# Deep Learning for Computer Vision

Dr. Konda Reddy Mopuri
Mehta Family School of Data Science and Artificial Intelligence
IIT Guwahati
Aug-Dec 2022

# So far in the class..

- Feedforward NNs

# So far in the class..

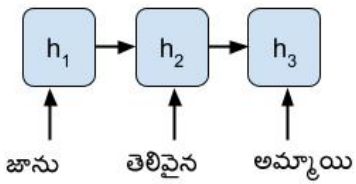
- Feedforward NNs
- RNNs

# Sequence-to-Sequence tasks using RNNs

Input sequence: $x_1, x_2, \ldots x_T$

Input sequence: $y_1, y_2, \ldots y_{T'}$

Encoder: $h_t = E(x_t, h_{t-1})$




Sequence to sequence learning by Sutskever et al. NeurIPS 2014

# Sequence-to-Sequence tasks using RNNs

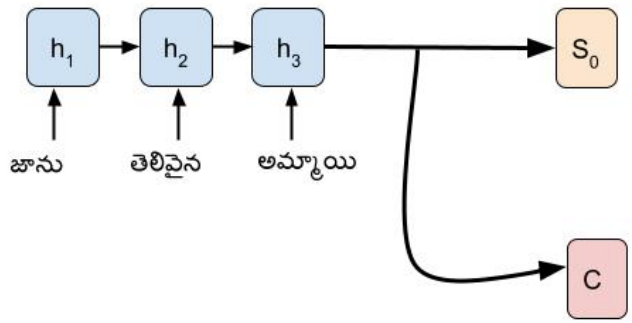
Input sequence: $x_1, x_2, \ldots x_T$

Input sequence: $y_1, y_2, \ldots y_{T'}$

Last hidden state $h_T \rightarrow$ Initial state of the Decoder
$S_0$ and the context information C
E.g. $S_0 \leftarrow h_T$ + dense layers, and $C \leftarrow h_T$

Decoder: $s_t = D(y_{t-1}, s_{t-1}, C)$

Encoder: $h_t = E(x_t, h_{t-1})$




Sequence to sequence learning by Sutskever et al. NeurIPS 2014

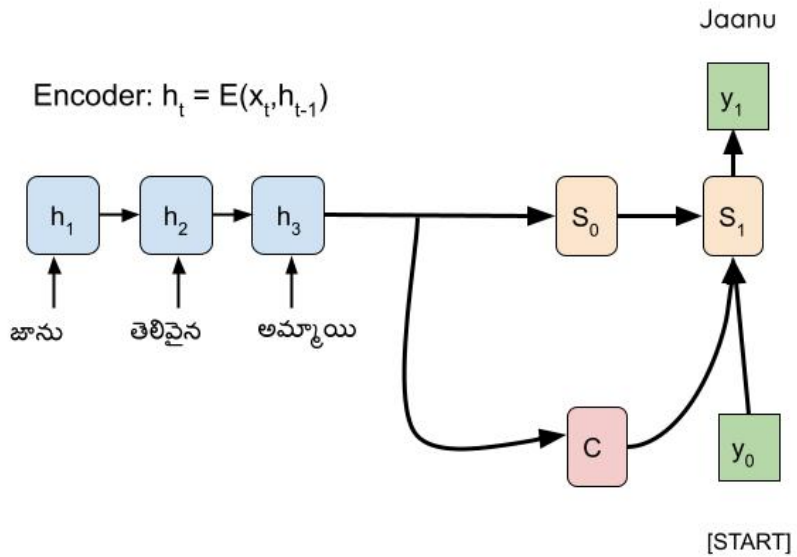
# Sequence-to-Sequence tasks using RNNs

Input sequence: $x_1, x_2, \ldots x_T$

Input sequence: $y_1, y_2, \ldots y_{T'}$

Last hidden state $h_T \rightarrow$ Initial state of the Decoder $S_0$ and the context information C
E.g. $S_0 \leftarrow h_T$ + dense layers, and $C \leftarrow h_T$

Decoder: $s_t = D(y_{t-1}, s_{t-1}, C)$




Sequence to sequence learning by Sutskever et al. NeurIPS 2014

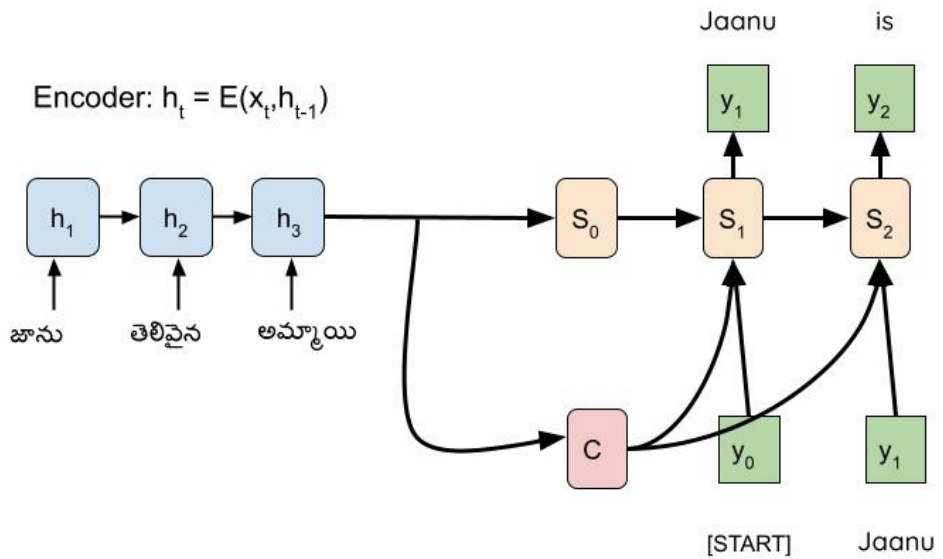
# Sequence-to-Sequence tasks using RNNs

Input sequence: $x_1, x_2, \ldots x_T$

Input sequence: $y_1, y_2, \ldots y_{T'}$

Last hidden state $h_T \rightarrow$ Initial state of the Decoder $S_0$ and the context information C
E.g. $S_0 \leftarrow h_T$ + dense layers, and $C \leftarrow h_T$

Decoder: $s_t = D(y_{t-1}, s_{t-1}, C)$




Sequence to sequence learning by Sutskever et al. NeurIPS 2014

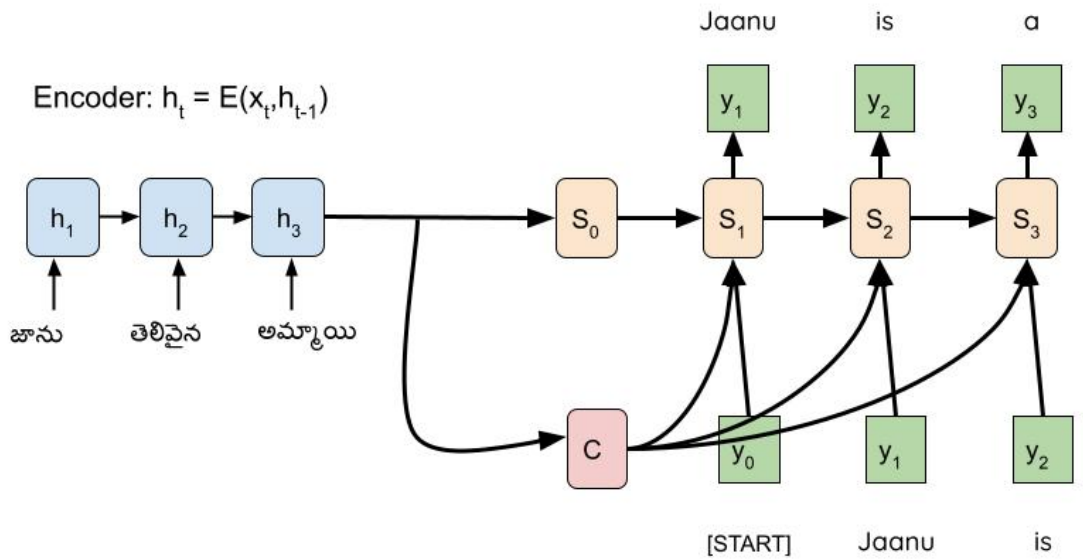
# Sequence-to-Sequence tasks using RNNs

Input sequence: $x_1, x_2, \ldots x_T$

Input sequence: $y_1, y_2, \ldots y_{T'}$

Last hidden state $h_T \rightarrow$ Initial state of the Decoder $S_0$ and the context information C
E.g. $S_0 \leftarrow h_T$ + dense layers, and $C \leftarrow h_T$

