STREAMING AUTOMATIC SPEECH RECOGNITION WITH THE TRANSFORMER MODEL

Niko Moritz, Takaaki Hori, Jonathan Le Roux

Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA

ASR (Automatic Speech Recognition)

Traditional

HMM-DNN

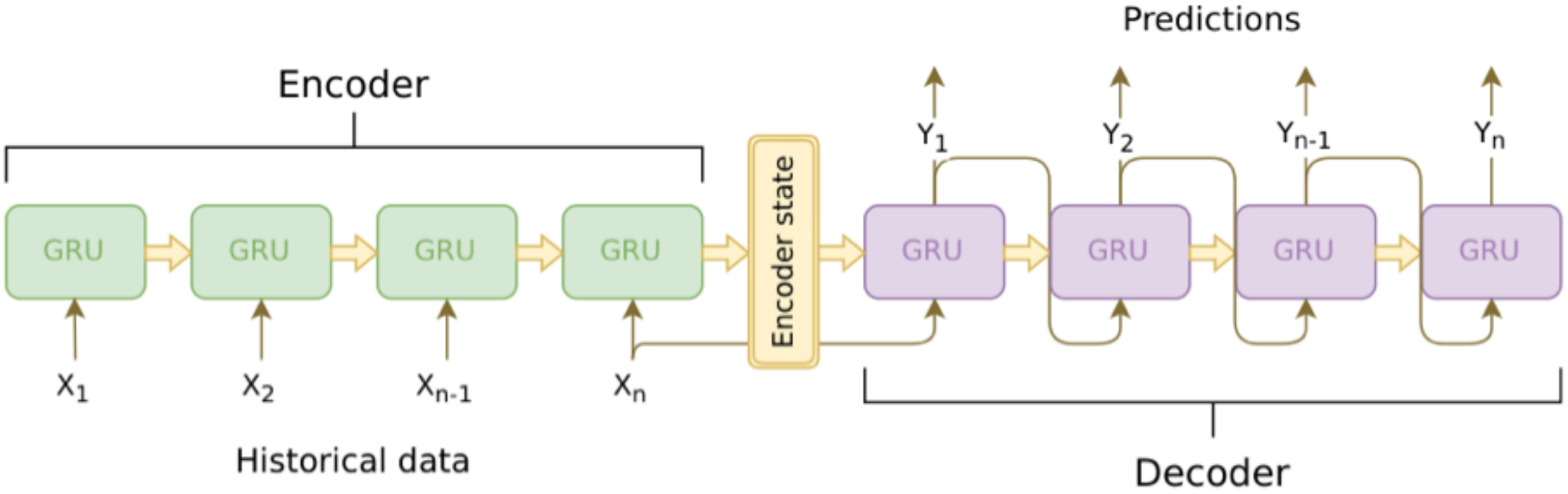
End-to-End

CTC (Connectionist Temporal Classification)

