Summarization with Pointer-Generator Networks

Anđelka Zečević
andjelkaz@matf.bg.ac.rs

Summarization with Pointer-Generator Networks

This talk is based on paper:
Abigail See, Peter J. Lui, Christopher D. Manning. **Get to the Point: Summarization with Pointer Generator Networks**. ACL 2017.

Source code & data available at GitHub repository:
https://github.com/abisee/pointer-generator

Summarization

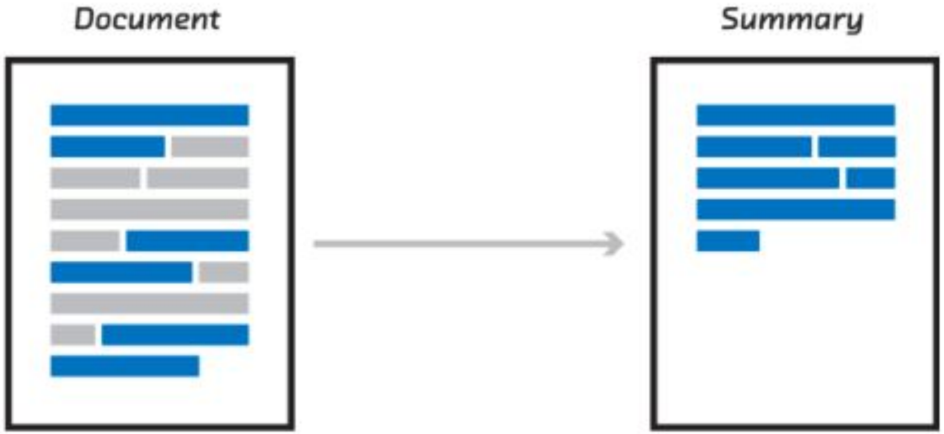
Goal:

For the given document/document collection create a summary with all salient information.

Approaches differ for:

- purpose: **generic** vs query-based
- input type: **single document** vs multi-document
- output type: extractive vs **abstractive**

Extractive Summarization



Created summary is coherent, grammatical, acurite.

Abstractive Summarization

Created summary is sophisticated, includes paraphrasing, new words, real-world knowledge, but suffers from factoid inaccuracy, repetition, and OOV handling.

Original Text (truncated): lagos, nigeria (cnn) a day after winning nigeria's presidency, ***muhammadu buhari*** told cnn's christiane amanpour that **he plans to aggressively fight corruption that has long plagued nigeria** and go after the root of the nation's unrest. *buhari* said he'll "rapidly give attention" to curbing violence in the northeast part of nigeria, where the terrorist group boko haram operates. by cooperating with neighboring nations chad, cameroon and niger, **he said his administration is confident it will be able to thwart criminals** and others contributing to nigeria's instability. for the first time in nigeria's history, the opposition defeated the ruling party in democratic elections. *buhari* defeated incumbent goodluck jonathan by about 2 million votes, according to nigeria's independent national electoral commission. **the win comes after a long history of military rule, coups and botched attempts at democracy in africa's most populous nation.**

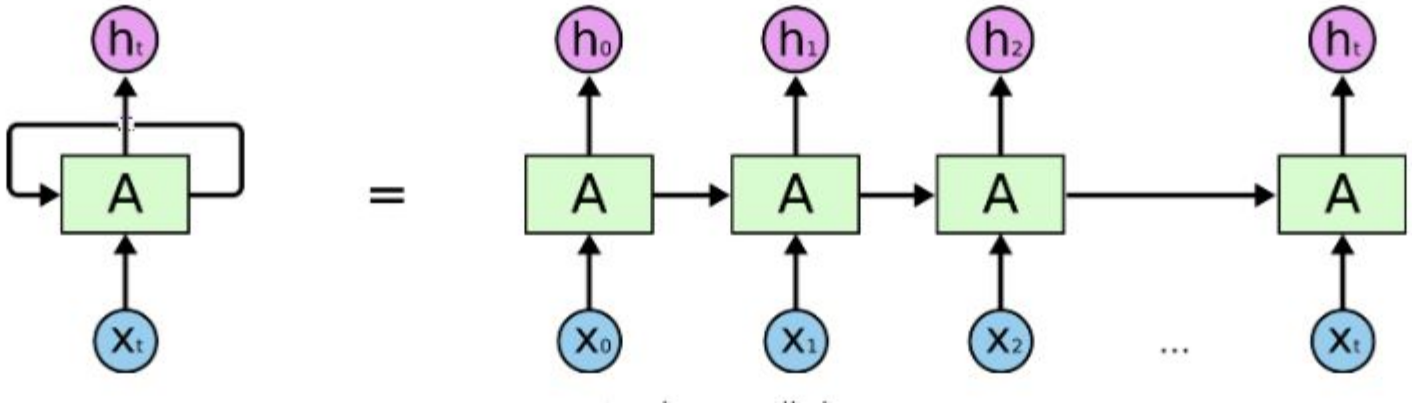
Baseline Seq2Seq + Attention: UNK UNK says his administration is confident it will be able to destabilize nigeria's economy. UNK says his administration is confident it will be able to thwart criminals and other nigerians. he says the country has long nigeria and nigeria's economy.