Decoder: $s_t = D(y_{t-1}, s_{t-1}, C)$




Sequence to sequence learning by Sutskever et al. NeurIPS 2014

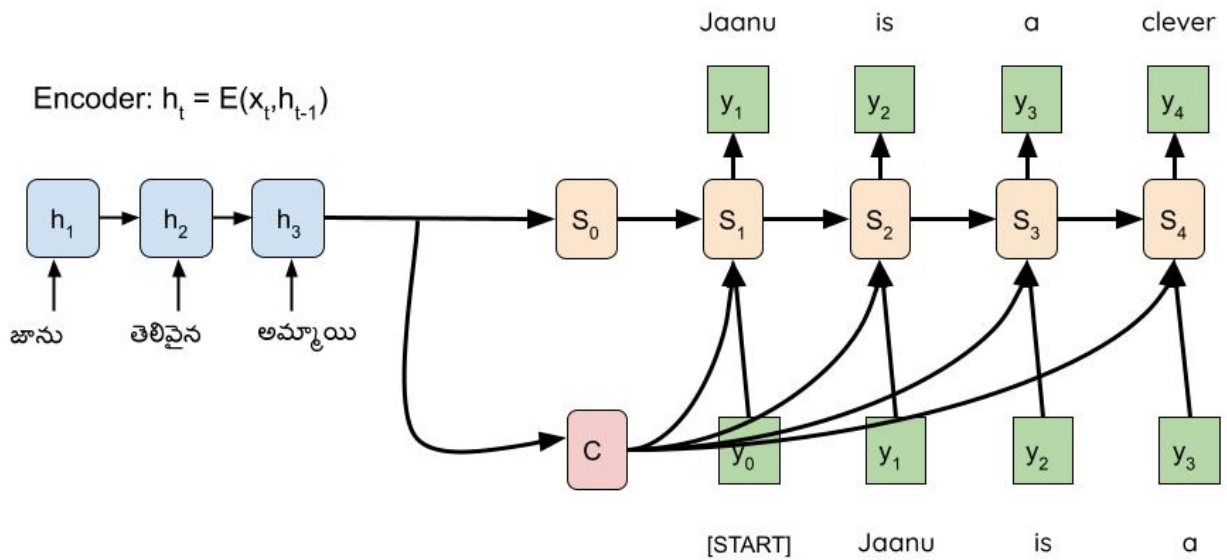
# Sequence-to-Sequence tasks using RNNs

Input sequence: $x_1, x_2, \ldots x_T$

Input sequence: $y_1, y_2, \ldots y_{T'}$

Last hidden state $h_T \rightarrow$ Initial state of the Decoder $S_0$ and the context information C
E.g. $S_0 \leftarrow h_T$ + dense layers, and $C \leftarrow h_T$

Decoder: $s_t = D(y_{t-1}, s_{t-1}, C)$




Sequence to sequence learning by Sutskever et al. NeurIPS 2014

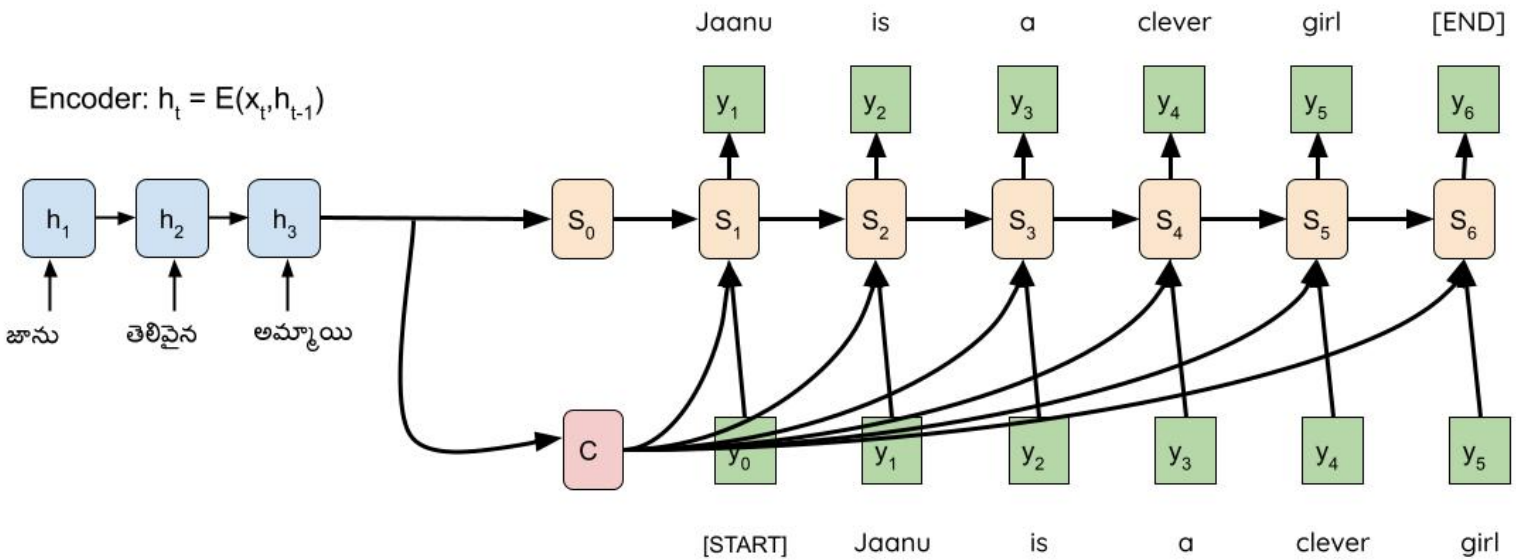
# Sequence-to-Sequence tasks using RNNs

Input sequence: $x_1, x_2, \ldots x_T$

Input sequence: $y_1, y_2, \ldots y_{T'}$

Last hidden state $h_T \rightarrow$ Initial state of the Decoder $S_0$ and the context information C
E.g. $S_0 \leftarrow h_T$ + dense layers, and $C \leftarrow h_T$