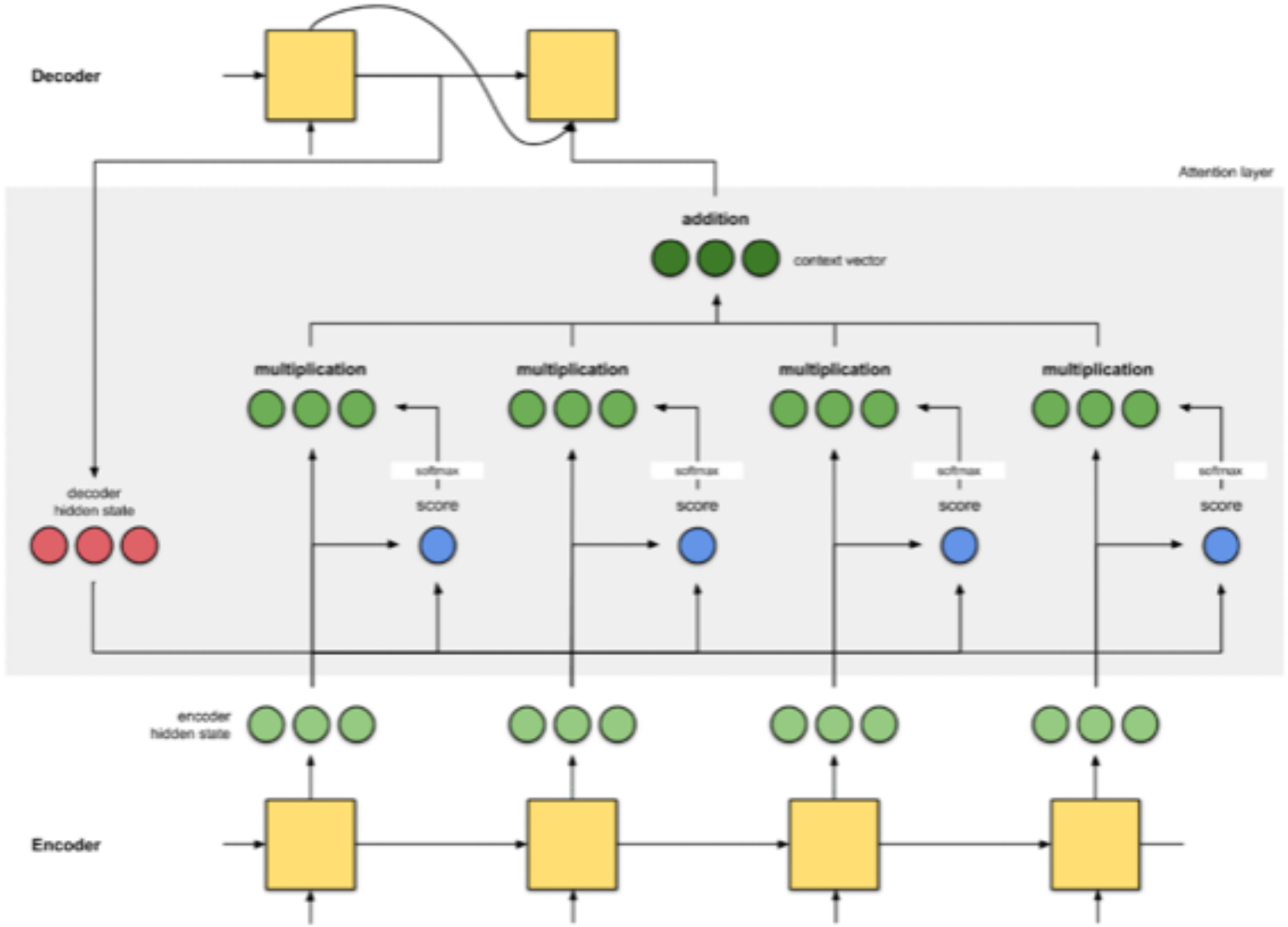
Attention Seq-to-Seq

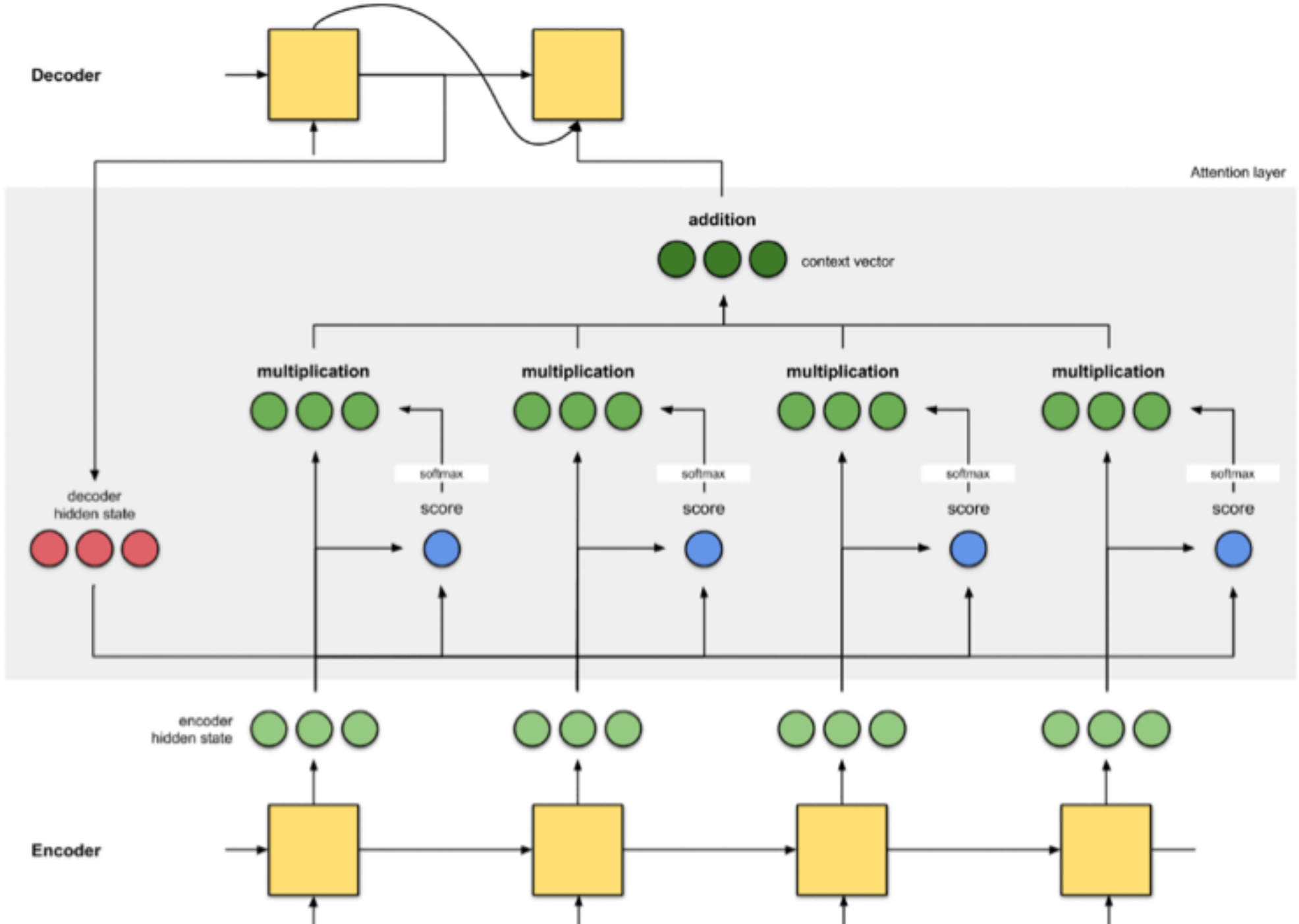
RNN-T

Seq-to-Seq

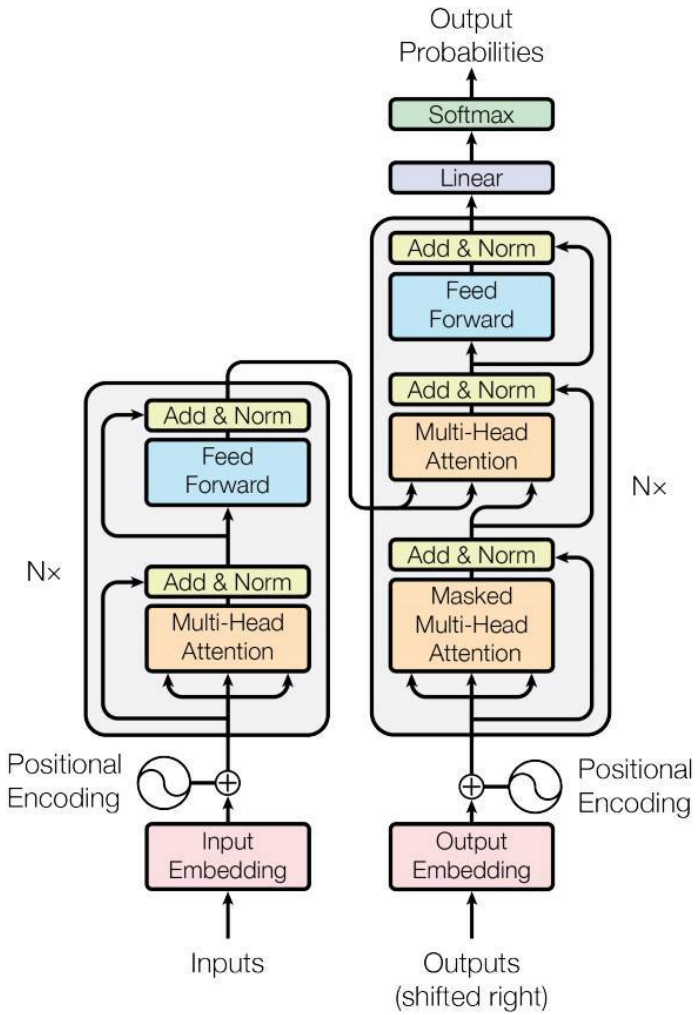

Predictions
Encoder
GRU
GRU
GRU
GRU
Encoder state
GRU
GRU
GRU
GRU
Y_1
Y_2
Y_{n-1}
Y_n
X_1
X_2
X_{n-1}
X_n
Historical data
Decoder

Attention Seq-to-Seq (RNN)