Pointer-Gen: *muhammadu buhari* says he plans to aggressively fight corruption **in the northeast part of nigeria**. he says he'll "rapidly give attention" to curbing violence **in the northeast part of nigeria**. he says his administration is confident it will be able to thwart criminals.

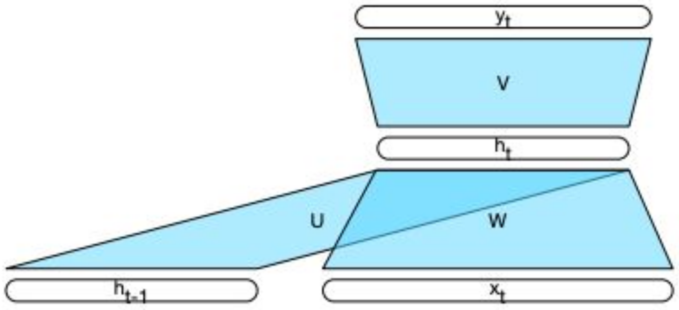
Pointer-Gen + Coverage: *muhammadu buhari* says he plans to aggressively fight corruption that has long plagued nigeria. he says his administration is confident it will be able to thwart criminals. the win comes after a long history of military rule, coups and botched attempts at democracy in africa's most populous nation.

Recurrent Neural Networks - RNNs

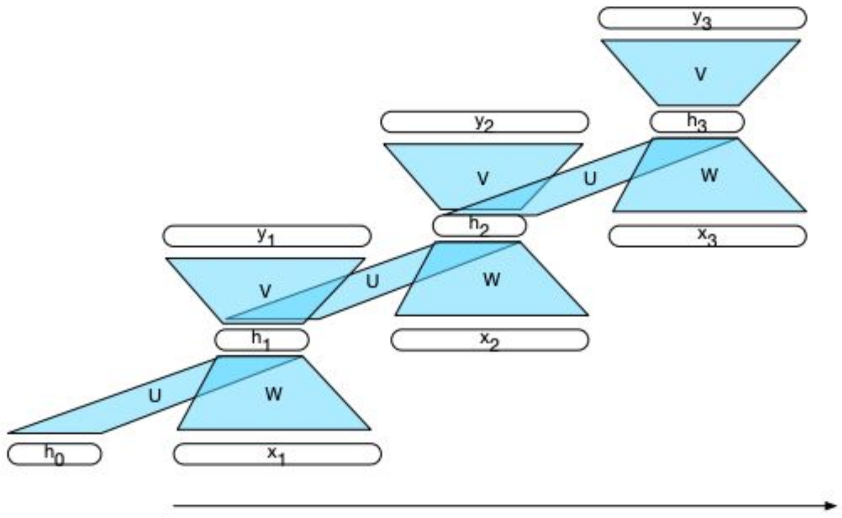
RNNs are a class of neural networks designed for a sequence processing.



Recurrent Neural Networks - RNNs



$$
\begin{aligned}
h_t &= g(Uh_{t-1} + Wx_t) \\
y_t &= f(Vh_t)
\end{aligned}
$$



from **Speech and Language Processing** by Dan Jurafsky and James H. Martin.

Recurrent Neural Networks - RNNs

Input:

X_t - word vector

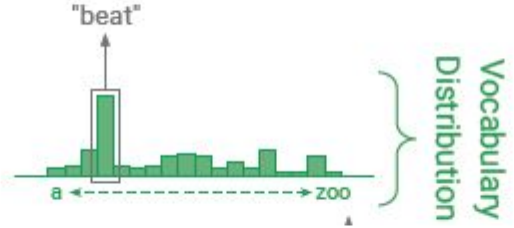
In most of the cases: word embeddings such as word2vec or Glove in combination with POS, discretized TF-IDF values, …

Output:

Y_t - output vector

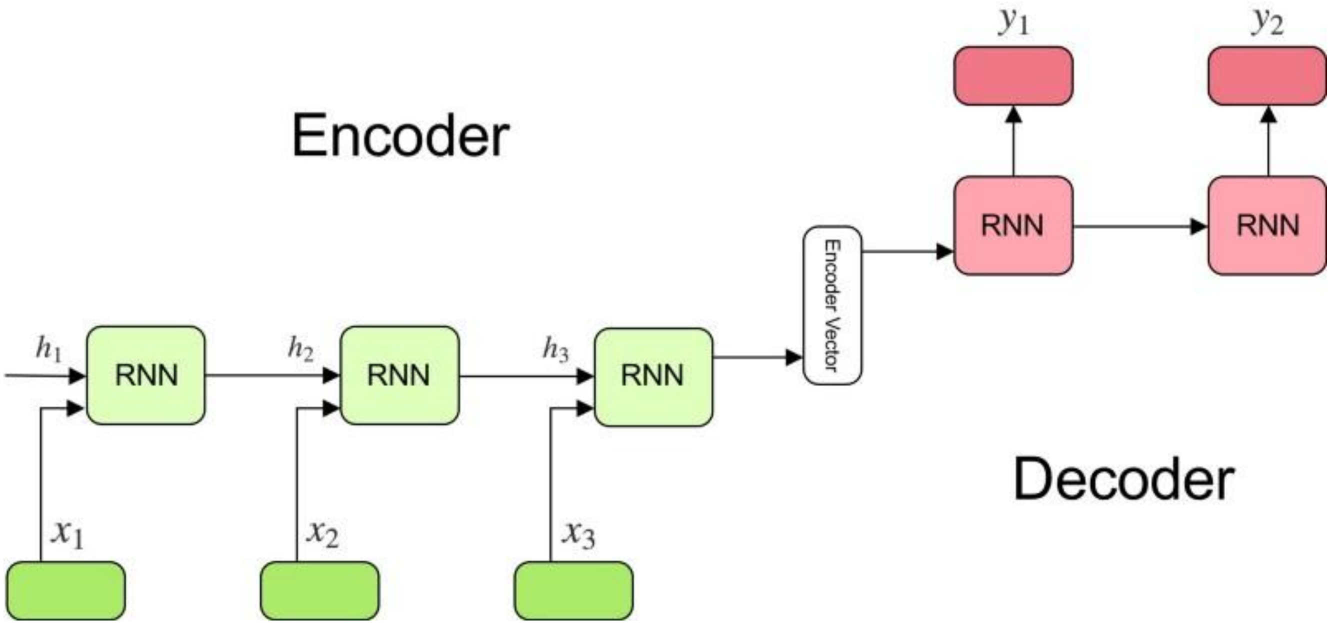
In most of the cases: f is softmax function

Interpretation: probability distribution over the possible output classes



Sequence to Sequence Model (Seq2Seq)

many-to-many mapping (many-to-one + one-to-many)



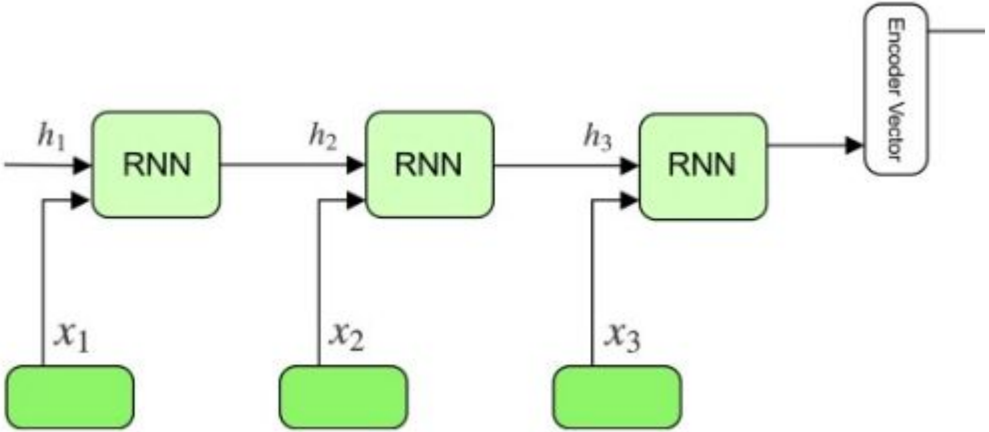
Sequence to Sequence Model

Encoder part:

Input sequence: $x = (x_1, x_2, \ldots, x_{T_x})$

Hidden state at time t: $h_t = g_e(x_t, h_{t-1})$

Context vector: $c = q(\{h_1, h_2, \ldots, h_{T_x}\})$,
for instance, $c = h_{T_x}$



Sequence to Sequence Model

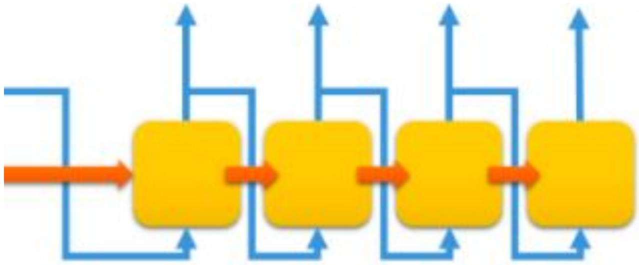
Decoder part:

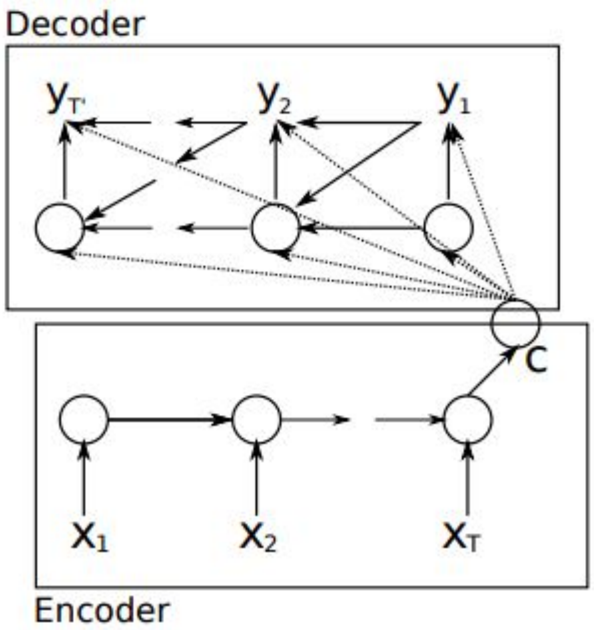
Output sequence: $y = (y_1, y_2, \ldots, y_{T_y})$

Hidden state at time t: $s_t = g_d(y_{t-1}, s_{t-1}, c)$

with $p(y_t|\{y_1, y_2, \ldots, y_{t-1}\}, c) = f_d(y_{t-1}, s_t, c)$

and goal to maximize $p(y) = \prod_{t=1}^{t=T_y} p(y_t|\{y_1, y_2, \ldots, y_{t-1}\}, c)$

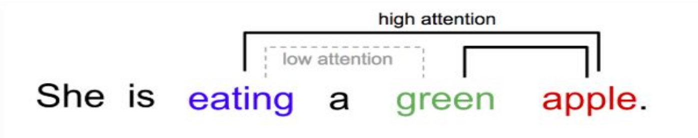


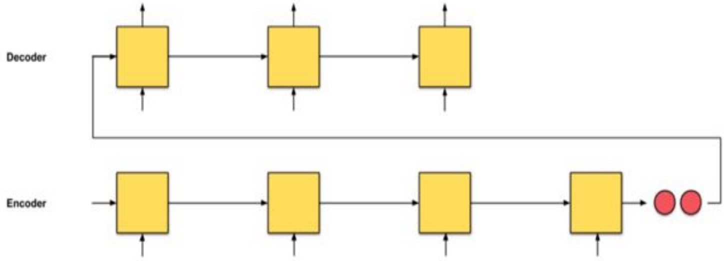


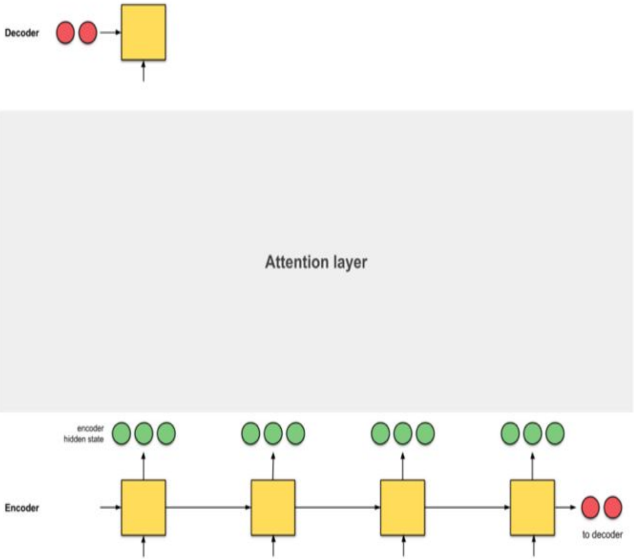
Attention

Sequence to sequence models perform badly on long sentences.

Intuition:





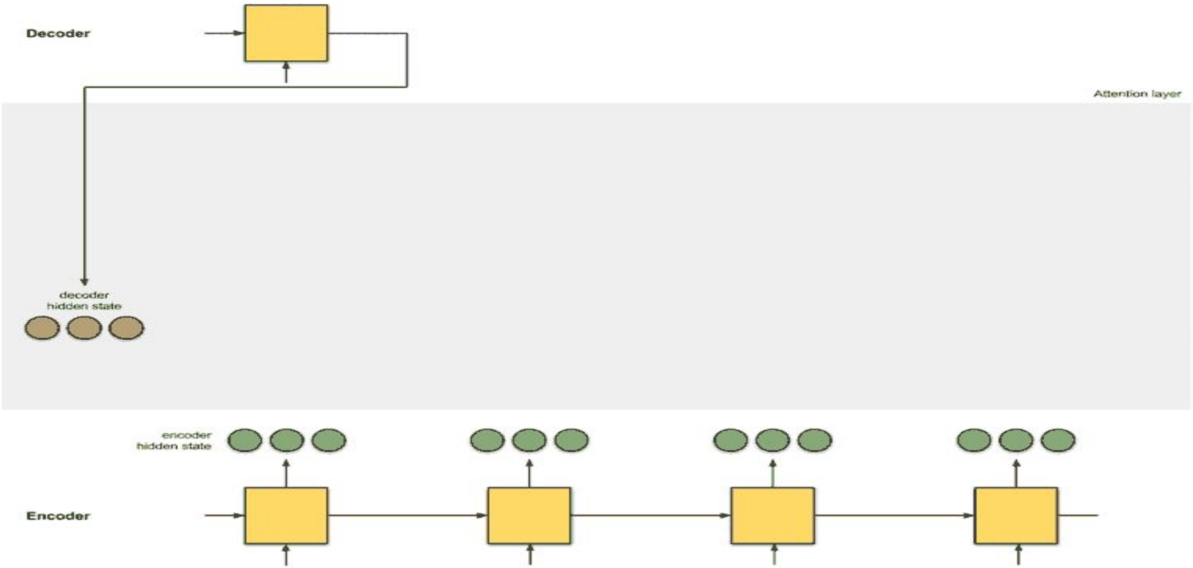


Sequence to Sequence Model with Attention

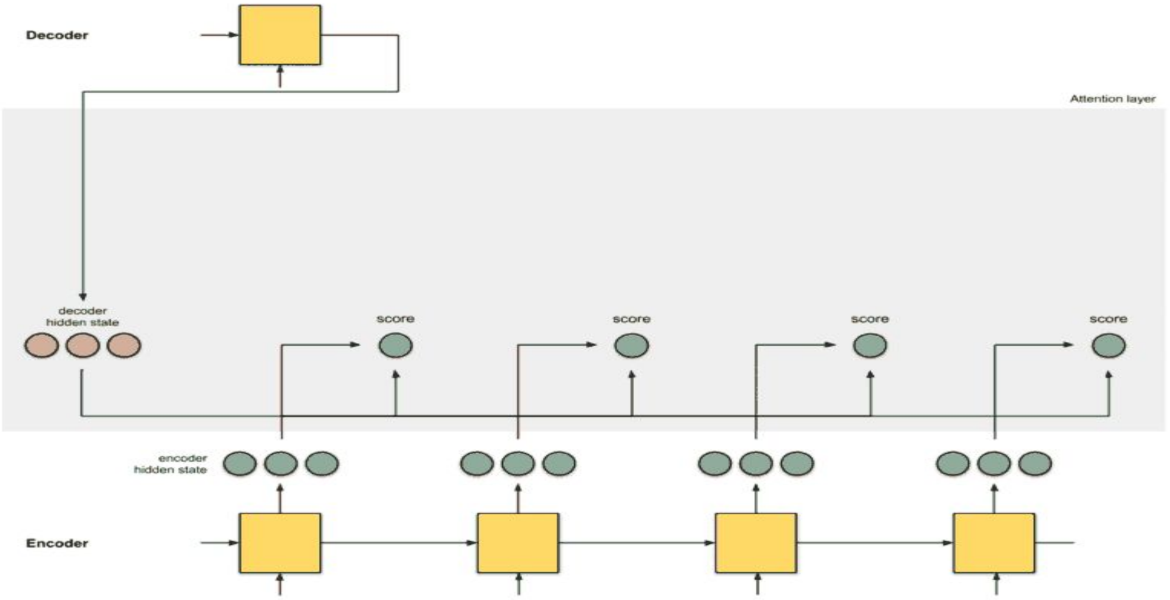
Encoder

Bahdanau et al. Neural Machine Translation by Jointly Learning to Align and Translate. ICLR 2015.
Animations are taken from https://towardsdatascience.com/attn-illustrated-attention-5ec4ad276ee3 \o/

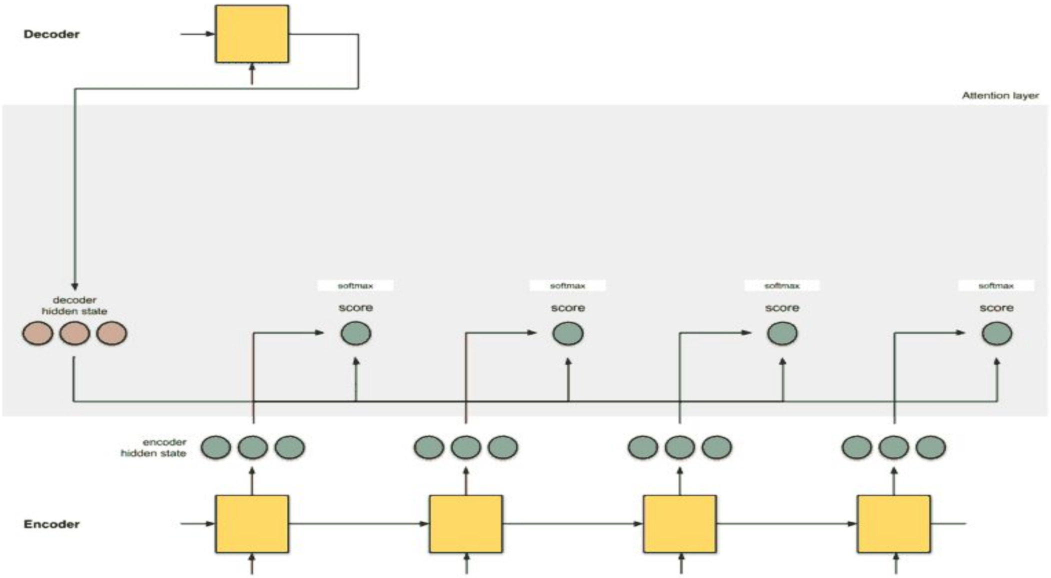
Sequence to Sequence Model with Attention