Decoder: $s_t = D(y_{t-1}, s_{t-1}, C)$




Sequence to sequence learning by Sutskever et al. NeurIPS 2014

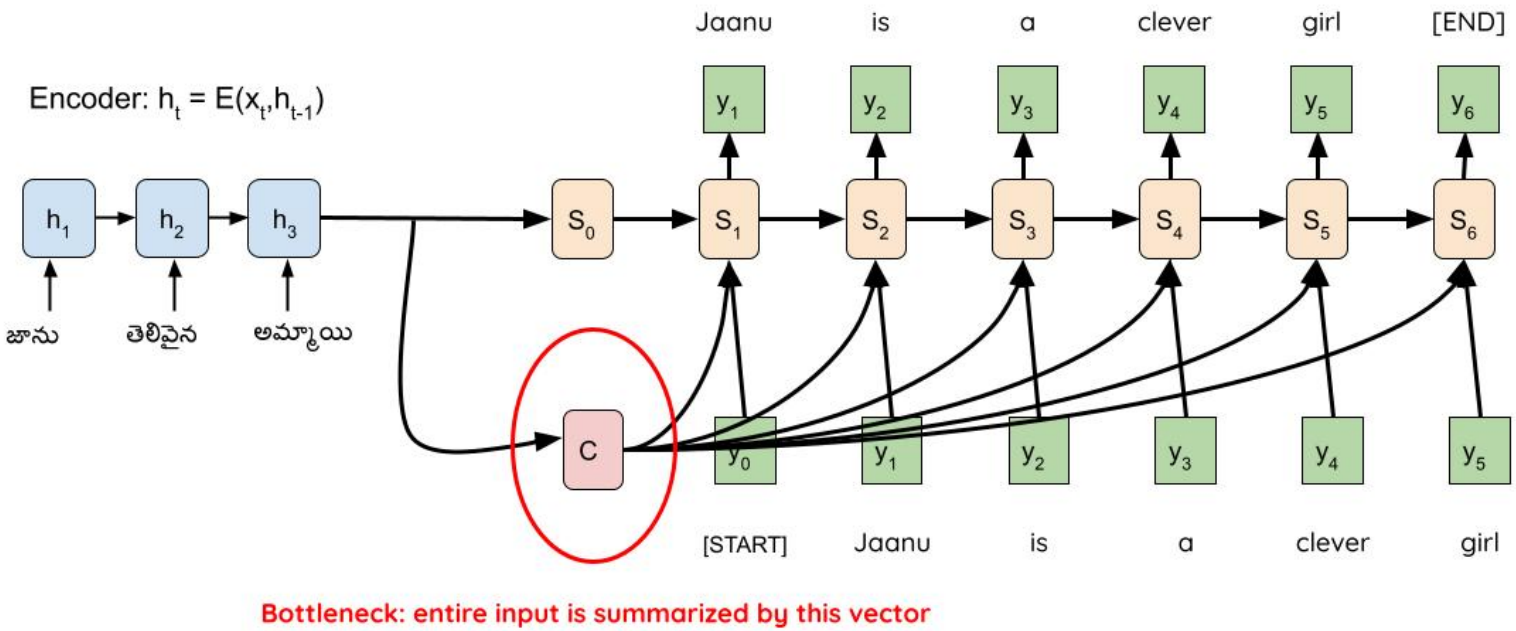
# Sequence-to-Sequence tasks using RNNs

Input sequence: $x_1, x_2, \ldots x_T$

Input sequence: $y_1, y_2, \ldots y_{T'}$

Last hidden state $h_T \rightarrow$ Initial state of the Decoder $S_0$ and the context information C
E.g. $S_0 \leftarrow h_T$ + dense layers, and $C \leftarrow h_T$

Decoder: $s_t = D(y_{t-1}, s_{t-1}, C)$




Sequence to sequence learning by Sutskever et al. NeurIPS 2014

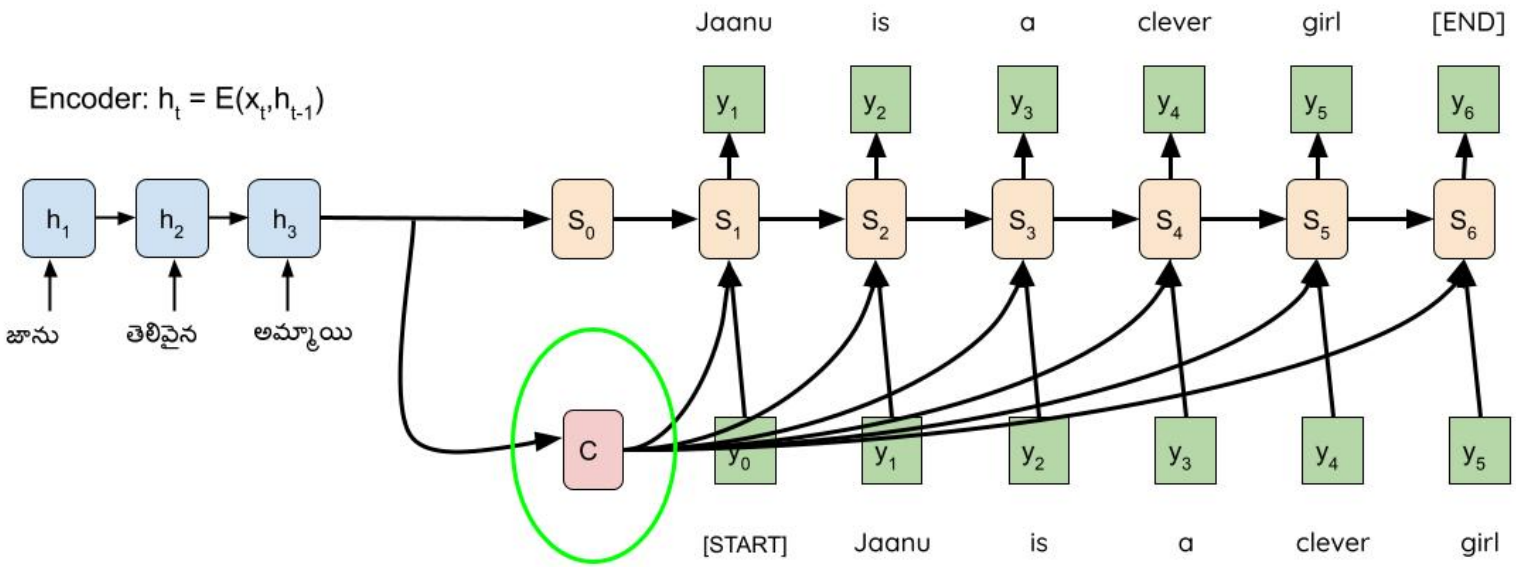
# Sequence-to-Sequence tasks using RNNs

Input sequence: $x_1, x_2, \ldots x_T$

Input sequence: $y_1, y_2, \ldots y_{T'}$

Last hidden state $h_T \rightarrow$ Initial state of the Decoder $S_0$ and the context information C
E.g. $S_0 \leftarrow h_T$ + dense layers, and $C \leftarrow h_T$

Decoder: $s_t = D(y_{t-1}, s_{t-1}, C)$




Solution: use different context at each time step!

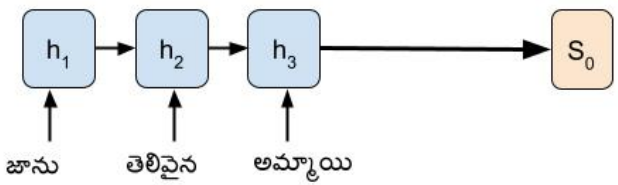
Sequence to sequence learning by Sutskever et al. NeurIPS 2014