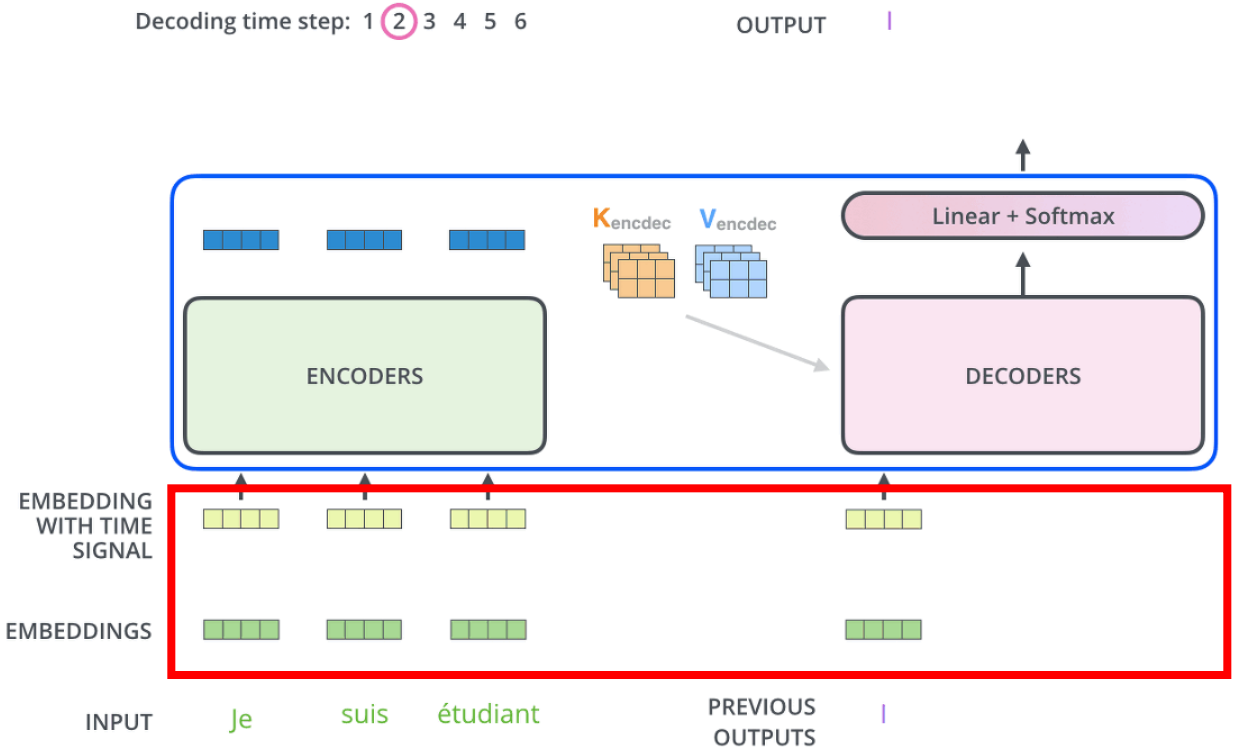


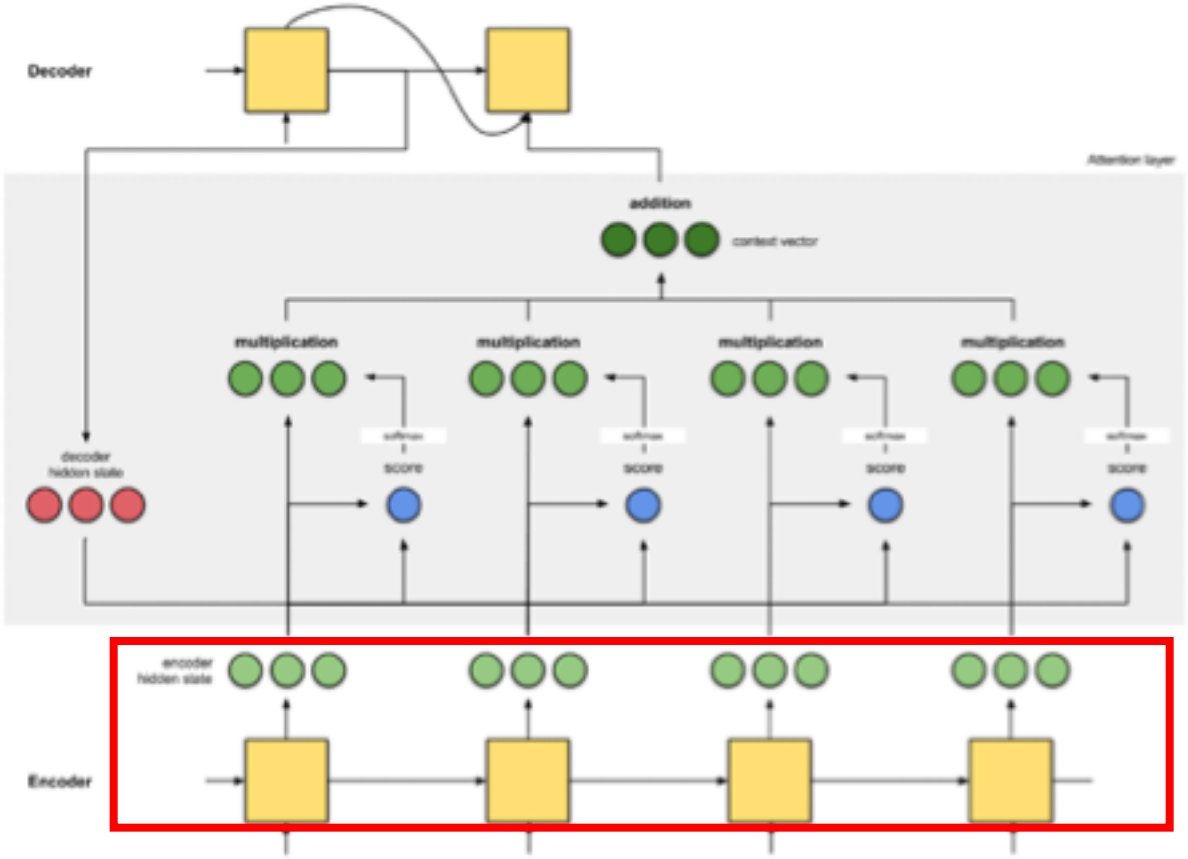

Decoder
Attention layer
addition
context vector
multiplication
multiplication
multiplication
multiplication
softmax
softmax
softmax
softmax
score
score
score
score
decoder
hidden state
encoder
hidden state
Encoder

Attention Seq-to-Seq (Transformer)

