SEQ2SEQ Modeling

SEQ2LABEL: $P(y^{(T)} \mid x^{(1)}, \ldots, x^{(T+1)})$

SEQ2SEQ: $P(y^{(1)}, \ldots, y^{(T_y)} \mid x^{(1)}, \ldots, x^{(T_x)})$

↳ APPLICATIONS

① **NMT** INPUT FRENCH SENT
OUTPUT ENGLISH SENT

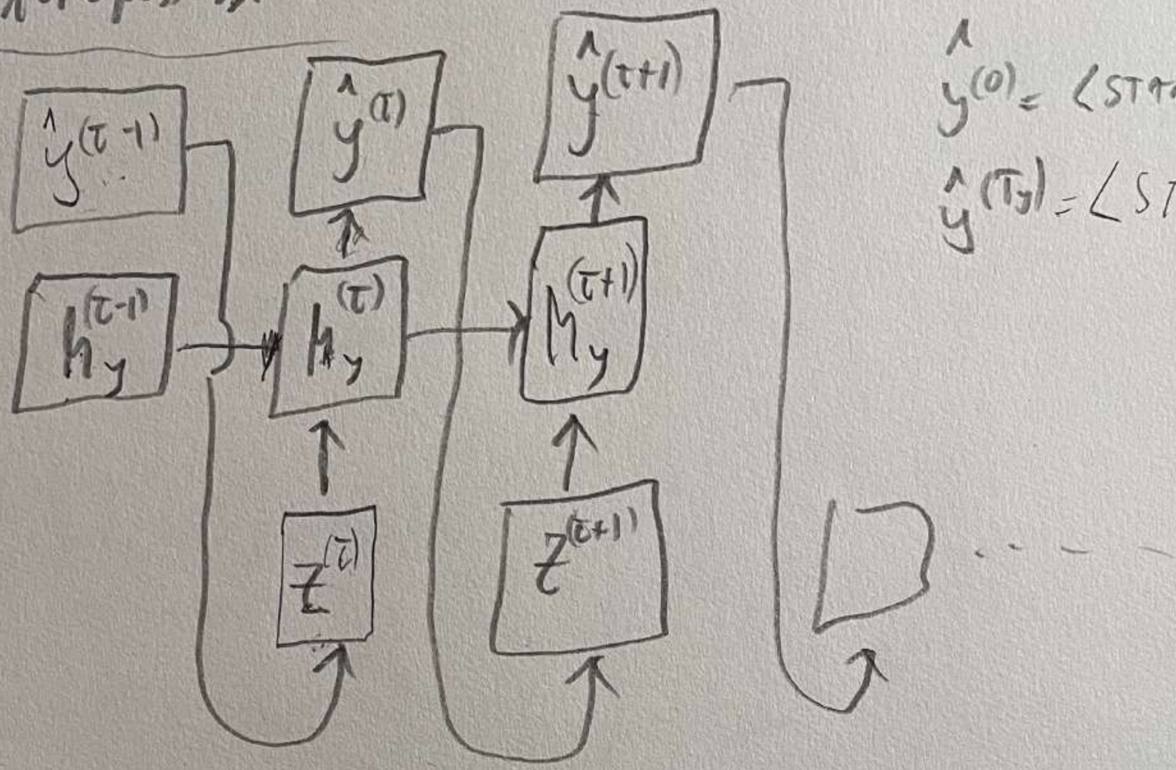
② TEXT GEN INPUT PREFIX
OUTPUT SUFFIX

↳ Factor the joint dist $P(y^{(1)} \ldots y^{(T_y)} \mid x^{(1)} \ldots x^{(T_x)})$

INTO $\prod_{t=1}^{T_y} P(y^{(t)} \mid x^{(1)}, \ldots, x^{T_x}, y^{(1)}, \ldots, y^{(t-1)})$

↳

AUTOREGRESSION



$\hat{y}^{(0)} = \langle START \rangle$

$\hat{y}^{(T_y)} = \langle STOP \rangle$

SEQ2SEQ ALGORITHM

① $h_y^{(0)} = h_x^{(T_x)} = f_h(x^{(1)}, \dots x^{(T_x)})$

② SET $\tau = 1$, $\hat{y}^{(0)} = \langle START \rangle$

③ while TRUE:

$h^{(\tau)} = f_h(h_y^{(\tau-1)}, y^{(\tau-1)})$ → Computing Hidden State

$\hat{y}^{(\tau)} = f_y(h_y^{(\tau)})$ → PREDICT τ^{th} token

if $\hat{y}^{(\tau)} = \langle STOP \rangle$;
Break

$\tau = \tau + 1$

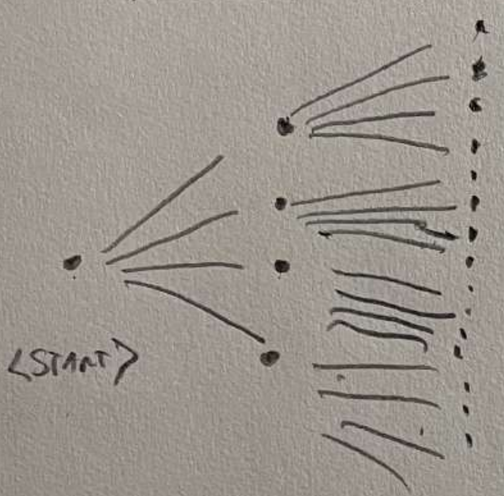
SEQ2LABEL	SEQ2SEQ
$h^{(\tau)} = f_h(h^{(\tau-1)}, \underline{x^{(\tau)}})$	$h^{(\tau)} = f_h(h^{(\tau-1)}, \underline{y^{(\tau-1)}})$
COMES from the DATA!	COMES from the MODEL!

The INFERENCE GAP

INFERENCE TASK: $\widehat{y^{(1)} \ldots y^{(T_y)}} = \underset{y^{(1)} \ldots y^{T_y}}{\text{ARGMAX}}\ P(y^{(1)} \ldots y^{(T_y)} \mid X^{(1)} \ldots X^{(T_x)})$

Training: $\hat{\theta} = \underset{\theta}{\text{Argmax}}\ P(y^{(t)} \mid f_h(h^{(t-1)}, y^{(t-1)}))$

SEARCH PROBLEM



$N = 10^5$

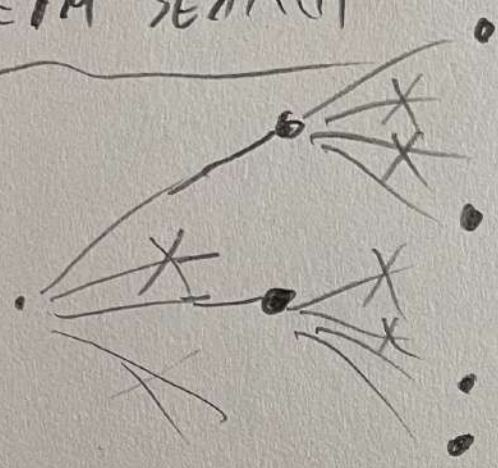
$T_y = 5$

$$P(y^{(1)}, \ldots, y^{(T_y)}) = \prod_{t=1}^{T_y} P(y^{(t)} \mid y^{(1)} \ldots y^{(t-1)}) \Rightarrow \text{MLE}$$

\# Trajectories: N^{T_y}

↳ Eg.: $(10^5)^5 = 10^{25}$ evaluations

BEAM SEARCH



Reduce branching factor from $N \rightarrow K$, where $K \ll N$.

$$N = 10^5, \quad T_y = 5 \rightarrow 10^{25}$$

$$N = 10, \quad T_y = 5 \rightarrow 10^{5}$$

INFERENCE GAP

During training: $L_{CE}\left(y^{(t)}, P\left(y^{(t)} \mid \underbrace{y^{(1)} \ldots y^{(t-1)}}\right)\right)$

LABEL PROB LABELS

"TEACHER FORCING"