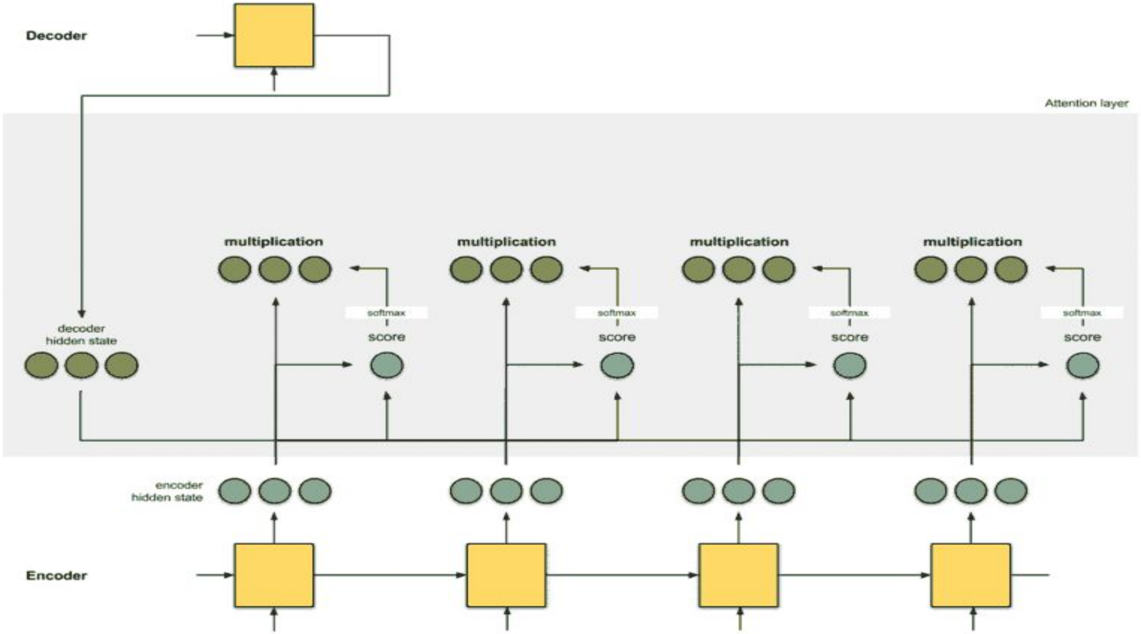
Sequence to Sequence Model with Attention



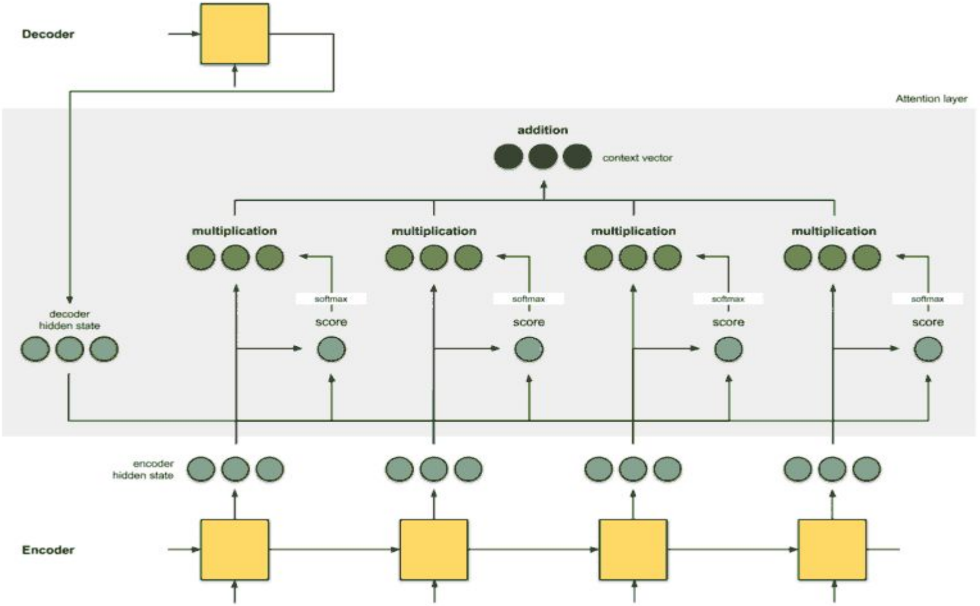
Sequence to Sequence Model with Attention



Sequence to Sequence Model with Attention



Sequence to Sequence Model with Attention



Sequence to Sequence Model with Attention

Decoder part:

Output sequence: $y = (y_1, y_2, \ldots, y_{T_y})$

Hidden state at time t: $s_t = g_d(y_{t-1}, s_{t-1}, c_t)$

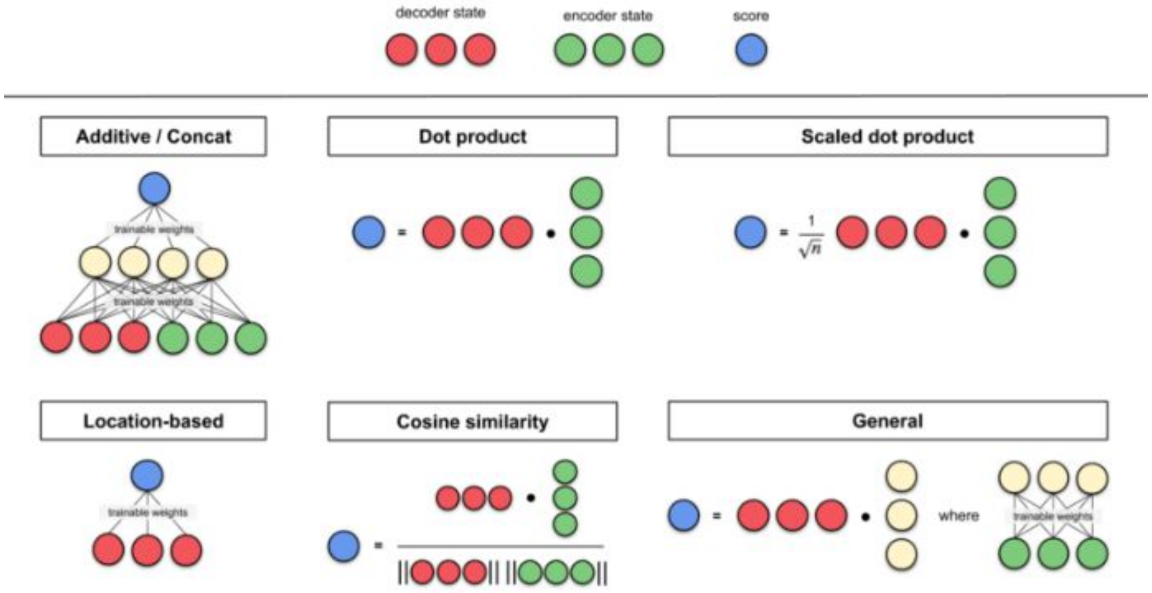
with $p(y_t | \{y_1, y_2, \ldots, y_{t-1}\}, x) = f_d(y_{t-1}, s_t, c_t)$

$c_t = \sum_{j=1}^{T_x} \alpha_{tj} h_j$ with $\alpha_{tj} = \frac{exp(e_{tj})}{\sum_{k=1}^{T_x} exp(e_{tk})}$ and $e_{tk} = a(s_{t-1}, h_k)$

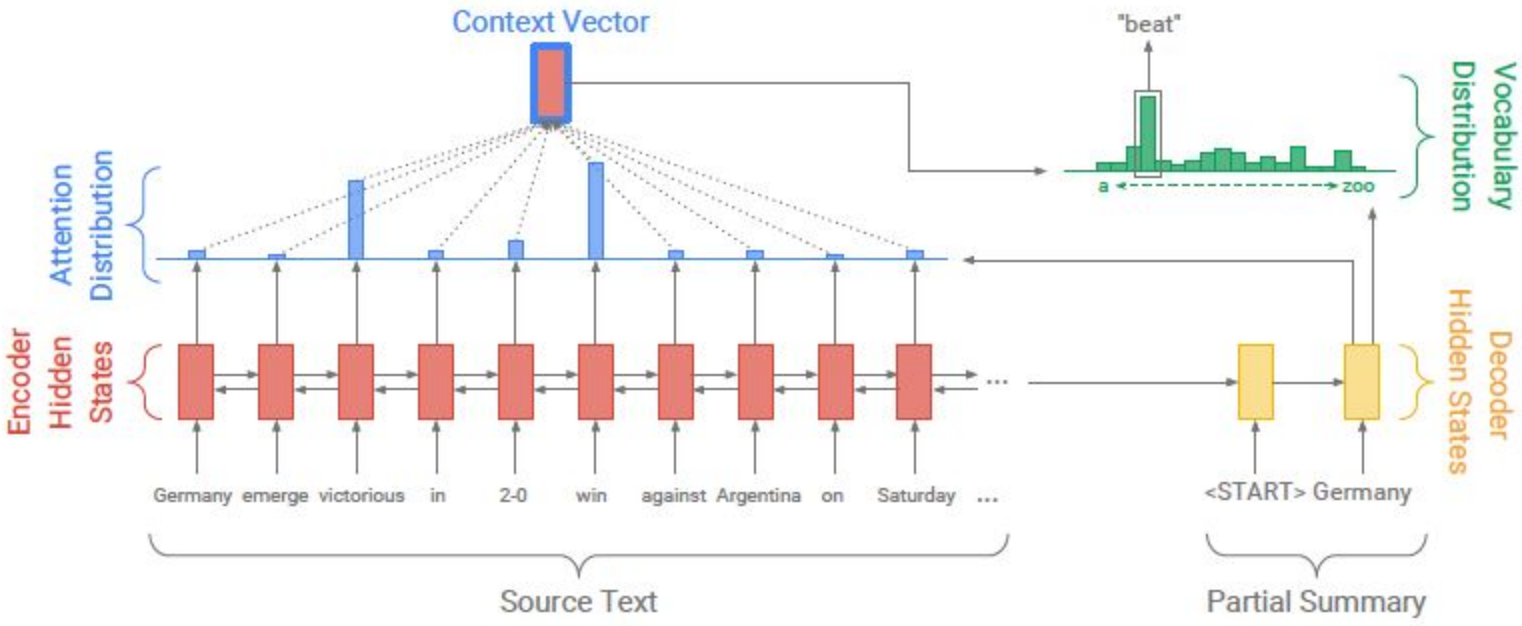
a is alignment function that scores how well the inputs around position k and the outputs at position t match.