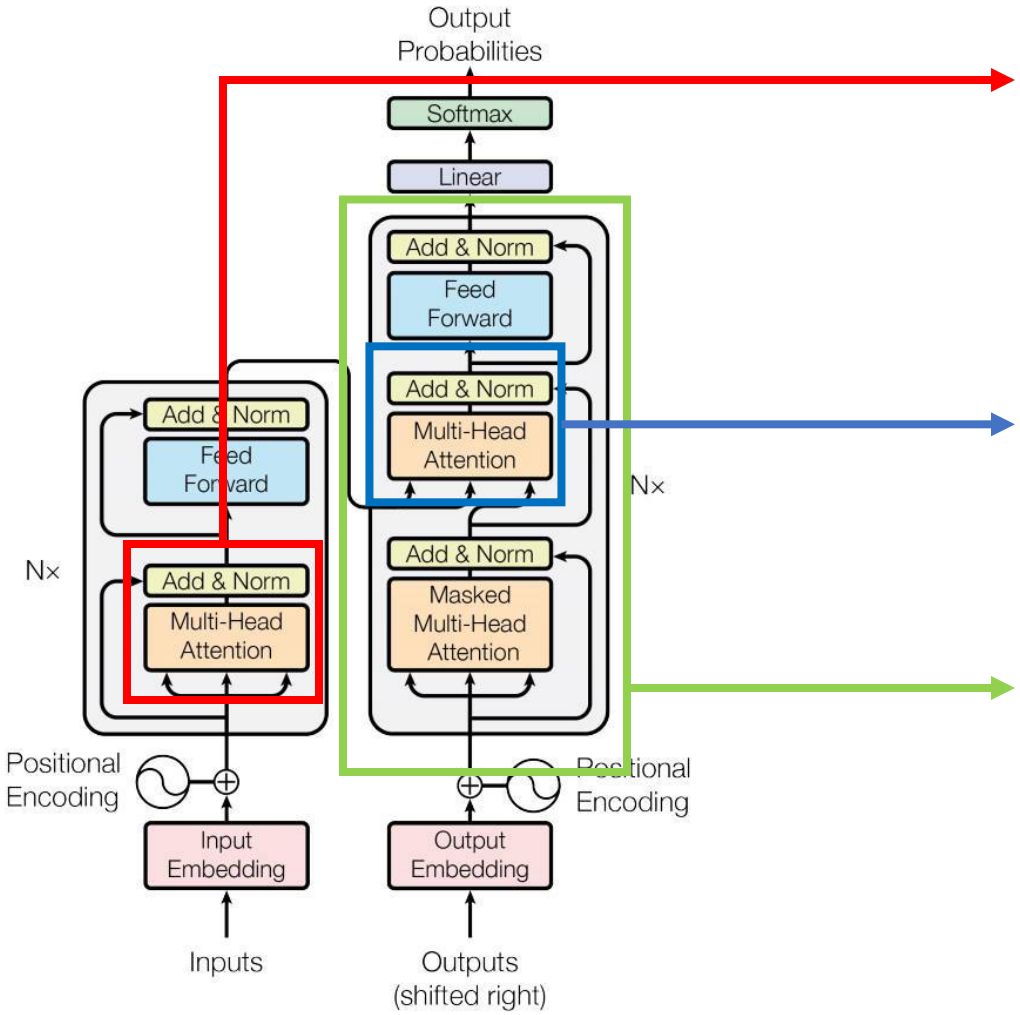


Attention Seq-to-Seq





Transformer in streaming fashion



1. Time-restricted self-attention

2. Triggered attention

3. Joint CTC-triggered attention decoding

Time-restricted self-attention

$$X_0 = \textsc{EncCNN}(X),$$

$$X_E = \textsc{EncSA}(X_0),$$

$$X'_e = X_{e-1} + \mathrm{MHA}_e(X_{e-1}, X_{e-1}, X_{e-1}),$$

$$X_e = X'_e + \mathrm{FF}_e(X'_e),$$

$$\text{with } \mathrm{FF}_e(X'_e) = \mathrm{ReLU}(X'_e W^{\mathrm{ff}}_{e,1} + b^{\mathrm{ff}}_{e,1}) W^{\mathrm{ff}}_{e,2} + b^{\mathrm{ff}}_{e,2},$$

$$\boldsymbol{x}^E_{1:n} = \textsc{EncSA}^{\mathrm{tr}}(\boldsymbol{x}^0_{1:n+\varepsilon^{\mathrm{enc}}}),$$