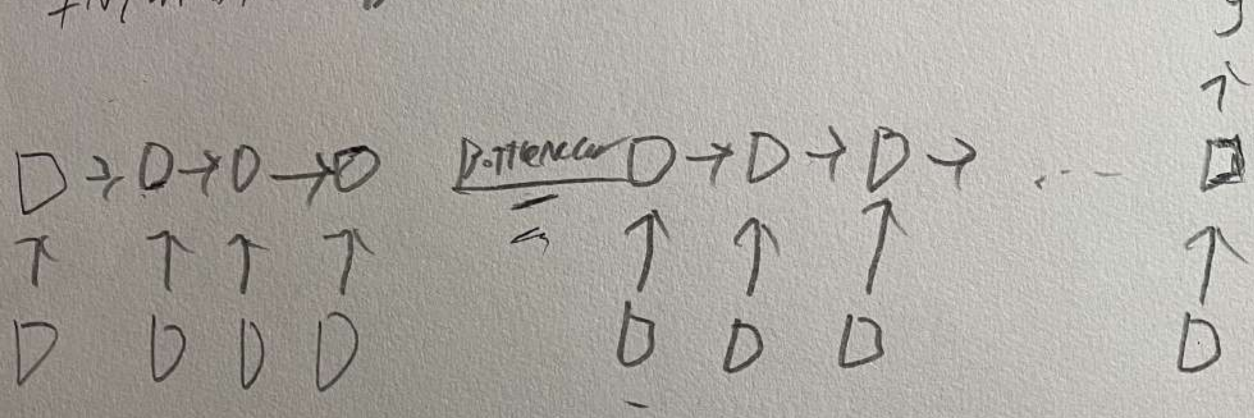
During test: $P\left(y^{(t)} \mid \hat{y}^{(1)} \ldots \hat{y}^{(t-1)}\right)$

Scheduled Sampling

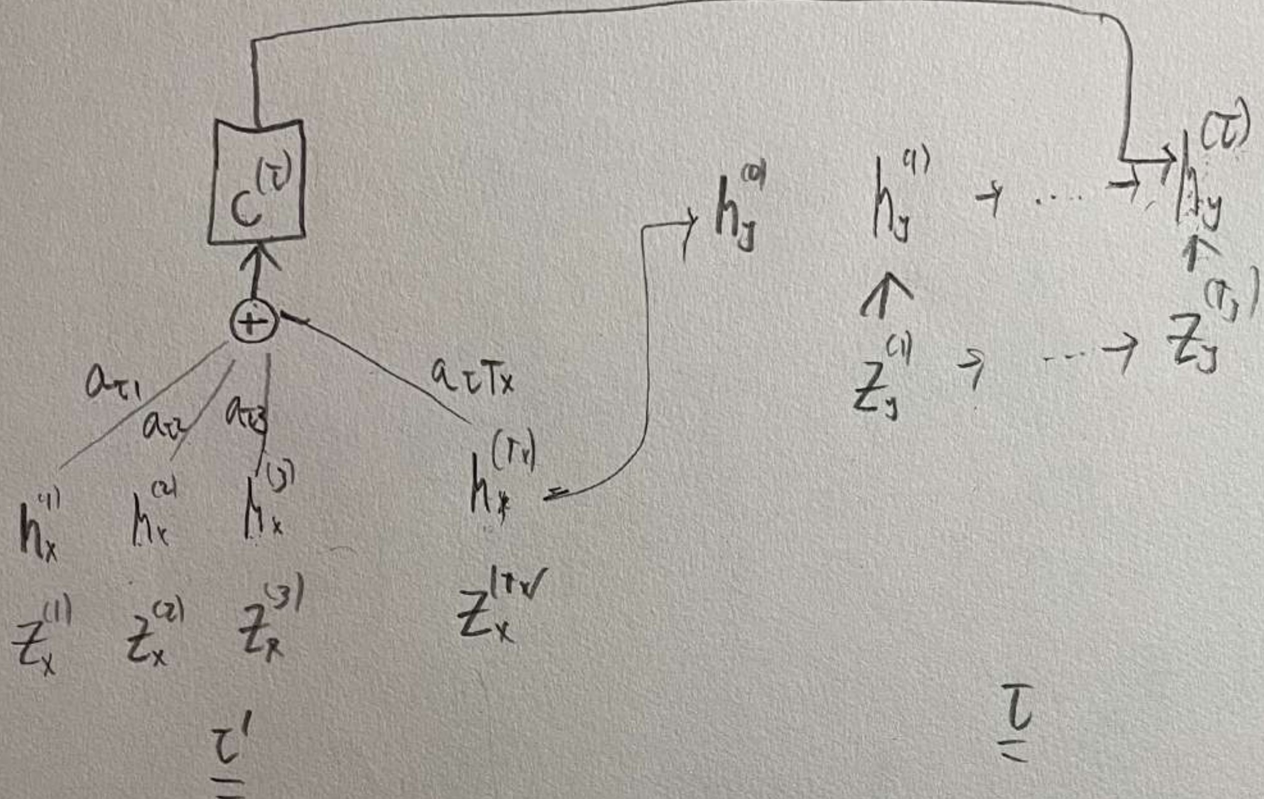
$$L_{CE}\left(y^{(t)}, P(y^{(t)} \mid \tilde{y}^{(1)}, \ldots, \tilde{y}^{(t-1)}\right)$$

