{\leftarrow} h_{t+1}^{\leftarrow}\right)$$

$$h_t = [h_t^{\rightarrow}, h_t^{\leftarrow}]$$

$$\overline{h} = \frac{1}{T}\sum_{t=1}^{T} h_t$$

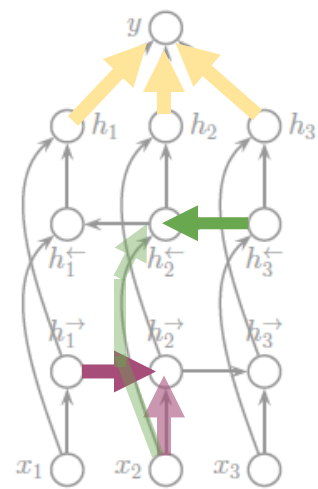
$$\hat{y} = \mathrm{softmax}(W\overline{h})$$

$$p(y|x_{1:T}) = \mathrm{Categorical}(y|\hat{y})$$

Bi-directional RNN:

the hidden states of the RNN depend on the past and future context

gives better results



Recurrent neural networks (RNNs)

- Seq2Vec (sequence classification) $f_\theta : \mathbb{R}^{TD} \rightarrow \mathbb{R}^C$
  - predict a single fixed-length output vector given a variable length sequence as input

Example:

Sentiment classification with word level **bidirectional** LSTM trained on a subset of the Internet Movie Database (IMDB) reviews. (20k positive and 20k negative examples)

IMDb

Prediction examples for two inputs:

'this movie is so great' $\Rightarrow$ 'positive'

'this movie is so bad' $\Rightarrow$ 'negative'

see the code here, and read more here

Recurrent neural networks (RNNs)

- Seq2Seq (sequence translation)
  - aligned: $T = T'$
  - unaligned: $T \neq T'$

$$f_\theta : \mathbb{R}^{TD} \rightarrow \mathbb{R}^{T'C}$$

Recurrent neural networks (RNNs)

- Seq2Seq (sequence translation)
  - aligned: $T = T'$

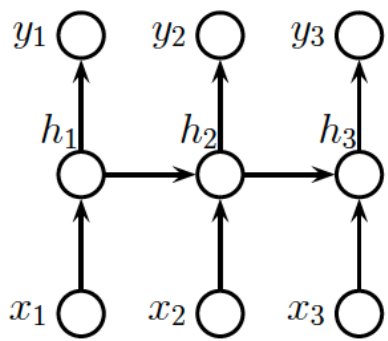
$$f_\theta : \mathbb{R}^{TD} \rightarrow \mathbb{R}^{TC}$$

modify the RNN as:

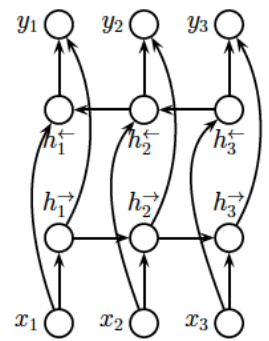
$$p\left(y_{1:T} \mid x_{1:T}\right) = \sum_{h_{1:T}} \prod_{t=1}^{T} p\left(y_t \mid h_t\right) \mathbb{I}\left(h_t = f\left(h_{t-1}, x_t\right)\right)$$

initial state: $h_1 = f\left(h_0, x_1\right) = f_0\left(x_1\right)$

dense sequence labeling:
predict one label per location



Bi-directional



Recurrent neural networks (RNNs)

- Seq2Seq (sequence translation)
  - aligned: $T = T'$

$$f_\theta : \mathbb{R}^{TD} \rightarrow \mathbb{R}^{TC}$$

modify the RNN as:

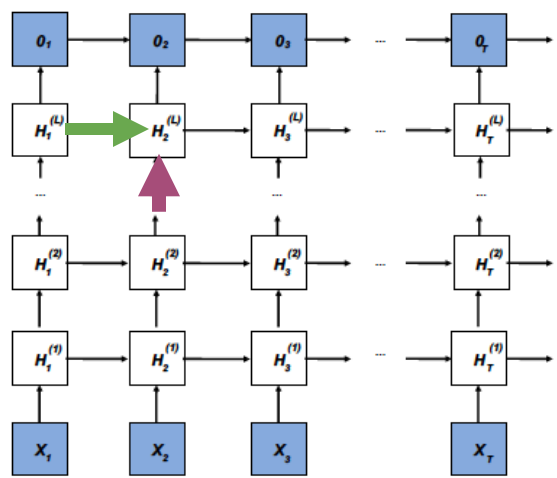
$$p\left(y_{1:T} \mid x_{1:T}\right) = \sum_{h_{1:T}} \prod_{t=1}^{T} p\left(y_t \mid h_t\right) \mathbb{I}\left(h_t = f\left(h_{t-1}, x_t\right)\right)$$

more depth to be more

input-to-hidden weights　　hidden-to-hidden weights

$$h_t^l = \varphi_l\left(W_{xh}^l h_t^{l-1} + W_{hh}^l h_{t-1}^l\right)$$

$$y_t = W_{hy} h_t^L$$



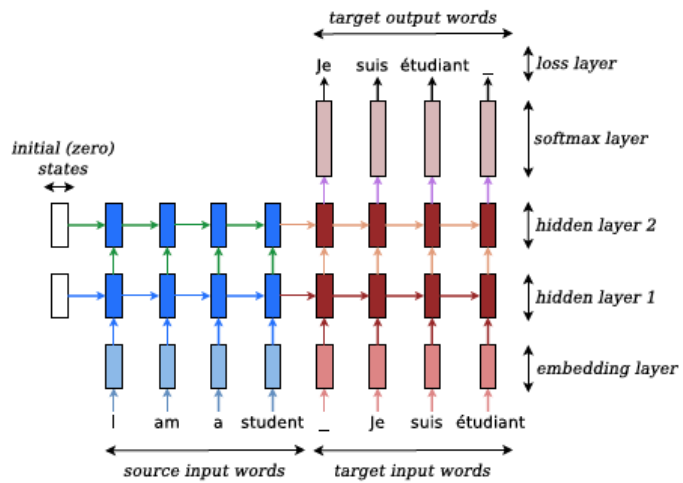
Recurrent neural networks (RNNs)

- Seq2Seq (sequence translation)
  - unaligned: $T \neq T'$

$$f_\theta : \mathbb{R}^{TD} \rightarrow \mathbb{R}^{T'C}$$

- encode the input sequence to get the context vector, the last state of an RNN, $c = f_e(x_{1:T})$
- generate the output sequence using an RNN decoder, $y_{1:T'} = f_d(c)$