# Sequence-to-Sequence tasks using RNNs and Attention

Input sequence: $x_1, x_2, \ldots x_T$

Input sequence: $y_1, y_2, \ldots y_{T'}$

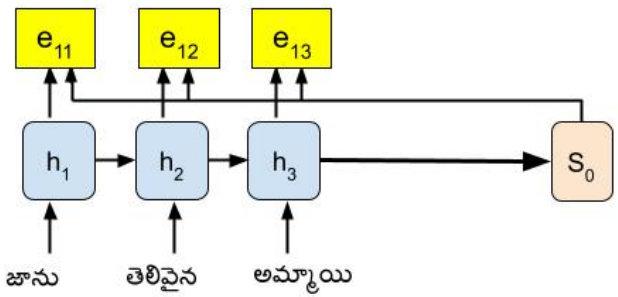
Encoder: $h_t = E(x_t, h_{t-1})$



# Sequence-to-Sequence tasks using RNNs and Attention

Compute the alignment scores

$e_{t,i} = f_{att}(s_{t-1}, h_i)$ $f_{att}$ - couple of dense layers

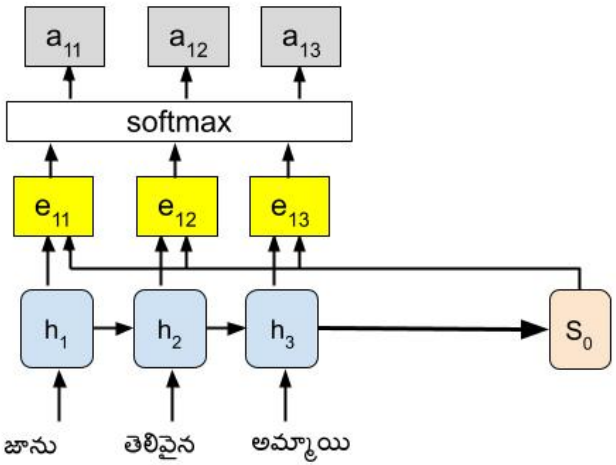



Neural Machine Translation with aligning by Bahdanau et al. ICLR 2015

# Sequence-to-Sequence tasks using RNNs and Attention

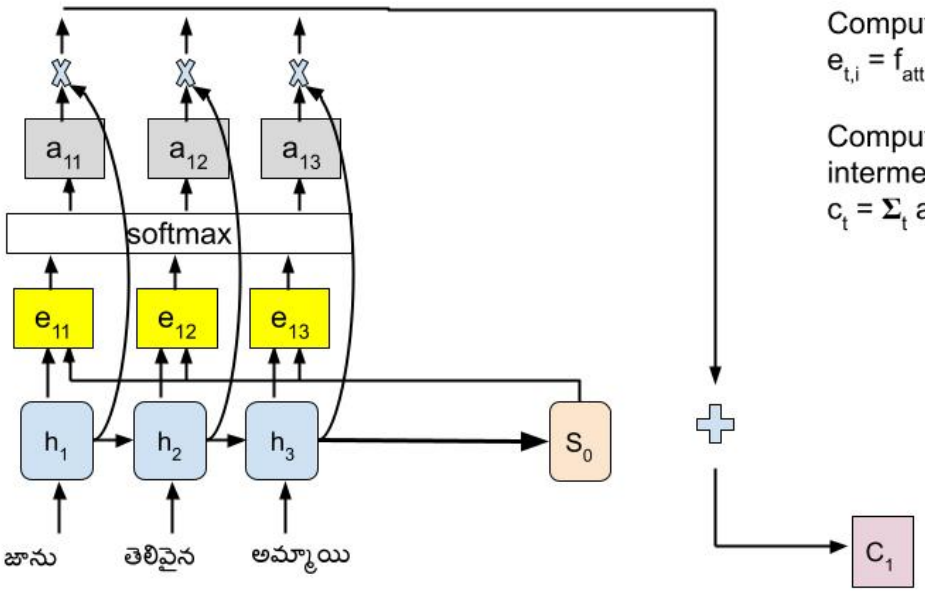
Compute the alignment scores

$e_{t,i} = f_{att}(s_{t-1}, h_i)$ $f_{att}$ - couple of dense layers




Neural Machine Translation with aligning by Bahdanau et al. ICLR 2015

# Sequence-to-Sequence tasks using RNNs and Attention




Compute the alignment scores

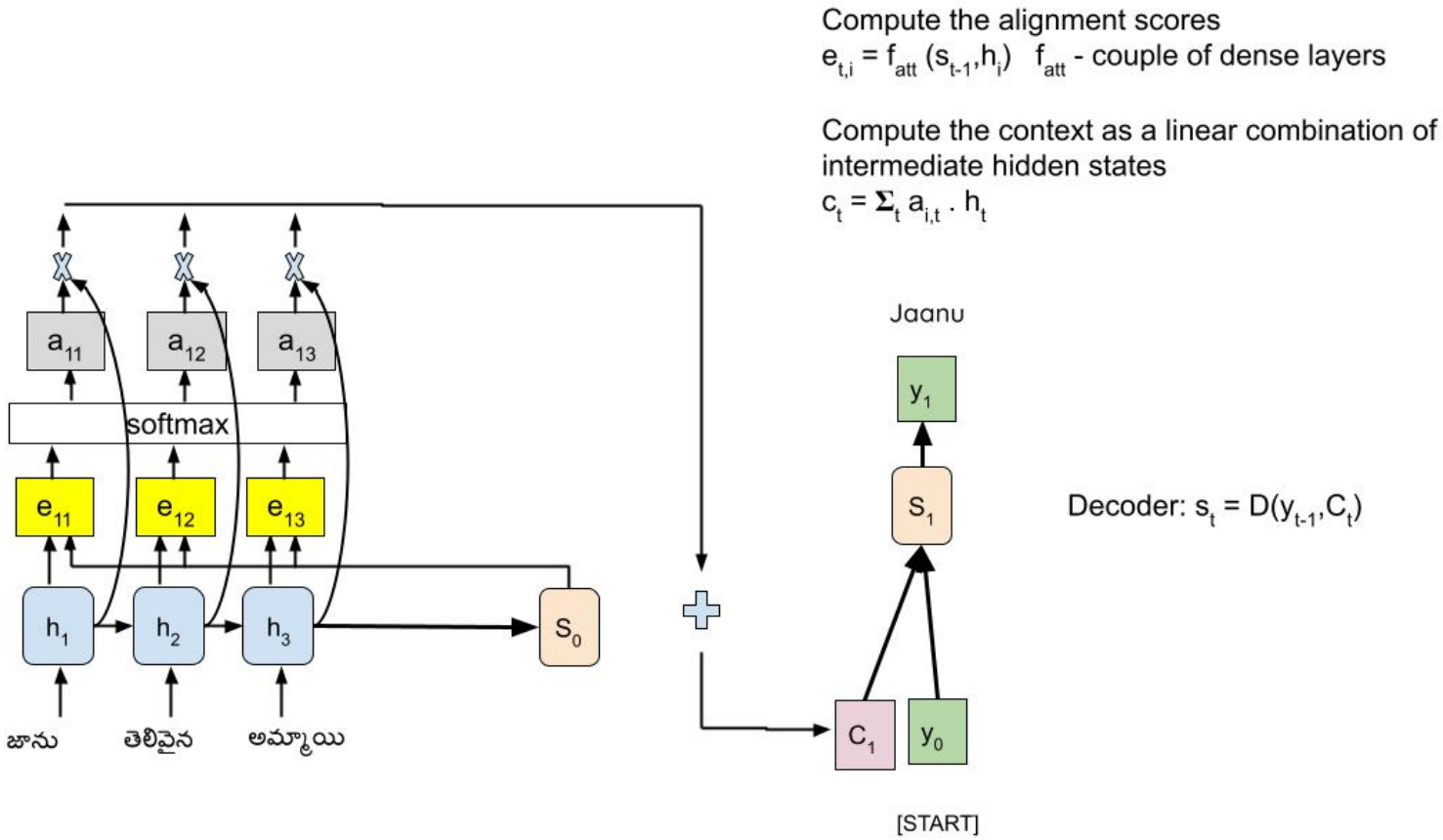
$e_{t,i} = f_{att}(s_{t-1}, h_i)$ $f_{att}$ - couple of dense layers

Compute the context as a linear combination of intermediate hidden states

$c_t = \Sigma_t a_{i,t} . h_t$

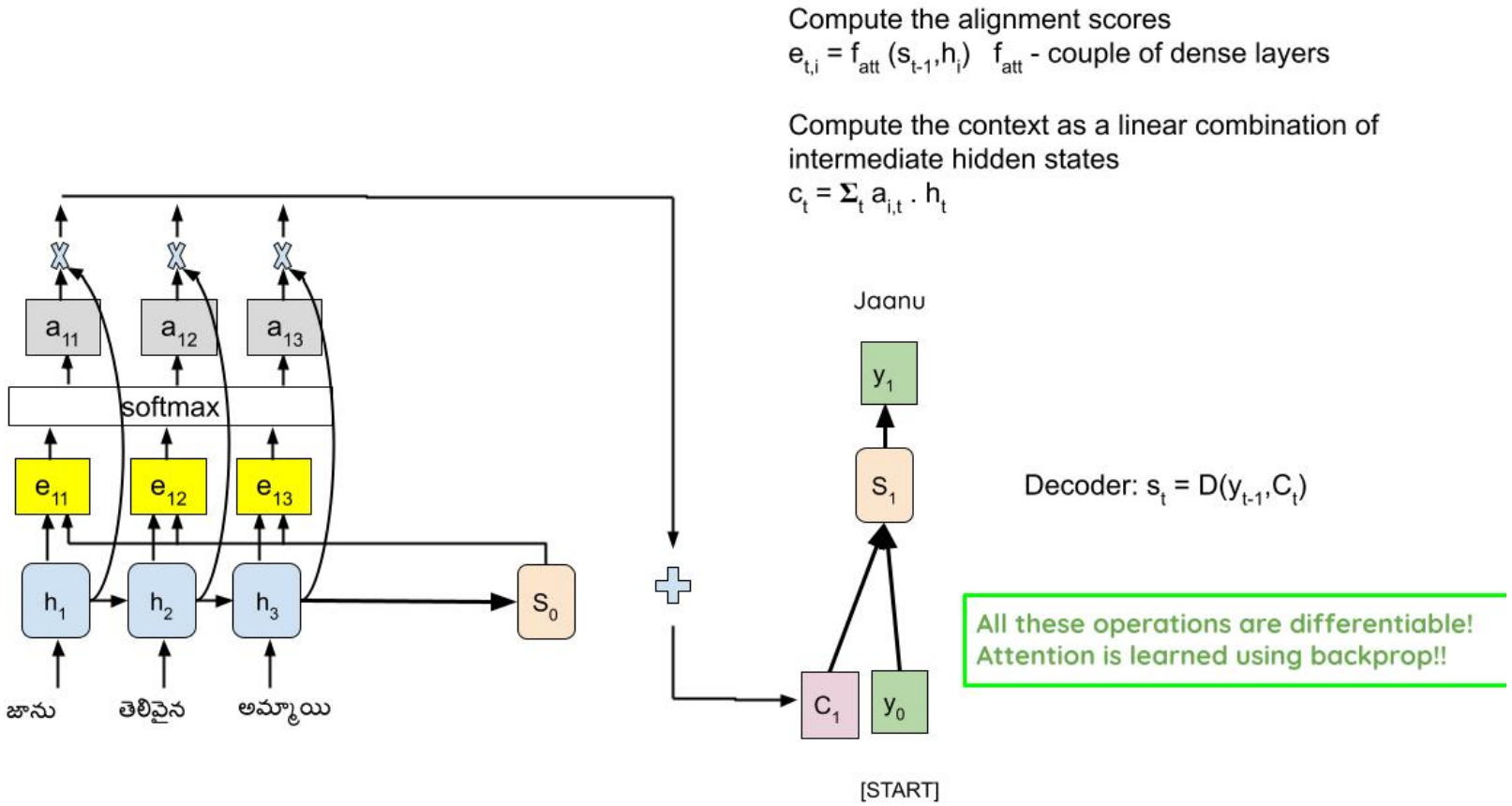
Neural Machine Translation with aligning by Bahdanau et al. ICLR 2015