$$\tilde{y}^{(t)} = \begin{cases} \hat{y}^{(t)} & \text{with probability } p \\ y^{(t)} & \text{with probability } 1-p \end{cases}$$

Information Bottleneck



ATTENTION BASED SEQ2SEQ MODELS



$$\sum_{\tau'=1}^{T_x} a_{\tau,\tau'} = 1$$

$$\rightarrow A \in [0,1]^{T_y \times T_x}$$

$a_{\tau\tau'}$ is the degree of attention that $h_y^{(\tau)}$ has on $h_x^{\tau'}$.

$$c^{(\tau)} = \sum_{\tau'=1}^{T_x} a_{\tau,\tau'} h_x^{(\tau')}$$

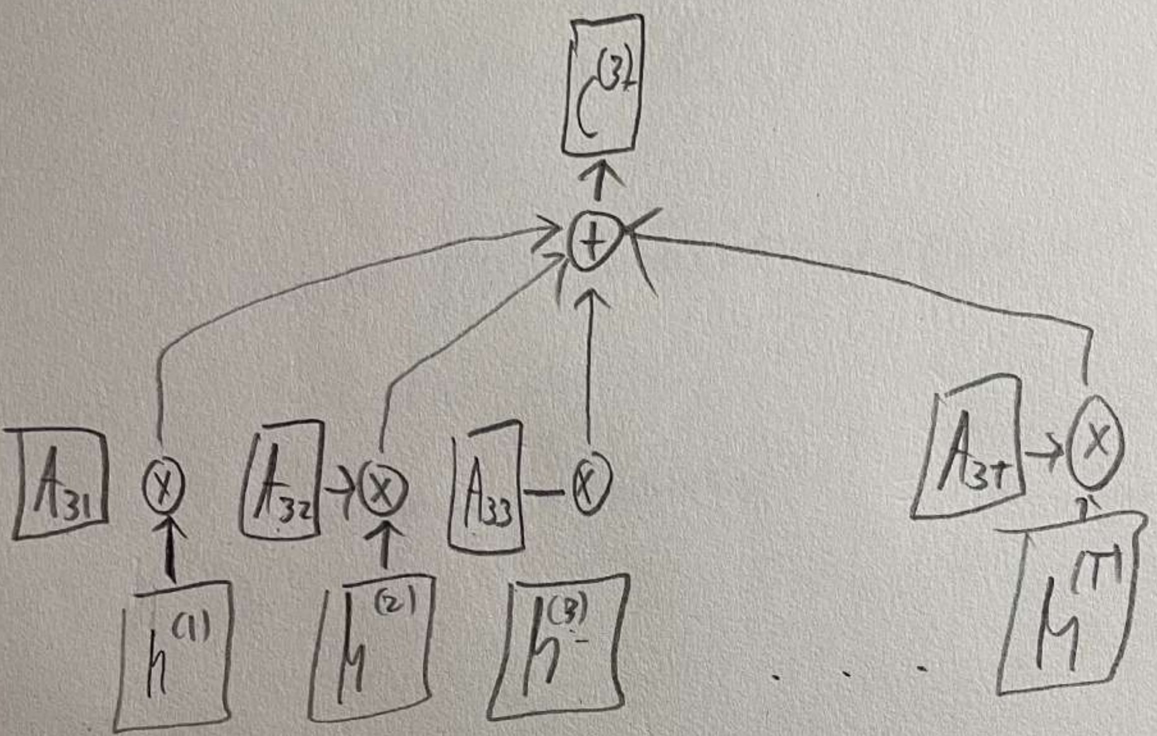
$$a_{\tau\tau'} = \frac{e^{\alpha_{\tau,\tau'}}}{\sum_{\tau''}^{T_x} e^{\alpha_{\tau,\tau''}}}$$