Example: translating English to French



see code here

optional

Training: Backpropagation through time (BPTT)

unroll the computation graph, then apply the backpropagation

Example:

model

$$h_t = W_{hx} x_t + W_{hh} h_{t-1}$$

$$\hat{y}_t = W_{hy} h_t$$

loss

$$L = \frac{1}{T} \sum_{t=1}^{T} \ell\left(y_t, \hat{y}_t\right)$$

derivatives

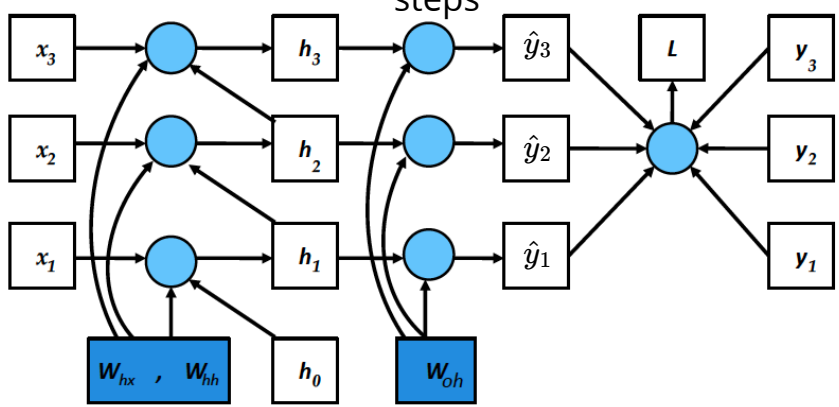
$$\frac{\partial L}{\partial W_{hx}}$$

$$\frac{\partial L}{\partial W_{hh}}$$

$$\frac{\partial L}{\partial W_{hy}}$$

Example:

An RNN unrolled (vertically) for 3 time steps



optional

Training: Backpropagation through time (BPTT)

unroll the computation graph, then apply the backpropagation

Example:

model

$$h_t = W_{hx}x_t + W_{hh}h_{t-1} = f\left(x_t, h_{t-1}, w_h\right)$$

$[\mathrm{vec}(W_{hx}); \mathrm{vec}(W_{hh})]$

$$\hat{y}_t = W_{hy}h_t = g(h_t, w_y)$$

loss

$$L = \frac{1}{T}\sum_{t=1}^{T} \ell\left(y_t, \hat{y}_t\right)$$

derivatives

$$\left.\begin{matrix}\frac{\partial L}{\partial W_{hx}} \\ \frac{\partial L}{\partial W_{hh}} \\ \frac{\partial L}{\partial W_{hy}}\end{matrix}\right\} \frac{\partial L}{\partial w_h} = \frac{1}{T}\sum_{t=1}^{T} \frac{\partial \ell(y_t,\hat{y}_t)}{\partial w_h} = \frac{1}{T}\sum_{t=1}^{T} \frac{\partial \ell(y_t,\hat{y}_t)}{\partial \hat{y}_t} \frac{\partial g(h_t,w_y)}{\partial h_t} \frac{\partial h_t}{\partial w_h}$$

$$\frac{\partial f(x_t,h_{t-1},w_h)}{\partial w_h} + \frac{\partial f(x_t,h_{t-1},w_h)}{\partial h_{t-1}} \frac{\partial h_{t-1}}{\partial w_h}$$

expand this recursively

$$\frac{\partial h_t}{\partial w_h} = \frac{\partial f(x_t,h_{t-1},w_h)}{\partial w_h} + \sum_{i=1}^{t-1}\left(\prod_{j=i+1}^{t} \frac{\partial f(x_j,h_{j-1},w_h)}{\partial h_{j-1}}\right) \frac{\partial f(x_i,h_{i-1},w_h)}{\partial w_h}$$

see code here

optional

Gating and long term memory

Vanishing and exploding gradients

activations can decay or explode as we go forwards and backwards in time

RNN variations that circumvent this:

- Gated recurrent units (GRU)
  - learns when to update the hidden state, by using a gating unit
- Long short term memory (LSTM)
  - augments the hidden state with a memory cell

Limitations of RNNS

We saw that RNNs have some important limitations that result in suboptimal performance:

1. All the information about the past of a sequence is encoded in the hidden state

Result: Loss of context when processing long sequences

2. The sequence is processed one token at the time using the current token and past hidden-state

Result: Processing cannot be parallelized in time and is slow for long sequences

Attention

$$z = g(Wx)$$

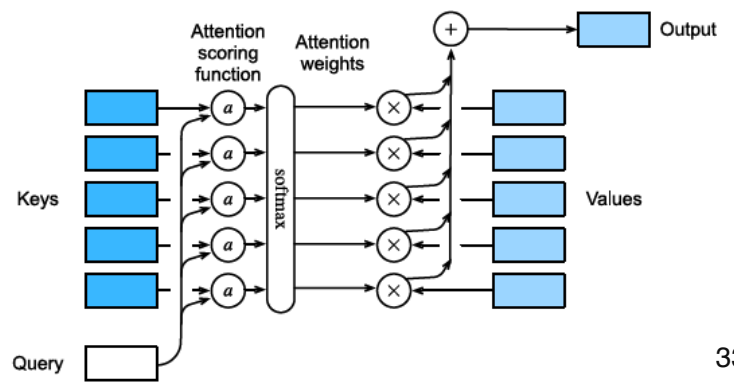
Instead of linear combination of the input activations, the model dynamically decides (in an input dependent way) which one to use based on how similar the input **query** vector $q \in \mathbb{R}^q$ is to a set of m **keys** $K \in \mathbb{R}^{m \times k}$. If q is most similar to key i, then we use value v_i.

$$\text{Attn}\left(q, \left(k_1, v_1\right), \ldots, \left(k_m, v_m\right)\right) = \text{Attn}\left(q, \left(k_{1:m}, v_{1:m}\right)\right) = \sum_{i=1}^{m} \alpha_i \left(q, k_{1:m}\right) v_i \in \mathbb{R}^v$$

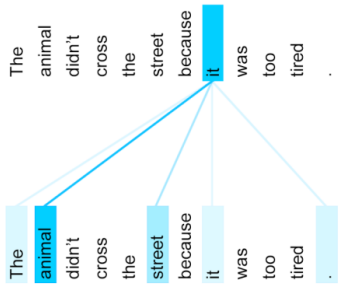
$$\alpha_i \left(q, k_{1:m}\right) = \text{softmax}_i \left(\left[a\left(q, k_1\right), \ldots, a\left(q, k_m\right)\right]\right) = \frac{\exp(a(q,k_i))}{\sum_{j=1}^{m} \exp(a(q,k_j))}$$

attention weight

The attention weights are computed from an attention score function $a(q, k_i) \in \mathbb{R}$, which gives the similarity of query q to key k_i.