# Sequence-to-Sequence tasks using RNNs and Attention



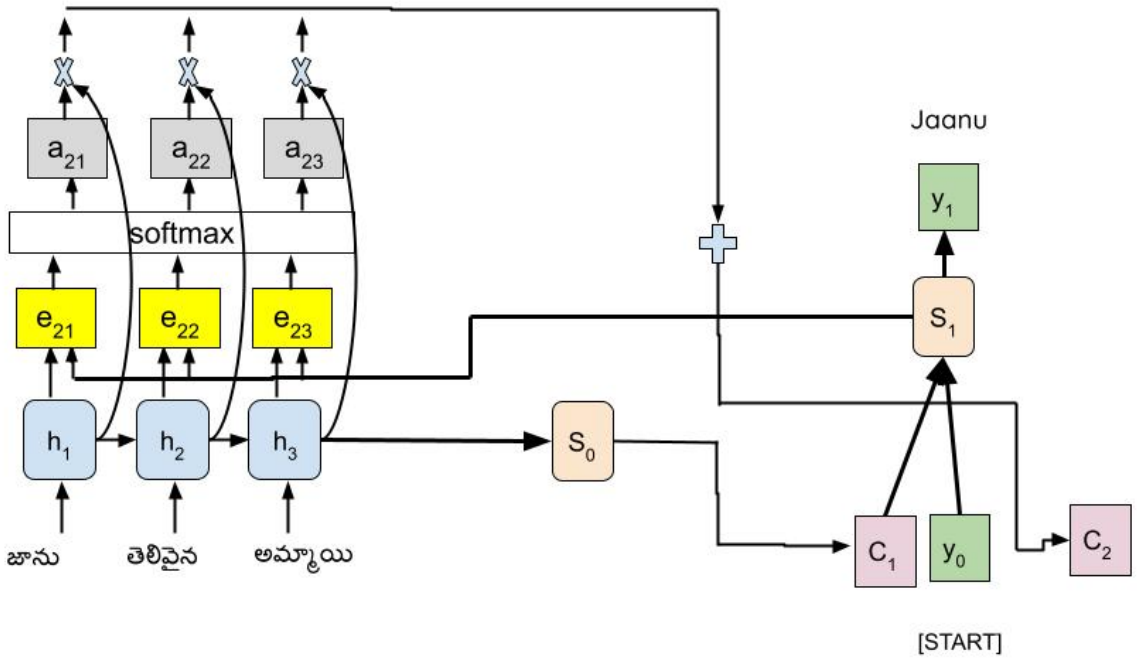

Neural Machine Translation with aligning by Bahdanau et al. ICLR 2015

# Sequence-to-Sequence tasks using RNNs and Attention



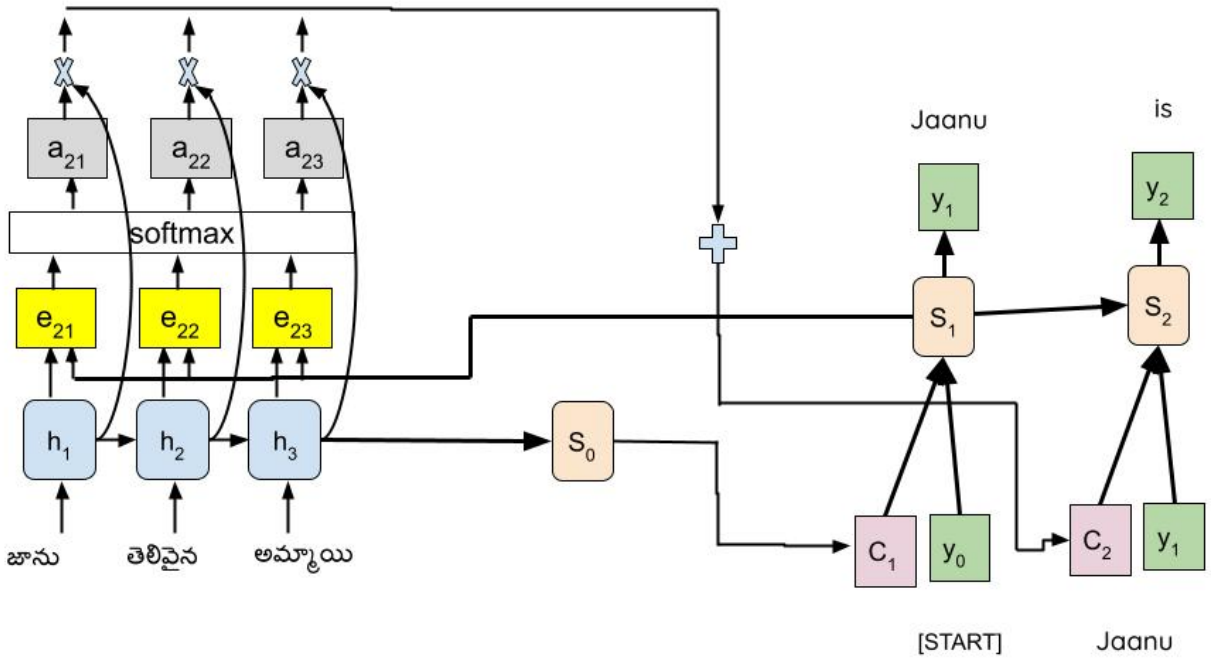

Neural Machine Translation with aligning by Bahdanau et al. ICLR 2015

# Sequence-to-Sequence tasks using RNNs and Attention



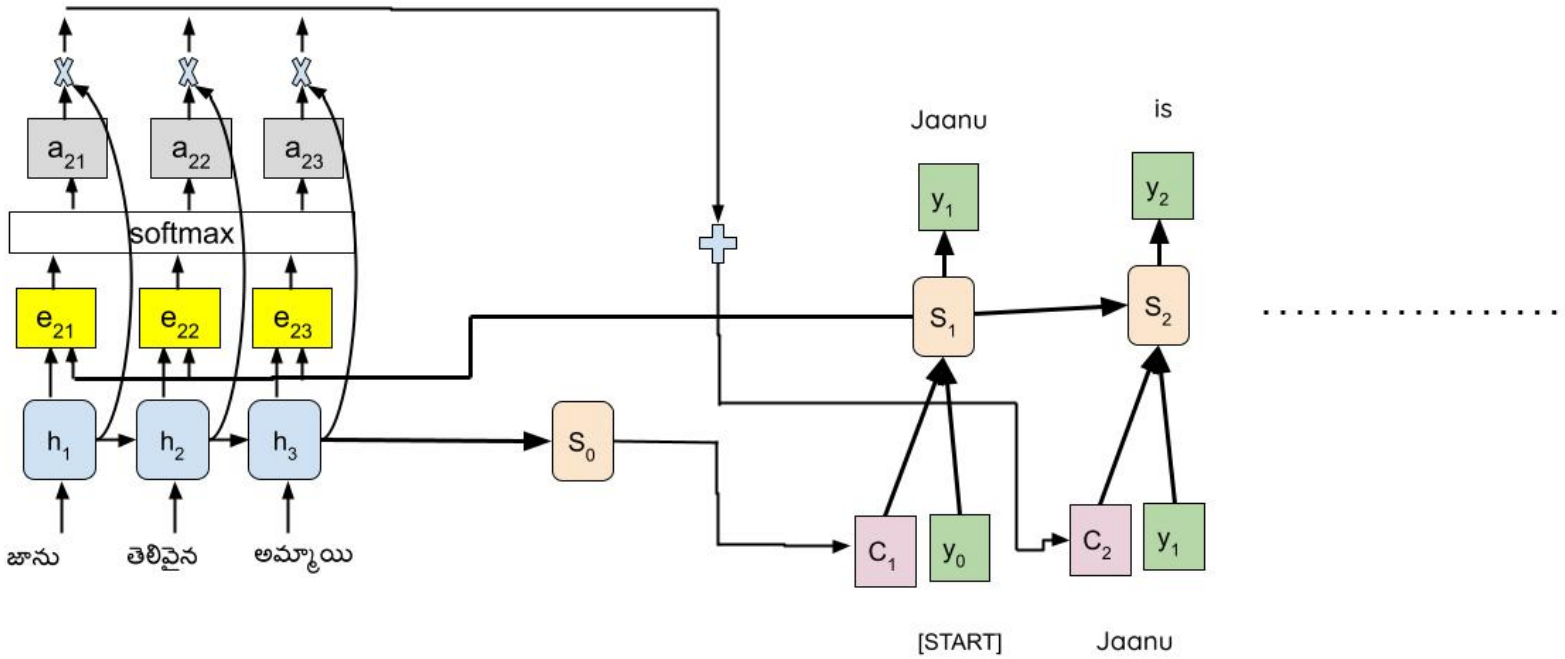

Neural Machine Translation with aligning by Bahdanau et al. ICLR 2015

# Sequence-to-Sequence tasks using RNNs and Attention




Neural Machine Translation with aligning by Bahdanau et al. ICLR 2015

# Sequence-to-Sequence tasks using RNNs and Attention




Neural Machine Translation with aligning by Bahdanau et al. ICLR 2015

# Sequence-to-Sequence tasks using RNNs and Attention

- Employs a different context at each time step of decoding