α? $\Rightarrow$ 1) $\alpha_{\tau\tau'} = h_y^{(\tau)} \cdot h_x^{(\tau')}$

2) $\alpha_{\tau\tau'} = f_{h_y}(h_y^{(\tau)}) \cdot f_{h_x}(h_x^{(\tau')})$

$\hookrightarrow f = NN$

Self Attention



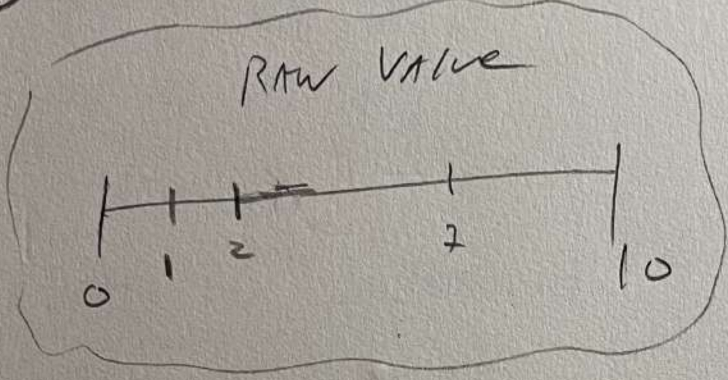
$$c^{(\tau)} = \sum_{\tau'} A_{\tau,\tau'} h^{(\tau')} \quad \text{where} \quad A_{\tau\tau'} = \frac{e^{h^{(\tau)} \cdot h^{(\tau')}}}{\sum_{\tau''} e^{h^{(\tau)} \cdot h^{(\tau'')}}}$$

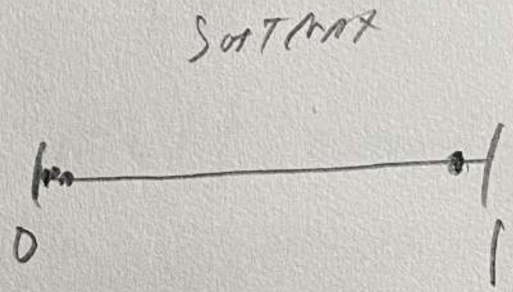
↳ Each Row $\sum A_{row} = 1$

ATTENTION IS ALL YOU NEED $\boxed{2017}$

TRICKS INTRODUCED:

① Scaled Dot Product Attention





$M \in \mathbb{R}^D \Rightarrow \|M\|$ vs D

$D=2$ $[1,1] = \sqrt{2}$

$D=3$ $[1,1,1] = \sqrt{3}$ $\sqrt{D}$

$D=4$ $[1,1,1,1] = \sqrt{4}$

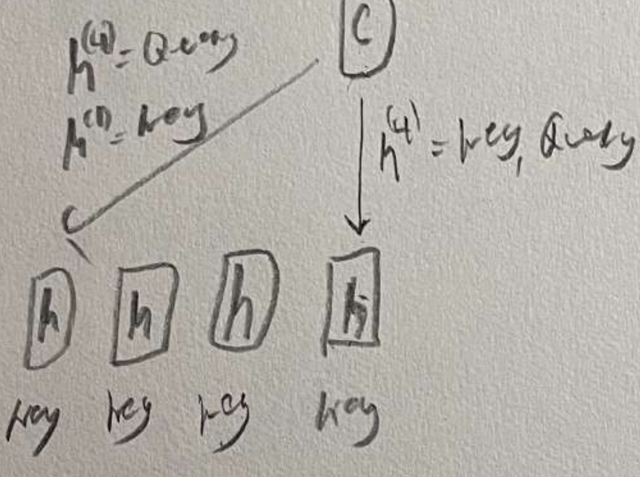
↳ Scaled Dot Product → $\boxed{\frac{1}{\sqrt{D}}}$!