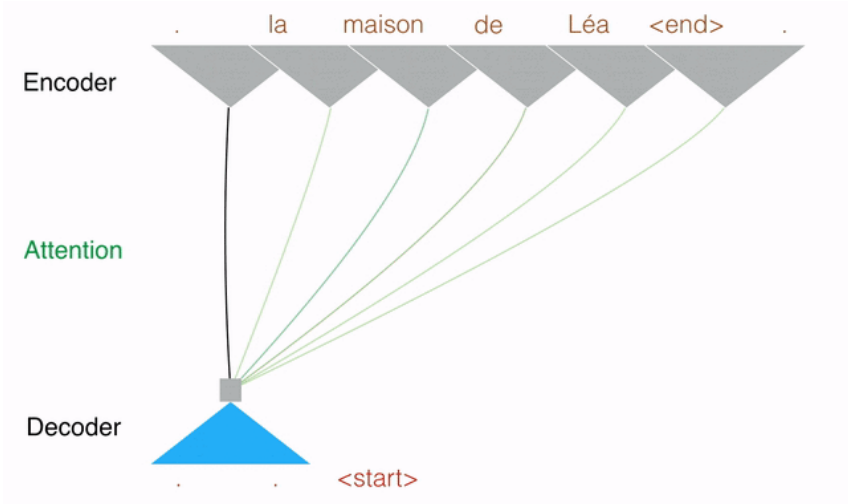


The Idea



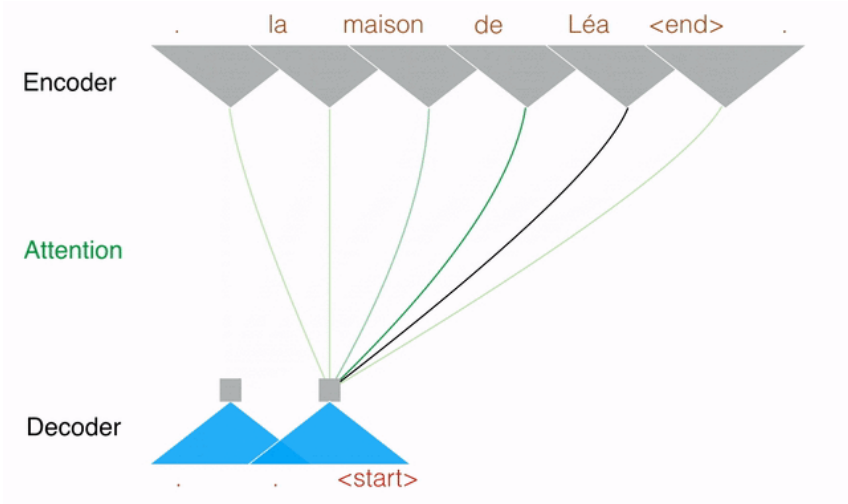
source

The Idea



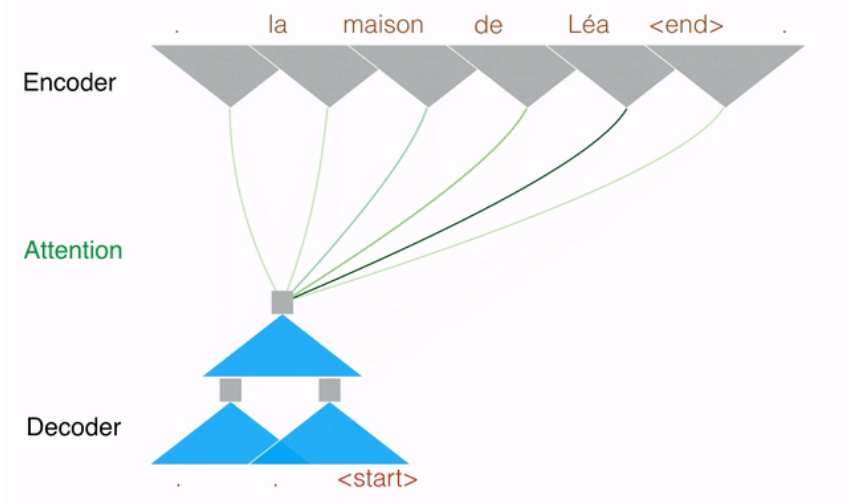
source

The Idea



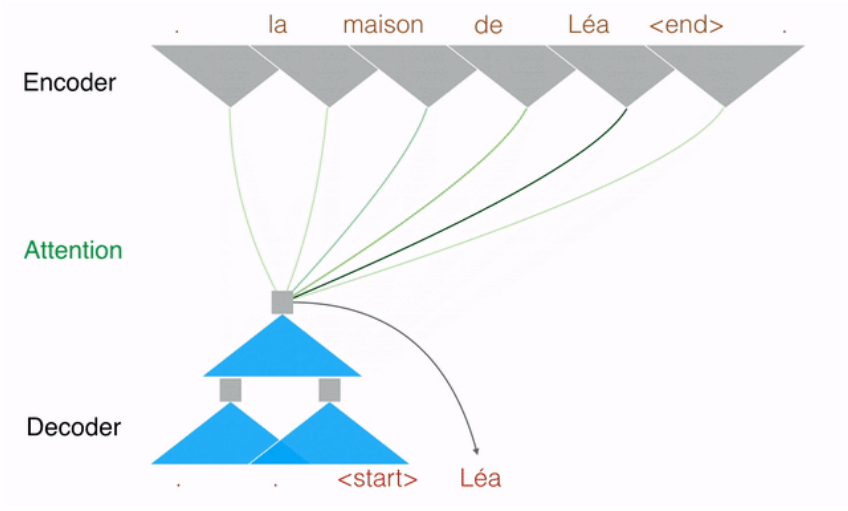
source

The Idea



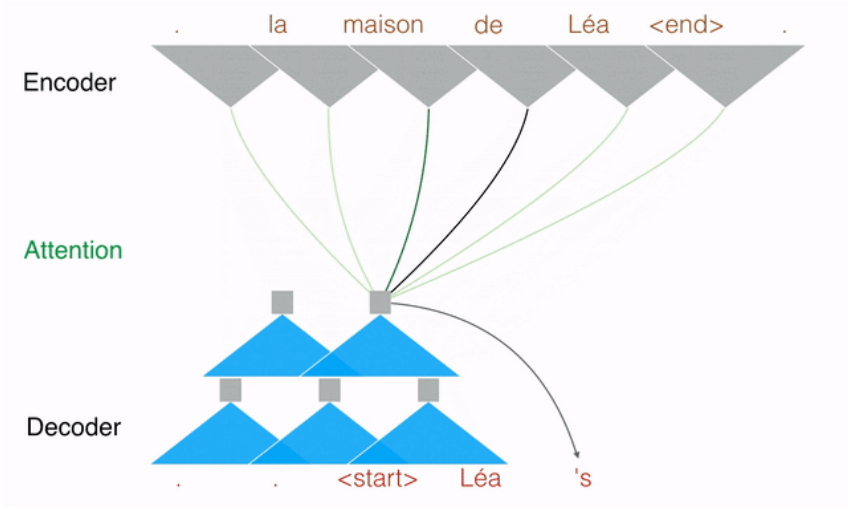
source

The Idea



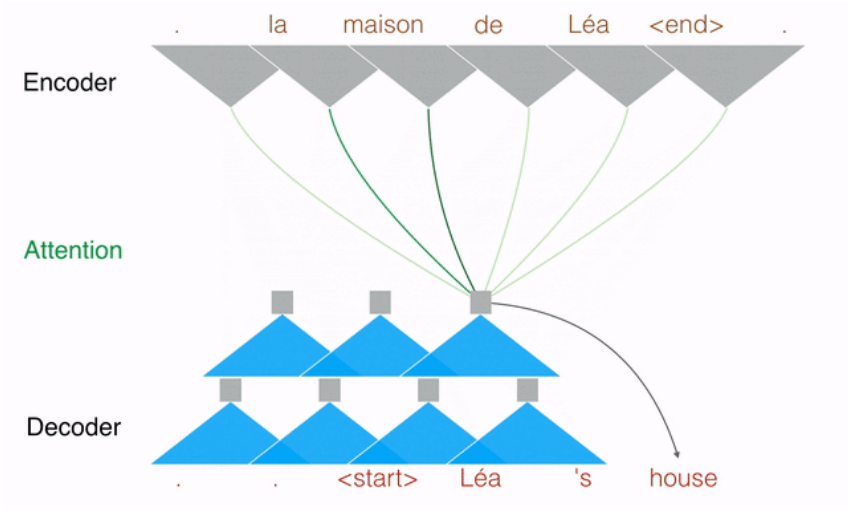
source

The Idea



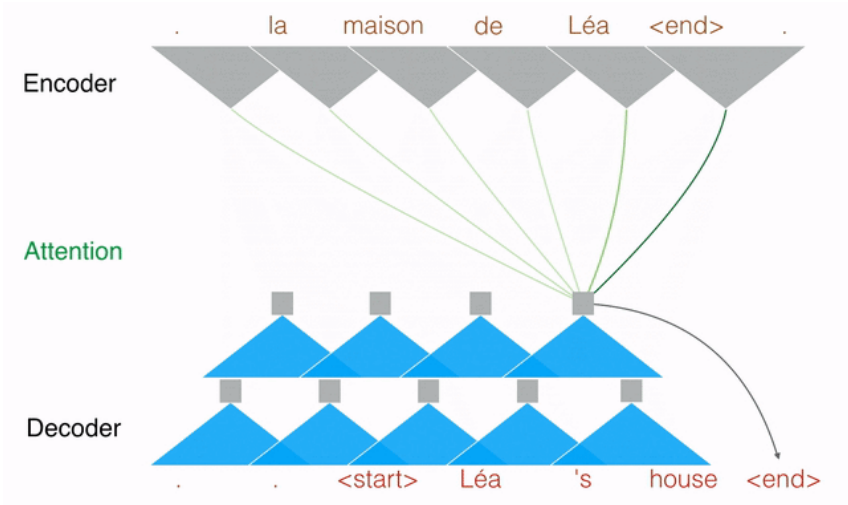
source

The Idea



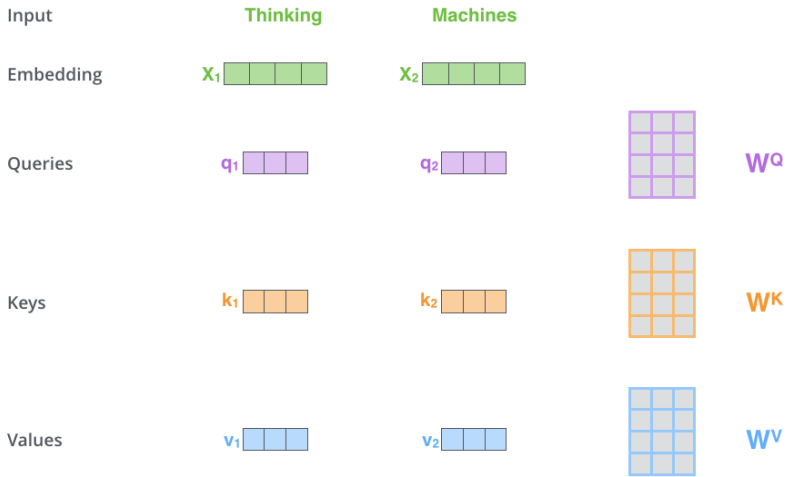
source

The Idea



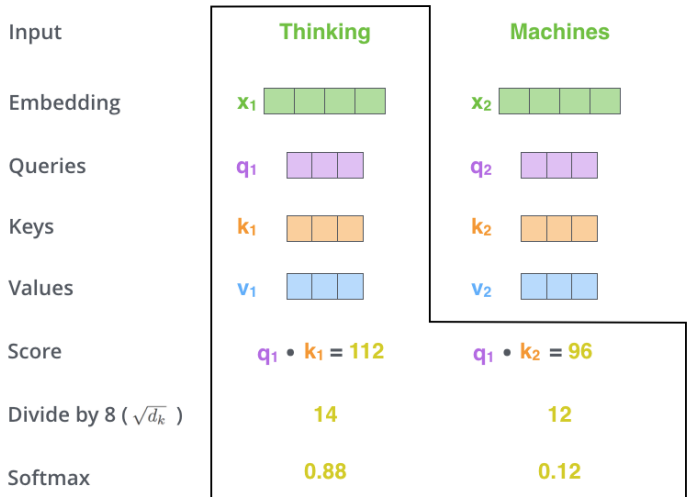
source

Attention Mechanism – High Level



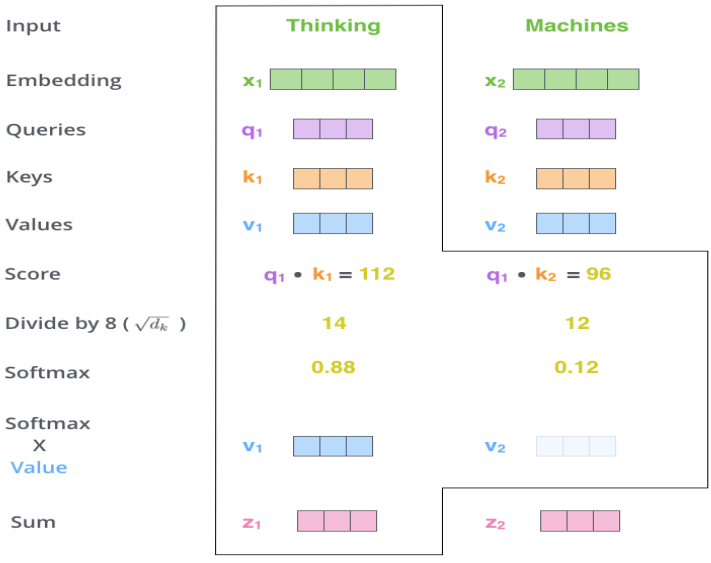
source

Attention Mechanism – High Level



source

Attention Mechanism – High Level



Attention Mechanism – Math

Let $\{x_t \in \mathbb{R}^n | t \in \{1, \dots T\}\}$ be a sequence of vectors being attended to, and $\{y_s \in \mathbb{R}^m | s \in \{1, \dots S\}\}$ being a sequence of vectors doing the attending. Then we have that:

Attention Mechanism – Math

Let $\{x_t \in \mathbb{R}^n | t \in \{1, \dots T\}\}$ be a sequence of vectors being attended to, and $\{y_s \in \mathbb{R}^m | s \in \{1, \dots S\}\}$ being a sequence of vectors doing the attending. Then we have that:

Query vector: $q_s = \underbrace{W^Q}_{d \times m} \cdot y_s \in \mathbb{R}^d$.

Key vector: $k_t = \underbrace{W^K}_{d \times n} \cdot x_t \in \mathbb{R}^d$.

Value vector: $v_t = \underbrace{W^V}_{r \times n} \cdot x_t \in \mathbb{R}^r$.

Attention Mechanism – Math

Let $\{x_t \in \mathbb{R}^n | t \in \{1, \dots T\}\}$ be a sequence of vectors being attended to, and $\{y_s \in \mathbb{R}^m | s \in \{1, \dots S\}\}$ being a sequence of vectors doing the attending. Then we have that:

Query vector: $q_s = \underbrace{W^Q}_{d \times m} \cdot y_s \in \mathbb{R}^d$.

Key vector: $k_t = \underbrace{W^K}_{d \times n} \cdot x_t \in \mathbb{R}^d$.

Value vector: $v_t = \underbrace{W^V}_{r \times n} \cdot x_t \in \mathbb{R}^r$.

The more q_s is aligned with k_t, the more the token at time s pays attention to the token at time t. The attention paid by the token at time s to the token at time t is

$$a_{st} = \frac{\exp(q_s \cdot k_t / \sqrt{d})}{\sum_{i=1}^{T} \exp(q_s \cdot k_t / \sqrt{d})}$$

Attention Mechanism – Math

Let $\{x_t \in \mathbb{R}^n | t \in \{1, \dots T\}\}$ be a sequence of vectors being attended to, and $\{y_s \in \mathbb{R}^m | s \in \{1, \dots S\}\}$ being a sequence of vectors doing the attending. Then we have that:

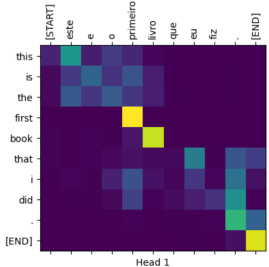
Query vector: $q_s = \underbrace{W^Q}_{d \times m} \cdot y_s \in \mathbb{R}^d$.

Key vector: $k_t = \underbrace{W^K}_{d \times n} \cdot x_t \in \mathbb{R}^d$.

Value vector: $v_t = \underbrace{W^V}_{r \times n} \cdot x_t \in \mathbb{R}^r$.

$$a_{st} = \frac{\exp(q_s \cdot k_t / \sqrt{d})}{\sum_{i=1}^{T} \exp(q_s \cdot k_t / \sqrt{d})}$$

The more q_s is aligned with k_t, the more the token at time s pays attention to the token at time t. The attention paid by the token at time s to the token at time t is



Attention Mechanism – Math

Let $\{x_t \in \mathbb{R}^n | t \in \{1, \dots T\}\}$ be a sequence of vectors being attended to, and $\{y_s \in \mathbb{R}^m | s \in \{1, \dots S\}\}$ being a sequence of vectors doing the attending. Then we have that:

Query vector: $q_s = \underbrace{W^Q}_{d \times m} \cdot y_s \in \mathbb{R}^d$.

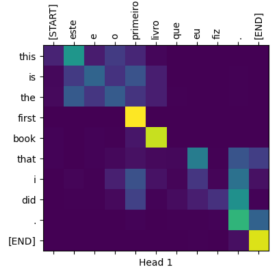
Key vector: $k_t = \underbrace{W^K}_{d \times n} \cdot x_t \in \mathbb{R}^d$.

Value vector: $v_t = \underbrace{W^V}_{r \times n} \cdot x_t \in \mathbb{R}^r$.

$$a_{st} = \frac{\exp(q_s \cdot k_t / \sqrt{d})}{\sum_{i=1}^{T} \exp(q_s \cdot k_t / \sqrt{d})}$$

$$z_s = \sum_{t=1}^{T} a_{st} v_t$$

The more q_s is aligned with k_t, the more the token at time s pays attention to the token at time t. The attention paid by the token at time s to the token at time t is



Attention Mechanism – Math

$$\underbrace{X}_{T\times n} = \begin{bmatrix} x_1 \\ \cdots \\ x_T \end{bmatrix}, \qquad \underbrace{Y}_{S\times m} = \begin{bmatrix} y_1 \\ \cdots \\ y_S \end{bmatrix}, \qquad \underbrace{Z}_{S\times r} = \begin{bmatrix} z_1 \\ \cdots \\ z_S \end{bmatrix},$$

Attention Mechanism – Math

$$\underbrace{X}_{T\times n} = \begin{bmatrix} x_1 \\ \cdots \\ x_T \end{bmatrix}, \qquad \underbrace{Y}_{S\times m} = \begin{bmatrix} y_1 \\ \cdots \\ y_S \end{bmatrix}, \qquad \underbrace{Z}_{S\times r} = \begin{bmatrix} z_1 \\ \cdots \\ z_S \end{bmatrix},$$

$$\underbrace{Q}_{S\times d} = Y \cdot W^{Q\top} = \begin{bmatrix} q_1 \\ \cdots \\ q_S \end{bmatrix}, \qquad \underbrace{K}_{T\times d} = X \cdot W^{K\top} = \begin{bmatrix} k_1 \\ \cdots \\ k_T \end{bmatrix}, \quad \underbrace{V}_{T\times r} = X \cdot W^{V\top} = \begin{bmatrix} v_1 \\ \cdots \\ v_T \end{bmatrix}$$