Self-attention

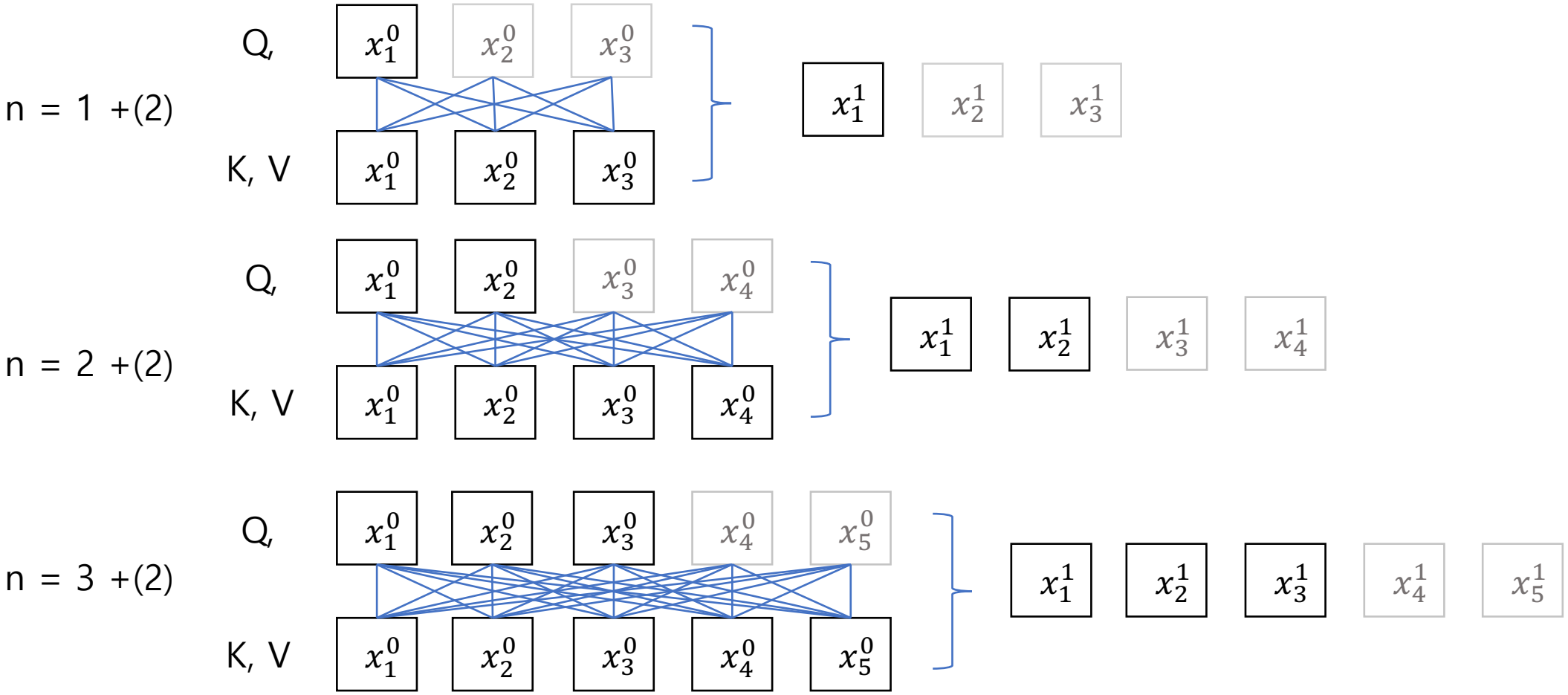
Residual connection

Look-ahead frame

Time-restricted self-attention

$$\boldsymbol{x}_{1:n}^{E} = \mathrm{EncSA}^{\mathrm{tr}}(\boldsymbol{x}_{1:n+\varepsilon^{\mathrm{enc}}}^{0}),$$

If, $\varepsilon^{enc} = 2$, E=1



Triggered attention

$$p_{\text{ta}}(Y|X_E) = \prod_{l=1}^{L} p(y_l|\boldsymbol{y}_{1:l-1}, \boldsymbol{x}^E_{1:\nu_l})$$

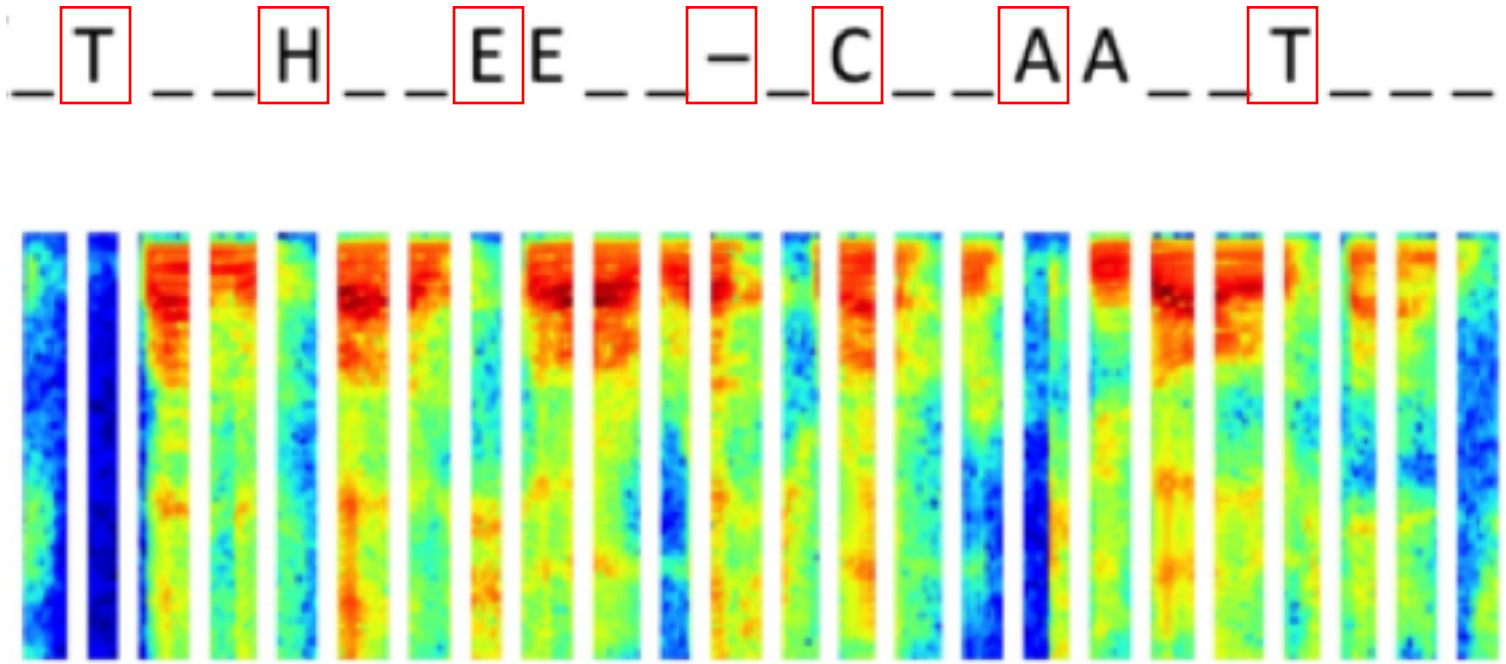
$$p(y_l|\boldsymbol{y}_{1:l-1}, \boldsymbol{x}^E_{1:\nu_l}) = \text{DECTA}(\boldsymbol{x}^E_{1:\nu_l}, \boldsymbol{y}_{1:l-1}),$$