② Attention is like a "soft" lookup table

for $c^{(t)} \Rightarrow$ $\boxed{c^{(t)}}$

$\boxed{h^{(t)}}$

③ Attention as a "Soft" Look up



hard

Value = f[key]

Soft

Value = f[Query, key]

For every $c^{(t)}$ there is:

① A Query $h^{(t)}$

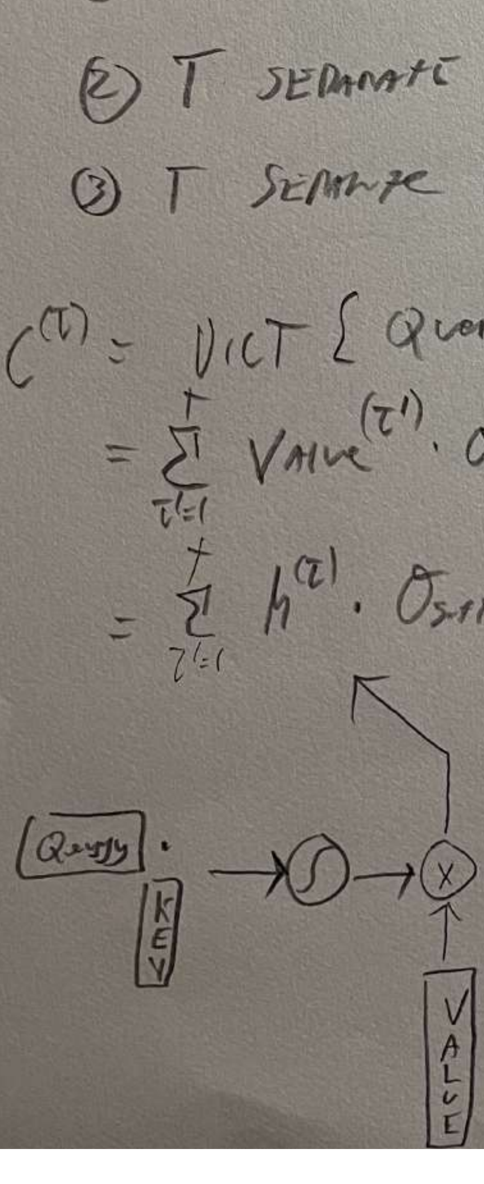
② T separate keys: $h^{(1)} \dots h^{(T)}$

③ T separate values: $h^{(1)}, \dots h^{(T)}$

$c^{(t)} = \text{DICT}[\text{Query}^{(t)}]$ ← "fuzzy" or "soft" lookup

$= \sum_{t'=1}^{T} \text{Value}^{(t')} \cdot \sigma_{softmax}(\text{Query}^{(t)} \cdot \text{key}^{(t')})$

$= \sum_{t'=1}^{T} h^{(t')} \cdot \sigma_{softmax}(h^{(t)} \cdot h^{(t')})$



Each input h serves as the key, value, query depending on which position we are at.

④ LEARNABLE SET OF KEYS, VALUES, QUERIES

$$\left.\begin{aligned} k_t &= K h^{(t)} + b_k \\ q_t &= Q h^{(t)} + b_q \\ v_t &= V h^{(t)} + b_v \end{aligned}\right\} \Rightarrow \theta = \{K, Q, V, b_k, b_q, b_v\}$$

⑤ MULTIHEADED SA:

$K = \{K^{(1)} \dots K^{(n)}\}$ n = NUMBER OF ATTENTION HEADS

$Q = \{Q^{(1)} \dots Q^{(n)}\}$

$V = \{V^{(1)} \dots V^{(n)}\}$

↳ n CONTEXT VECTORS AT EACH POSITION IN THE SEQUENCE $\Rightarrow C^{(t)} = [c_1^{(t)}, \dots, c_n^{(t)}]$

↳ $K, Q, V \in \mathbb{R}^{D/n}$ SUCH THAT WE CAN CONCAT TO GET $C^{(t)} \in \mathbb{R}^{D}$