Neural Machine Translation with aligning by Bahdanau et al. ICLR 2015

# Sequence-to-Sequence tasks using RNNs and Attention

- Employs a different context at each time step of decoding
- No more bottleneck-ing of the input

Neural Machine Translation with aligning by Bahdanau et al. ICLR 2015

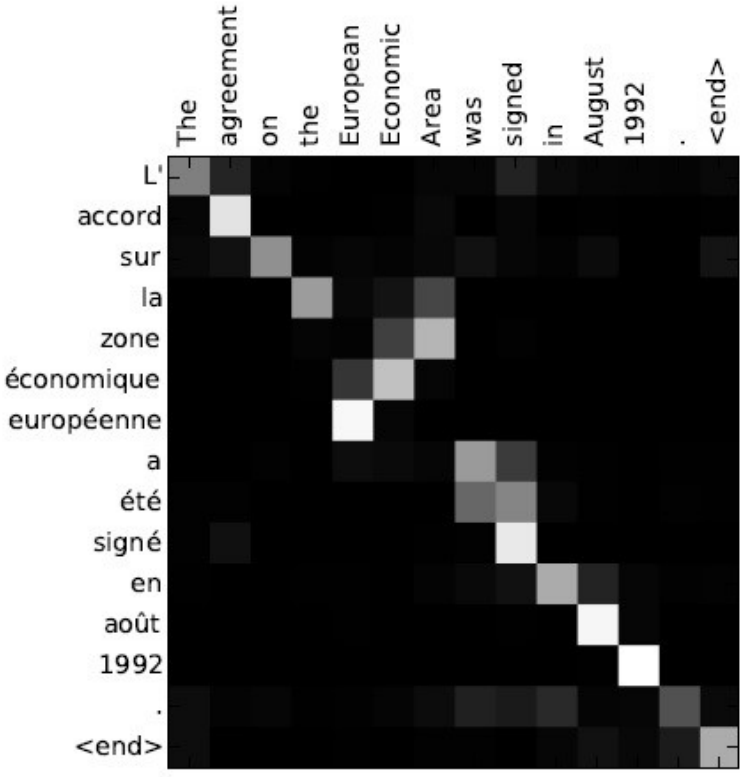
# Sequence-to-Sequence tasks using RNNs and Attention

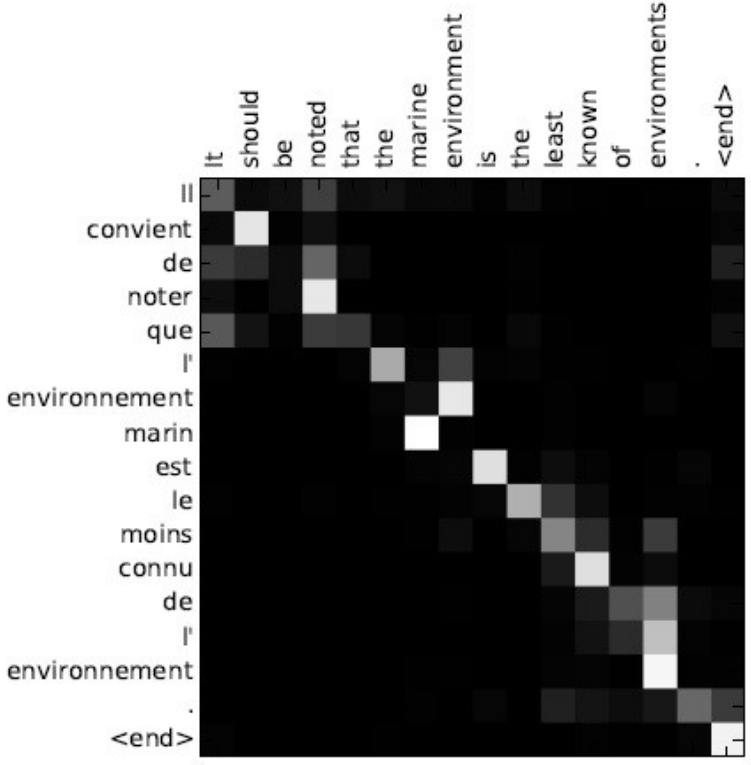
- Employs a different context at each time step of decoding
- No more bottleneck-ing of the input
- Decoder can 'attend' to different portions of the input at each time step

---

Neural Machine Translation with aligning by Bahdanau et al. ICLR 2015

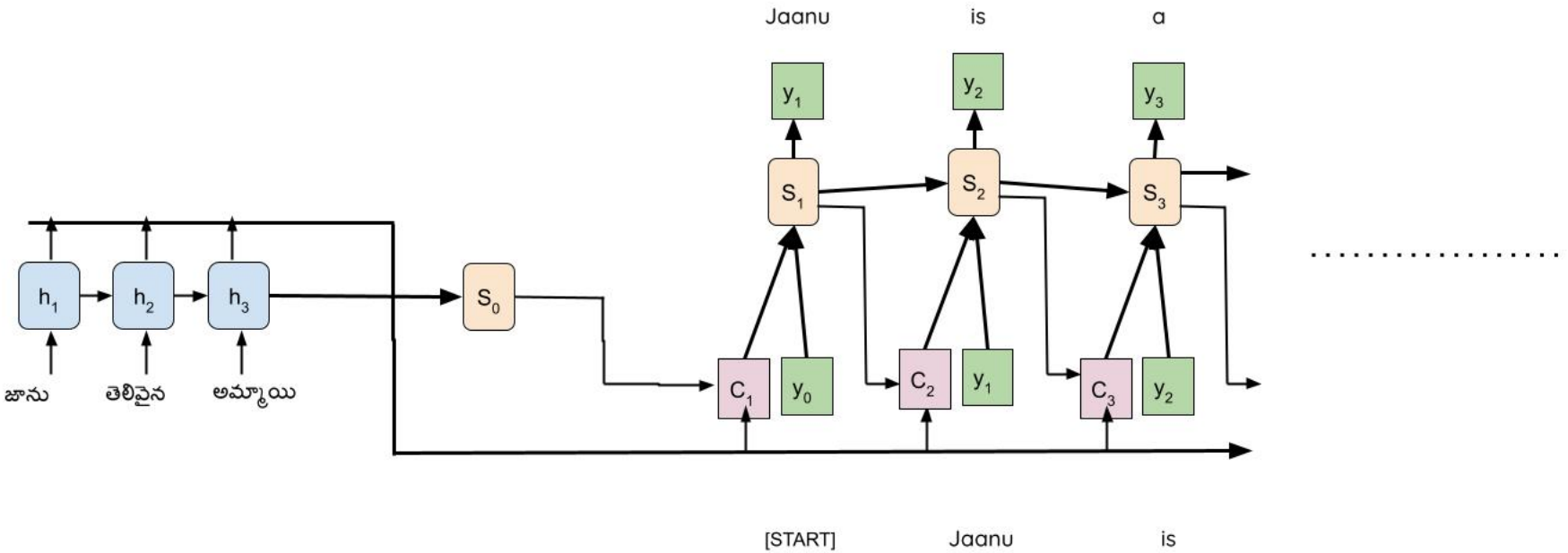
# Sequence-to-Sequence tasks using RNNs and Attention