$$\begin{aligned}
Y'_{d,l} &= \boldsymbol{y}^{d-1}_{1:l-1} + \text{MHA}^{\text{self}}_d(\boldsymbol{y}^{d-1}_{1:l-1}, \boldsymbol{y}^{d-1}_{1:l-1}, \boldsymbol{y}^{d-1}_{1:l-1}), \\
Y''_{d,l} &= Y'_{d,l} + \text{MHA}^{\text{dec}}_d(Y'_{d,l}, \boldsymbol{x}^E_{1:\nu_l}, \boldsymbol{x}^E_{1:\nu_l}), \\
\boldsymbol{y}^d_{1:l-1} &= Y''_{d,l} + \text{FF}_d(Y''_{d,l}),
\end{aligned}$$

$$\mathcal{L} = -\gamma \log p_{\text{ctc}} - (1-\gamma) \log p_{\text{ta}},$$

Triggered attention

with $\nu_l = n'_l + \varepsilon^{\text{dec}}$, where n'_l denotes the position of the first occurrence of label y_l in the CTC forced alignment sequence [12, 14],



Triggered attention

$$p_{\text{ta}}(Y|X_E) = \prod_{l=1}^{L} p(y_l|\boldsymbol{y}_{1:l-1}, \boldsymbol{x}^E_{1:\nu_l})$$

$$p(y_l|\boldsymbol{y}_{1:l-1}, \boldsymbol{x}^E_{1:\nu_l}) = \text{DECTA}(\boldsymbol{x}^E_{1:\nu_l}, \boldsymbol{y}_{1:l-1}),$$

$$\begin{aligned}
Y'_{d,l} &= \boldsymbol{y}^{d-1}_{1:l-1} + \text{MHA}^{\text{self}}_d(\boldsymbol{y}^{d-1}_{1:l-1}, \boldsymbol{y}^{d-1}_{1:l-1}, \boldsymbol{y}^{d-1}_{1:l-1}), \\
Y''_{d,l} &= Y'_{d,l} + \text{MHA}^{\text{dec}}_d(Y'_{d,l}, \boldsymbol{x}^E_{1:\nu_l}, \boldsymbol{x}^E_{1:\nu_l}), \\
\boldsymbol{y}^d_{1:l-1} &= Y''_{d,l} + \text{FF}_d(Y''_{d,l}),
\end{aligned}$$

$$\mathcal{L} = -\gamma \log p_{\text{ctc}} - (1-\gamma) \log p_{\text{ta}},$$

Joint CTC-triggered attention decoding

procedure DECODE(X_E, p_{ctc}, λ, α_0, α, β, K, P, θ_1, θ_2)
  $\ell \leftarrow (\langle \text{sos} \rangle,)$
  $\Omega \leftarrow \{\ell\}$, $\Omega_{\text{ta}} \leftarrow \{\ell\}$
  $p_{\text{nb}}(\ell) \leftarrow 0$, $p_{\text{b}}(\ell) \leftarrow 1$
  $p_{\text{ta}}(\ell) \leftarrow 1$
  for $n = 1, \ldots, N$ **do**
    $\Omega_{\text{ctc}}, p_{\text{nb}}, p_{\text{b}} \leftarrow$ CTCPREFIX($p_{\text{ctc}}(n), \Omega, p_{\text{nb}}, p_{\text{b}}$)
    for ℓ **in** Ω_{ctc} **do** ▷ Compute CTC prefix scores
      $p_{\text{prfx}}(\ell) \leftarrow p_{\text{nb}}(\ell) + p_{\text{b}}(\ell)$
      $\widehat{p}_{\text{prfx}}(\ell) \leftarrow \log p_{\text{prfx}}(\ell) + \alpha_0 \log p_{\text{LM}}(\ell) + \beta|\ell|$
    $\widehat{\Omega} \leftarrow$ PRUNE($\Omega_{\text{ctc}}, \widehat{p}_{\text{prfx}}, K, \theta_1$)
    for ℓ **in** $\widehat{\Omega}$ **do** ▷ Delete old prefixes in Ω_{ta}
      if ℓ **in** Ω_{ta} **and** DCOND($\ell, \widehat{\Omega}, p_{\text{ctc}}$) **then**
        delete ℓ in Ω_{ta}
    for ℓ **in** $\widehat{\Omega}$ **do** ▷ Compute transformer scores
      if ℓ **not in** Ω_{ta} **and** ACOND($\ell, \widehat{\Omega}, p_{\text{ctc}}$) **then**
        $p_{\text{ta}}(\ell) \leftarrow$ DECTA($\boldsymbol{x}^E_{1:n+c^{\text{dec}}}, \ell$)
        add ℓ to Ω_{ta}
    for ℓ **in** $\widehat{\Omega}$ **do** ▷ Compute joint scores
      $\widehat{\ell} \leftarrow \ell$ **if** ℓ **in** Ω_{ta} **else** $\ell_{:-1}$
      $p \leftarrow \lambda \log p_{\text{prfx}}(\ell) + (1 - \lambda) \log p_{\text{ta}}(\widehat{\ell})$
      $p_{\text{joint}}(\ell) \leftarrow p + \alpha \log p_{\text{LM}}(\ell) + \beta|\ell|$
    $\Omega \leftarrow$ MAX($\widehat{\Omega}, p_{\text{joint}}, P$)
    $\widehat{\Omega} \leftarrow$ PRUNE($\widehat{\Omega}, \widehat{p}_{\text{prfx}}, P, \theta_2$)
    $\Omega \leftarrow \Omega + \widehat{\Omega}$
    remove from Ω_{ta} prefixes rejected due to pruning
  return MAX($\widehat{\Omega}, p_{\text{joint}}, 1$)