Attention Mechanism – Math

$$\underbrace{X}_{T\times n} = \begin{bmatrix} x_1 \\ \cdots \\ x_T \end{bmatrix}, \qquad \underbrace{Y}_{S\times m} = \begin{bmatrix} y_1 \\ \cdots \\ y_S \end{bmatrix}, \qquad \underbrace{Z}_{S\times r} = \begin{bmatrix} z_1 \\ \cdots \\ z_S \end{bmatrix},$$

$$\underbrace{Q}_{S\times d} = Y \cdot {W^Q}^\top = \begin{bmatrix} q_1 \\ \cdots \\ q_S \end{bmatrix}, \qquad \underbrace{K}_{T\times d} = X \cdot {W^K}^\top = \begin{bmatrix} k_1 \\ \cdots \\ k_T \end{bmatrix}, \quad \underbrace{V}_{T\times r} = X \cdot {W^V}^\top = \begin{bmatrix} v_1 \\ \cdots \\ v_T \end{bmatrix}$$

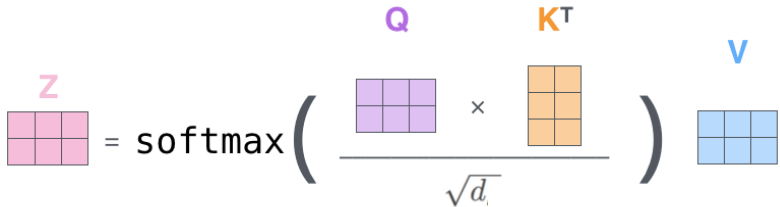
$$\underbrace{A}_{S\times T} = \mathrm{softmax}_{\mathrm{dim=T}}(Q \cdot K^\top / \sqrt{d}), \qquad \underbrace{Z}_{S\times r} = A \cdot V$$

Attention Mechanism – Math

$$\underbrace{X}_{T\times n} = \begin{bmatrix} x_1 \\ \cdots \\ x_T \end{bmatrix}, \qquad \underbrace{Y}_{S\times m} = \begin{bmatrix} y_1 \\ \cdots \\ y_S \end{bmatrix}, \qquad \underbrace{Z}_{S\times r} = \begin{bmatrix} z_1 \\ \cdots \\ z_S \end{bmatrix},$$

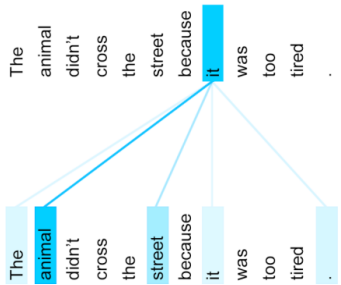
$$\underbrace{Q}_{S\times d} = Y \cdot W^{Q^\top} = \begin{bmatrix} q_1 \\ \cdots \\ q_S \end{bmatrix}, \qquad \underbrace{K}_{T\times d} = X \cdot W^{K^\top} = \begin{bmatrix} k_1 \\ \cdots \\ k_T \end{bmatrix}, \quad \underbrace{V}_{T\times r} = X \cdot W^{V^\top} = \begin{bmatrix} v_1 \\ \cdots \\ v_T \end{bmatrix}$$

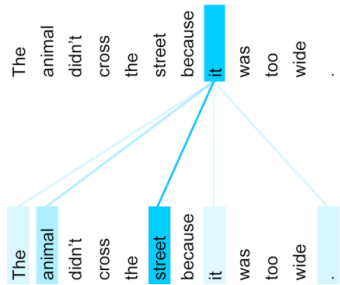
$$\underbrace{A}_{S\times T} = \text{softmax}_{\text{dim}=\text{T}}(Q \cdot K^\top/\sqrt{d}), \qquad \underbrace{Z}_{S\times r} = A \cdot V$$



Self-Attention

If the sequence $\{x_t \in \mathbb{R}^n | t \in \{1, \dots T\}\}$ being attended to, is the same as the sequence $\{y_s \in \mathbb{R}^m | s \in \{1, \dots S\}\}$ doing the attending, i.e. if $X = Y$, we call it **SELF-ATTENTION.**





Attention Mechanism – Example

$$x_t = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}, \quad y = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}.$$

$$W^K = \mathbb{I}, \quad , W^V = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix}, \quad W^Q = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & -1 \end{bmatrix}.$$

$$z =$$

Attention Mechanism – Example

$$x_t = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}, \quad y = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}.$$

$$W^K = \mathbb{I}, \quad , W^V = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix}, \quad W^Q = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & -1 \end{bmatrix}.$$

$$k_t = \left\{ \mathbb{I} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbb{I} \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbb{I} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\} = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}$$

Attention Mechanism – Example

$$x_t = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}, \quad y = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}.$$

$$k_t = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}, \quad W^V = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix}, \quad W^Q = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & -1 \end{bmatrix}.$$

$$v_t =$$

Attention Mechanism – Example

$$x_t = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}, \quad y = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}.$$

$$k_t = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}, \quad W^V = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix}, \quad W^Q = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & -1 \end{bmatrix}.$$

$$v_t = \left\{ W^V \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad W^V \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad W^V \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}$$

Attention Mechanism – Example

$$x_t = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}, \quad y = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}.$$

$$k_t = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}, \quad W^V = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix}, \quad W^Q = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & -1 \end{bmatrix}.$$

$$v_t = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad W^V \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad W^V \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}$$

Attention Mechanism – Example

$$x_t = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}, \quad y = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}.$$

$$k_t = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}, \quad W^V = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix}, \quad W^Q = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & -1 \end{bmatrix}.$$

$$v_t = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad W^V \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}$$