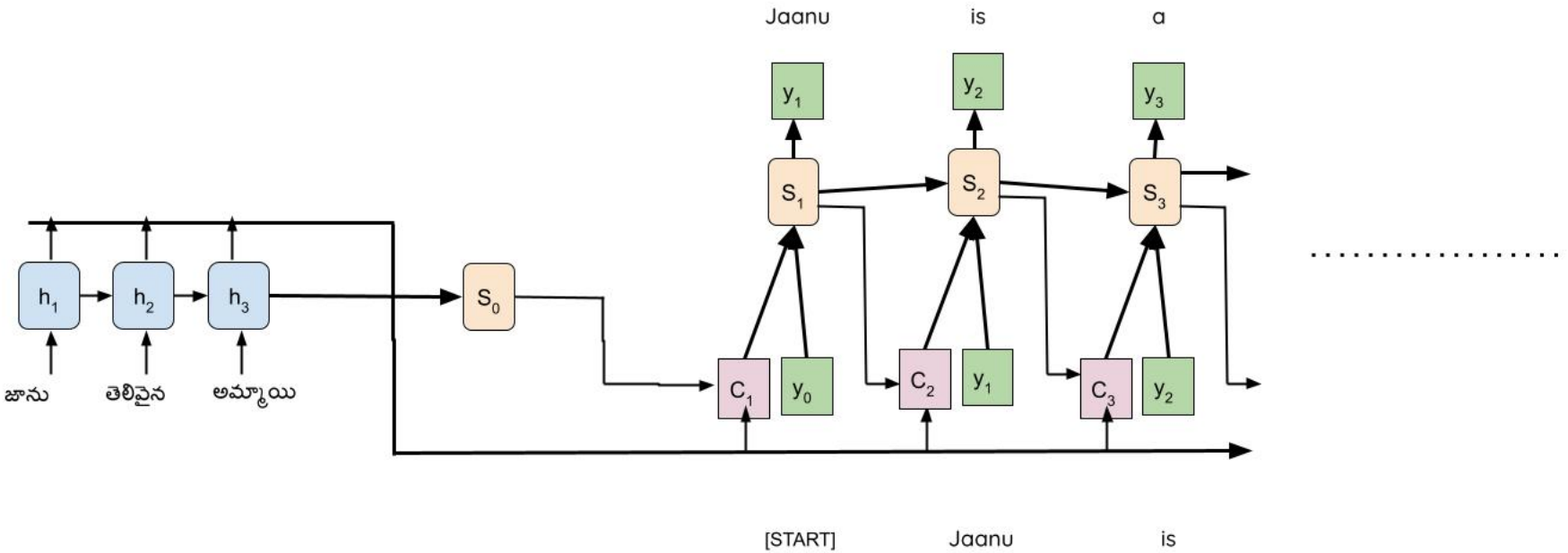

Neural Machine Translation with aligning by Bahdanau et al. ICLR 2015

# Sequence-to-Sequence tasks using RNNs and Attention




- Decoder doesn't consider the $h_i$ to be an ordered set

# Sequence-to-Sequence tasks using RNNs and Attention



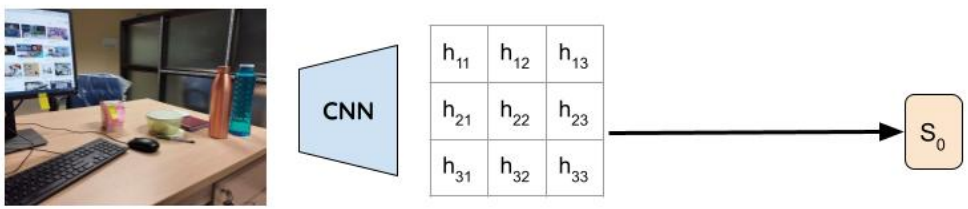

- Decoder doesn't consider the $h_i$ to be an ordered set
- This architecture can be exploited to process a set of inputs $h_i$

# Image captioning using RNNs and Attention


| | | |
|---|---|---|
| $h_{11}$ | $h_{12}$ | $h_{13}$ |
| $h_{21}$ | $h_{22}$ | $h_{23}$ |
| $h_{31}$ | $h_{32}$ | $h_{33}$ |