procedure PRUNE($\Omega_{\text{in}}, \widehat{p}_{\text{prfx}}, L, \theta$)
  $\Omega_{\text{in}} \leftarrow$ MAX($\Omega_{\text{in}}, \widehat{p}_{\text{prfx}}, L$), $\Omega_{\text{out}} \leftarrow \{\}$
  for ℓ **in** Ω_{in} **do**
    if $\max(\widehat{p}_{\text{prfx}}) - \theta < \widehat{p}_{\text{prfx}}(\ell)$ **then**
      add ℓ to Ω_{out}
  return Ω_{out}

procedure DCOND($\ell, \widehat{\Omega}, p_{\text{ctc}}$)
  $c \leftarrow \ell_{-1}$ ▷ last element of ℓ
  $n_{\text{prev}} \leftarrow$ time index of ℓ when added to $\widehat{\Omega}$
  if $|\ell| > 1$ **and** $n - n_{\text{prev}} > 2$ **and** $p_{\text{ctc}}(n, c) > 0.01 > \max(p_{\text{ctc}}(n_{\text{prev}}, c), p_{\text{ctc}}(n_{\text{prev}} + 1, c))$ **then**
    return true
  else
    return false

procedure ACOND($\ell, \widehat{\Omega}, p_{\text{ctc}}$)
  $c \leftarrow \ell_{-1}$ ▷ last element of ℓ
  if $p_{\text{ctc}}(n, c) > \max(p_{\text{ctc}}(n + 1, c), p_{\text{ctc}}(n + 2, c))$ **or** any($|\widetilde{\ell}| > |\ell| + 1$ **and** $\widetilde{\ell}$ starts with ℓ **for** $\widetilde{\ell}$ **in** $\widehat{\Omega}$) **then**
    return true
  else
    return false

Experiment

Dataset : LibriSpeech

Setting : transformer

small : d_model = 256, d_ff = 2048, d_h = 4, E = 12, D = 6
large : d_model = 512, d_h = 8

: joint CTC-TA decoding

CTC weight $\lambda = 0.5$, CTC LM weight $\alpha_0 = 0.7$, LM weight $\alpha = 0.5$,
pruning beam width $\theta_1 = 16.0$, pruning beam width $\theta_2 = 6.0$,
insertion bonus $\beta = 2.0$, pruning size $K = 300$, and pruning size $P = 30$.

Result

Theoretical delay

Encoder CNN : 30ms

Time restricted self-attention : $E \times \varepsilon^{enc} \times 40ms$

Triggered attention : $\varepsilon^{dec} \times 40ms$

WER : 2.8% / 7.3% → ε^{enc}=4, ε^{dec}=6 → 2190ms

Result

	CTC-attention dec.				CTC beam search				Att. beam search			
	clean		other		clean		other		clean		other	
System	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test
baseline	4.7	4.9	13.0	12.9	6.1	6.1	15.7	15.9	6.0	7.8	14.5	14.9
+RNN-LM	2.9	3.1	8.0	8.4	3.1	3.4	9.3	9.6	4.7	7.2	10.7	11.5
+SpecAug.	2.4	2.8	6.4	6.7	2.9	3.2	7.6	7.9	4.2	5.2	8.3	8.6
+large	2.4	2.7	6.0	6.1	2.5	2.8	6.9	7.0	4.1	5.0	7.9	8.0

	CTC beam search				TA: $\varepsilon^{\text{dec}} = 6$				TA: $\varepsilon^{\text{dec}} = 12$				TA: $\varepsilon^{\text{dec}} = 18$			
	clean		other		clean		other		clean		other		clean		other	
ε^{enc}	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test
1	3.0	3.3	8.4	8.6	2.9	3.2	8.1	8.2	2.8	3.1	7.5	8.1	2.8	3.0	7.5	7.8
2	2.9	3.1	8.0	8.2	2.8	2.9	7.4	7.8	2.7	2.9	7.2	7.6	-	-	-	-
4	2.8	2.9	7.8	7.9	2.6	2.8	7.2	7.3	-	-	-	-	-	-	-	-
∞	2.5	2.8	6.9	7.0	2.5	2.7	6.3	6.5	2.5	2.7	6.3	6.4	2.4	2.6	6.1	6.3