Attention Mechanism – Example

$$x_t = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}, \quad y = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}.$$

$$k_t = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}, \quad W^V = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix}, \quad W^Q = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & -1 \end{bmatrix}.$$

$$v_t = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\}$$

Attention Mechanism – Example

$$x_t = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}, \quad y = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}.$$

$$k_t = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}, \quad v_t = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\}, \quad W^Q = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & -1 \end{bmatrix}.$$

$$q =$$

Attention Mechanism – Example

$$x_t = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}, \quad y = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}.$$

$$k_t = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}, \quad v_t = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\}, \quad W^Q = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & -1 \end{bmatrix}.$$

$$q = W^Q y = \begin{bmatrix} -1 \\ 1 \\ 0 \\ 1 \end{bmatrix}$$

Attention Mechanism – Example

$$x_t = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}, \quad y = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}.$$

$$k_t = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}, \quad v_t = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\}, \quad q = \begin{bmatrix} -1 \\ 1 \\ 0 \\ 1 \end{bmatrix}.$$

$$q \cdot k_t =$$

Attention Mechanism – Example

$$x_t = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}, \quad y = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}.$$

$$k_t = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}, \quad v_t = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\}, \quad q = \begin{bmatrix} -1 \\ 1 \\ 0 \\ 1 \end{bmatrix}.$$

$$q \cdot k_t = [-1, 1, -1]$$

Attention Mechanism – Example

$$x_t = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}, \quad y = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}.$$

$$k_t = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}, \quad v_t = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\}, \quad q = \begin{bmatrix} -1 \\ 1 \\ 0 \\ 1 \end{bmatrix}.$$

$$q \cdot k_t = [-1, 1, -1], \quad \text{softmax}(q \cdot k_t / \sqrt{d}) = \frac{[e^{-1/2}, e^{+1/2}, e^{-1/2}]}{e^{-1/2} + e^{+1/2} + e^{-1/2}}$$

Attention Mechanism – Example

$$x_t = \left\{ \begin{bmatrix}1\\0\\0\\0\end{bmatrix}, \quad \begin{bmatrix}0\\1\\0\\0\end{bmatrix}, \quad \begin{bmatrix}1\\0\\0\\0\end{bmatrix} \right\}, \quad y = \begin{bmatrix}1\\-1\\0\end{bmatrix}.$$

$$k_t = \left\{ \begin{bmatrix}1\\0\\0\\0\end{bmatrix}, \quad \begin{bmatrix}0\\1\\0\\0\end{bmatrix}, \quad \begin{bmatrix}1\\0\\0\\0\end{bmatrix} \right\}, \quad v_t = \left\{ \begin{bmatrix}1\\0\end{bmatrix}, \quad \begin{bmatrix}0\\1\end{bmatrix}, \quad \begin{bmatrix}1\\0\end{bmatrix} \right\}, \quad q = \begin{bmatrix}-1\\1\\0\\1\end{bmatrix}.$$

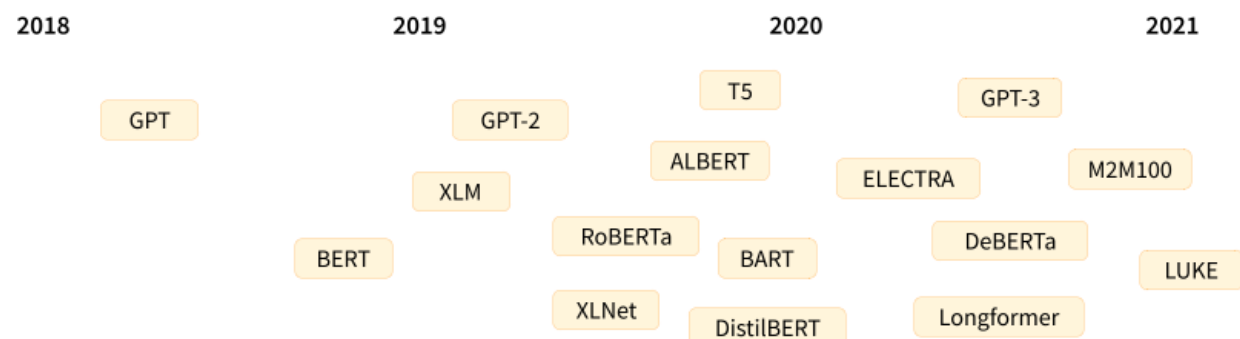
$$\text{softmax}(q \cdot k_t/\sqrt{d}) = \frac{[e^{-1/2}, e^{+1/2}, e^{-1/2}]}{e^{-1/2} + e^{+1/2} + e^{-1/2}}$$

$$z = \frac{e^{-1/2}\begin{bmatrix}1\\0\end{bmatrix} + e^{+1/2}\begin{bmatrix}0\\1\end{bmatrix} + e^{-1/2}\begin{bmatrix}1\\0\end{bmatrix}}{e^{-1/2} + e^{+1/2} + e^{-1/2}} = \begin{bmatrix}2e^{-1/2}\\e^{+1/2}\end{bmatrix} / (2e^{-1/2} + e^{+1/2})$$

Transformers

a seq2seq model which uses attention in the encoder as well as the decoder, thus eliminating the need for RNNs

- Self-attention
- Multi-headed attention
- Positional encoding



read more here

Transformers: self-attention

given a sequence of input tokens $x_1, \ldots, x_n$, generate a sequence of outputs of the same size with:

$$y_i = \mathrm{Attn}\left(x_i, \left(x_1, x_1\right), \ldots, \left(x_n, x_n\right)\right)$$

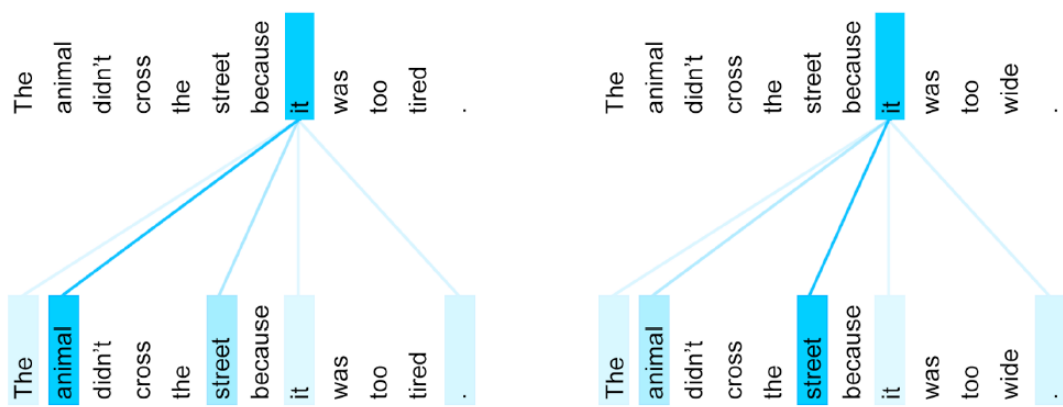
query $\in \mathbb{R}^d$; (key, value)s

for decoder we set $x_i = y_{i-1}$ and $n = i - 1$

this gives improved representations of context

Example:

coreference resolution: encoder self-attention for the word "it" differs depending on the input context which is important in translation, e.g. what pronoun to use in French



optional

Transformers: positional encoding

attention is permutation invariant, and hence ignores the input word ordering. To overcome this, we can concatenate the word embeddings with a positional embedding so that the model knows what order the words occur in