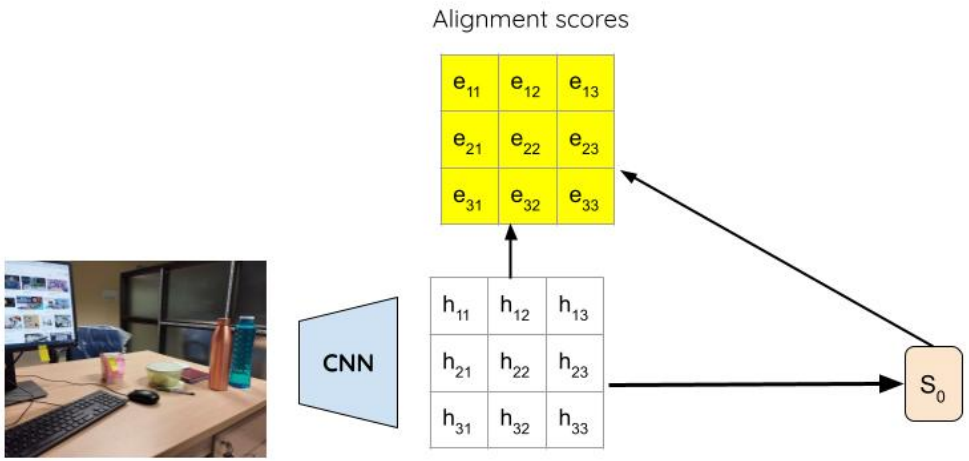
Show Attend and Tell by Xu et al. 2015

# Image captioning using RNNs and Attention



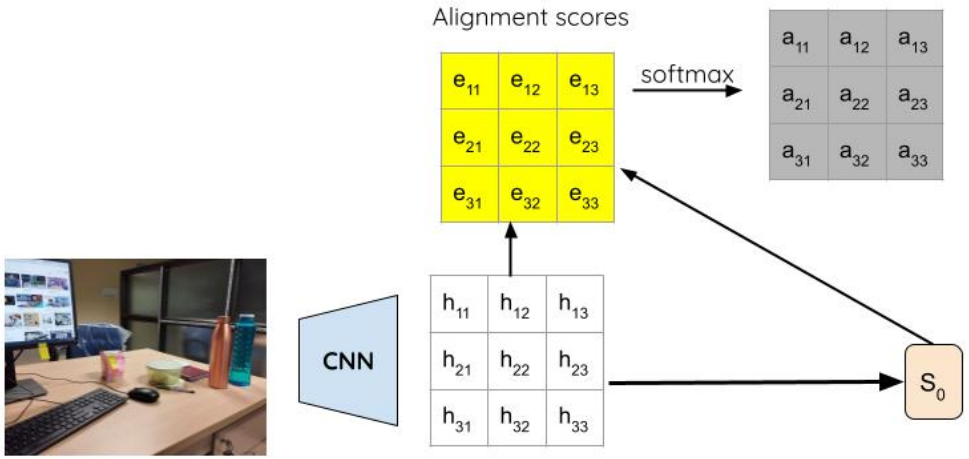

Show Attend and Tell by Xu et al. 2015

# Image captioning using RNNs and Attention



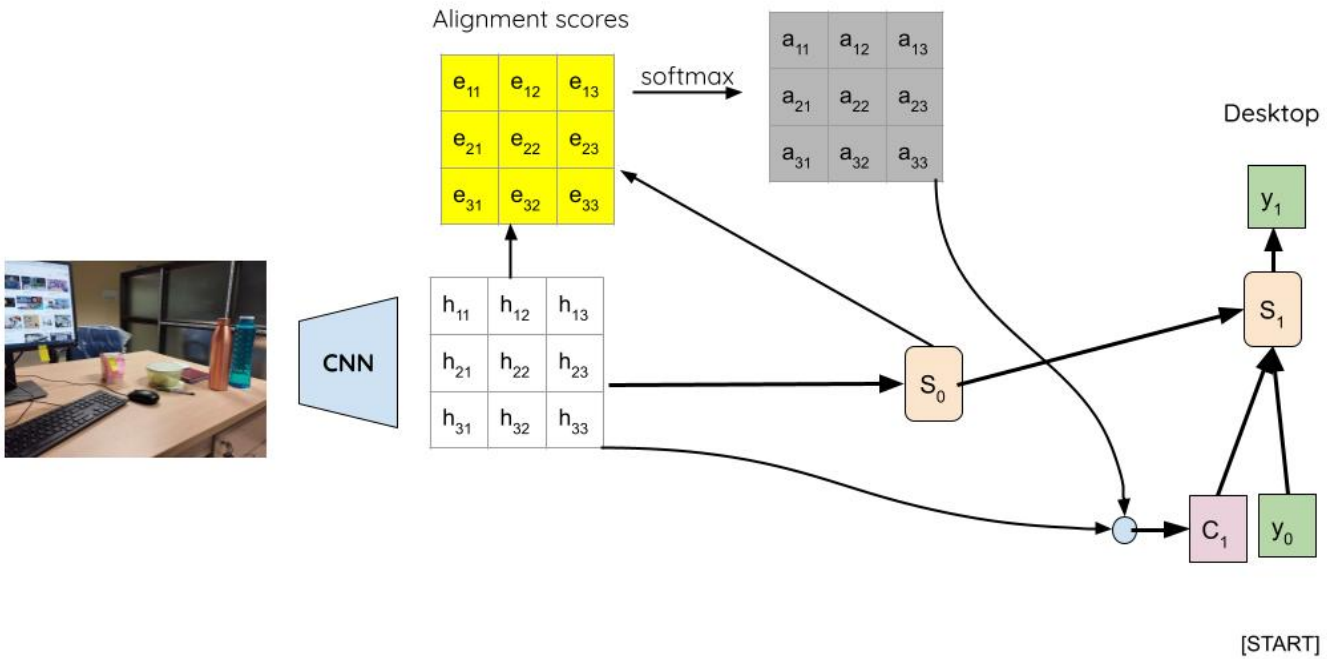

Show Attend and Tell by Xu et al. 2015

# Image captioning using RNNs and Attention



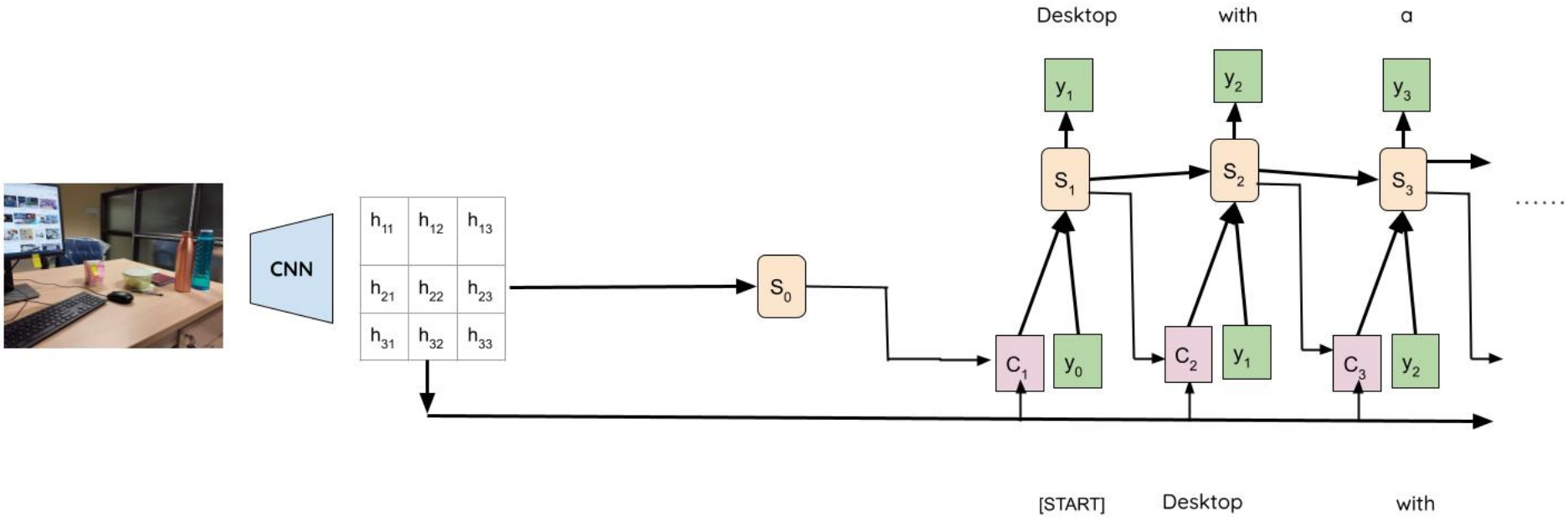

Show Attend and Tell by Xu et al. 2015

# Image captioning using RNNs and Attention




Show Attend and Tell by Xu et al. 2015

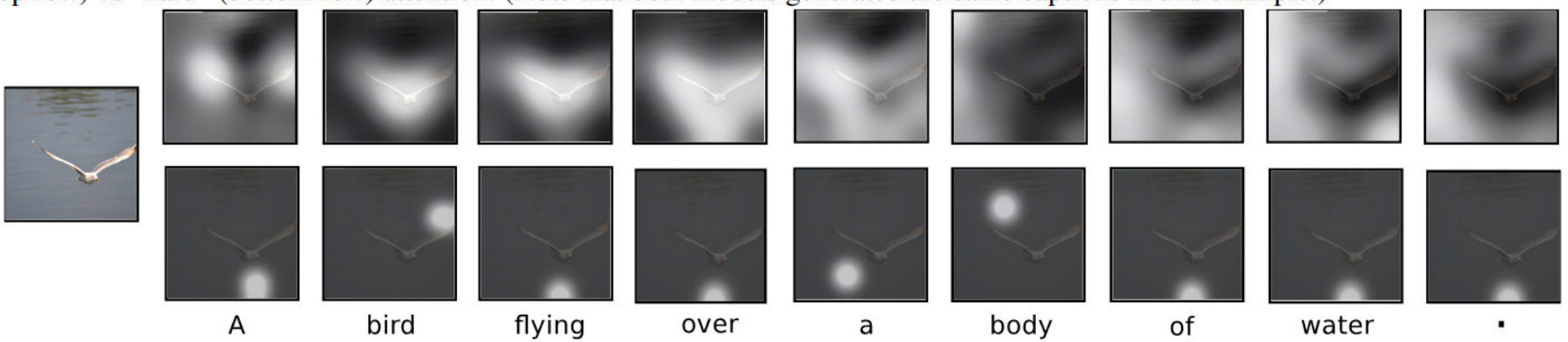
# Image captioning using RNNs and Attention




Show Attend and Tell by Xu et al. 2015

# Image captioning using RNNs and Attention

*Figure 2.* Attention over time. As the model generates each word, its attention changes to reflect the relevant parts of the image. "soft" (top row) vs "hard" (bottom row) attention. (Note that both models generated the same captions in this example.)




Show Attend and Tell by Xu et al. 2015

# Image captioning using RNNs and Attention

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.


A stop sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

Show Attend and Tell by Xu et al. 2015