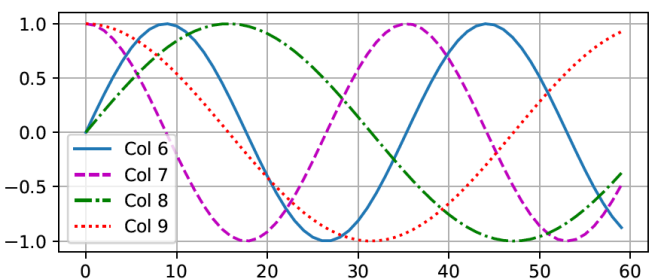
$$p_{i,2j} = \sin\left(\frac{i}{10000^{2j/d}}\right)$$
$$p_{i,2j+1} = \cos\left(\frac{i}{10000^{2j/d}}\right)$$

$$\mathrm{POS}(\mathrm{Embed}(X)) = \underbrace{X}_{\in \mathbb{R}^{n\times d}} + \underbrace{P}_{\in \mathbb{R}^{n\times d}}$$

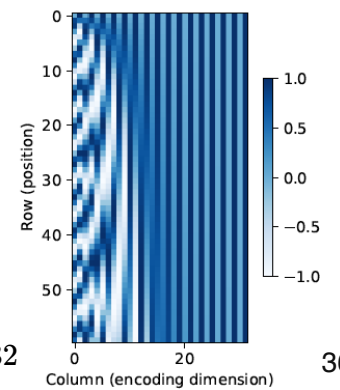
```
0 in binary is 000
1 in binary is 001
2 in binary is 010
3 in binary is 011
4 in binary is 100
5 in binary is 101
6 in binary is 110
7 in binary is 111
```

lower columns have higher frequencies

read more here



Example:



$n = 60,\ d = 32$

optional

Transformers: putting it all together

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

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

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

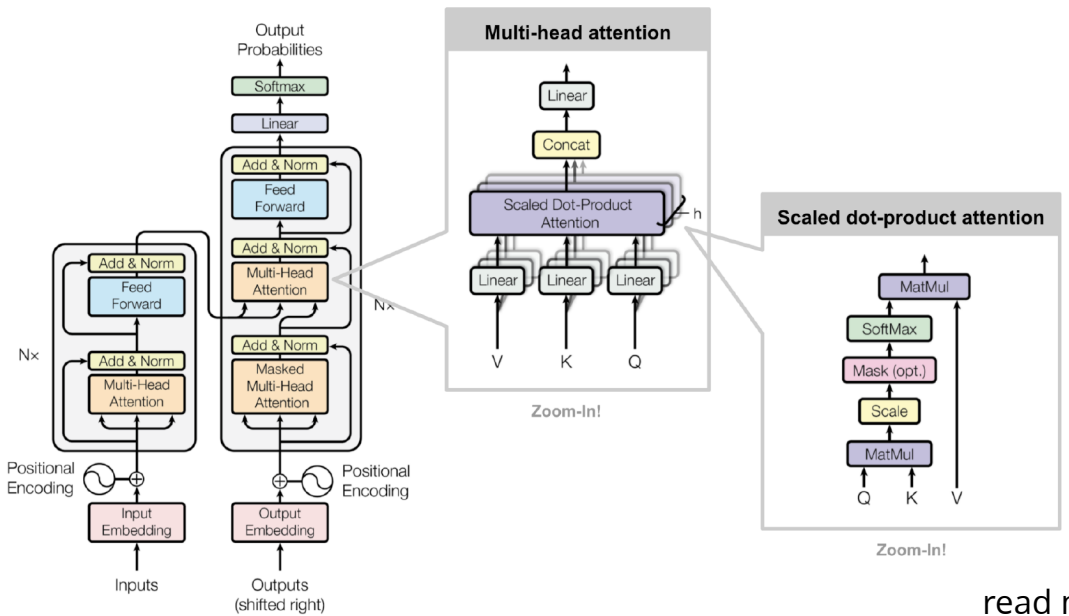
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task,

read more here, see the code here

optional

Transformers: putting it all together

A transformer is a seq2seq model that uses self-attention for the encoder and decoder rather than an RNN. The encoder uses a series of encoder blocks, each of which uses multi-headed attention, residual connections, and layer normalization



read more here, see the code here

Language models

- ELMo (Embeddings from Language Model)
  - RNN based, trained unsupervised to minimize the negative log likelihood of the input sentence, i.e. $y_t = x_{t-1}$
- BERT (Bidirectional Encoder Representations from Transformers)
  - Transformer-based: map a modified version of a sequence back to the unmodified form and compute the loss at the masked locations: fill-in-the-blank :

```
Let's make [MASK] chicken! [SEP] It [MASK] great with orange sauce
```

- GPT (Generative Pre-training Transformer)
  - uses a masked transformer as the decoder, see an open-source model here (20 billion parameters)

Generative Pre-trained Transformer (GPT)

Model	Architecture	Parameters	Training data	Release date	Training cost
GPT-1	12-level, 12-headed Transformer decoder (no encoder).	117 million	BookCorpus: 4.5 GB of text, from 7000 unpublished books.	June 11, 2018	30 days on 8 P600 GPUs, or 1 petaFLOP/s-day.
GPT-2	GPT-1, but with modified normalization	1.5 billion	WebText: 40 GB of text, 8 million documents, from 45 million webpages upvoted on Reddit.	February 14, 2019 (initial/limited version) and November 5, 2019 (full version)	"tens of petaflop/s-day", or 1.5e21 FLOP.
GPT-3	GPT-2, but with modification to allow larger scaling	175 billion	499 billion tokens of CommonCrawl (570 GB), WebText, English Wikipedia, book corpora (Books1 and Books2).	May 28, 2020	3640 petaflop/s-day.
GPT-3.5	Undisclosed	175 billion	Undisclosed	March 15, 2022	Undisclosed
GPT-4	Trained with both text prediction and RLHF.	Estimated 1.7 trillion.	Undisclosed	March 14, 2023	Undisclosed. Estimated 2.1 × 10 FLOP.

https://en.wikipedia.org/wiki/Generative_pre-trained_transformer

Summary

- Recurrent neural networks (RNNs)
  - Vec2Seq (sequence generation)
  - Seq2Vec (sequence classification)
  - Seq2Seq (sequence translation)
  - training with back propagation through time
- attention mechanisms, self-attention and multi-headed attention
- The architecture of transformer
- language models with transformer