Sequence to Sequence Model with Attention

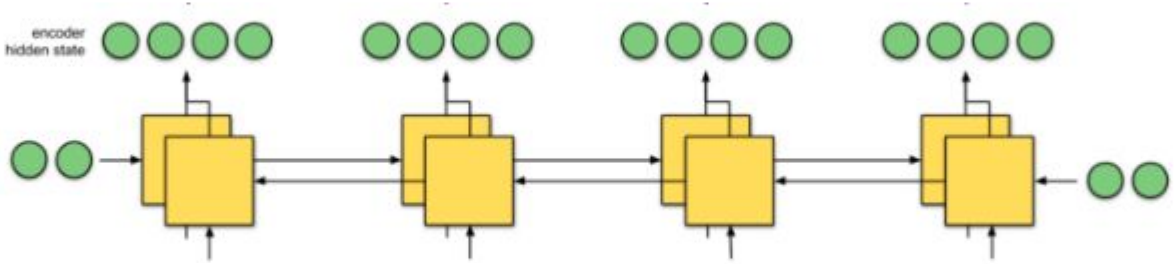
Variations of alignment functions:



Summarization - Seq2Seq Model with Attention



Bidirectional Recurrent Neural Networks - BiRNN



Feedforward RNN: hidden states $(\overrightarrow{h_1}, \overrightarrow{h_2}, \ldots, \overrightarrow{h_{T_x}})$

Backward RNN: hidden states $(\overleftarrow{h_1}, \overleftarrow{h_2}, \ldots, \overleftarrow{h_{T_x}})$

Hidden state at time t: $h = [\overrightarrow{h_t}, \overleftarrow{h_t}]$

Get to the point!

Seq2Seq with attention part:

Encoder hidden state: h_t
Decoder hidden state: s_t

$e_i^t = v^T tanh(W_h h_i + W_s s_t + b_{atten})$, i =1, 2, ..., T_x
attention vector: $a^t = softmax(e^t)$
context vector: $h_t^* = \sum_i a_i^t h_i$

decoder vocabulary distibution: $P_{vocab} = softmax(V[s_t, h_t^*] + b)$

learnable parameters: v, W_h, W_s, b_{atten}, V, b

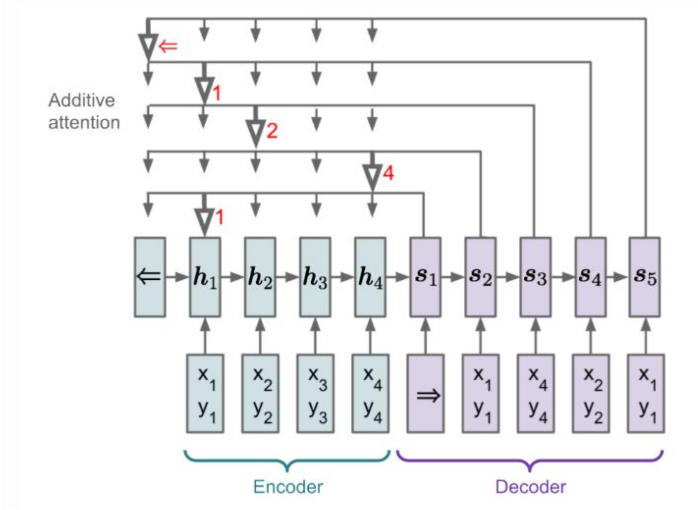
probability of word w: $P(w) = P_{vocab}(w)$

loss at time t for the target word w_t^*: $loss_t = -logP(w_t^*)$

overall loss: $loss = \frac{1}{T} \sum_{t=0}^{T} loss_t$

Pointer Networks (Ptr-Nets)

Sequence-to-sequence networks with the output elements that correspond to positions in an input sequence.



Vinyals et al. Pointer Networks. NIPS 2015.

Pointer Networks

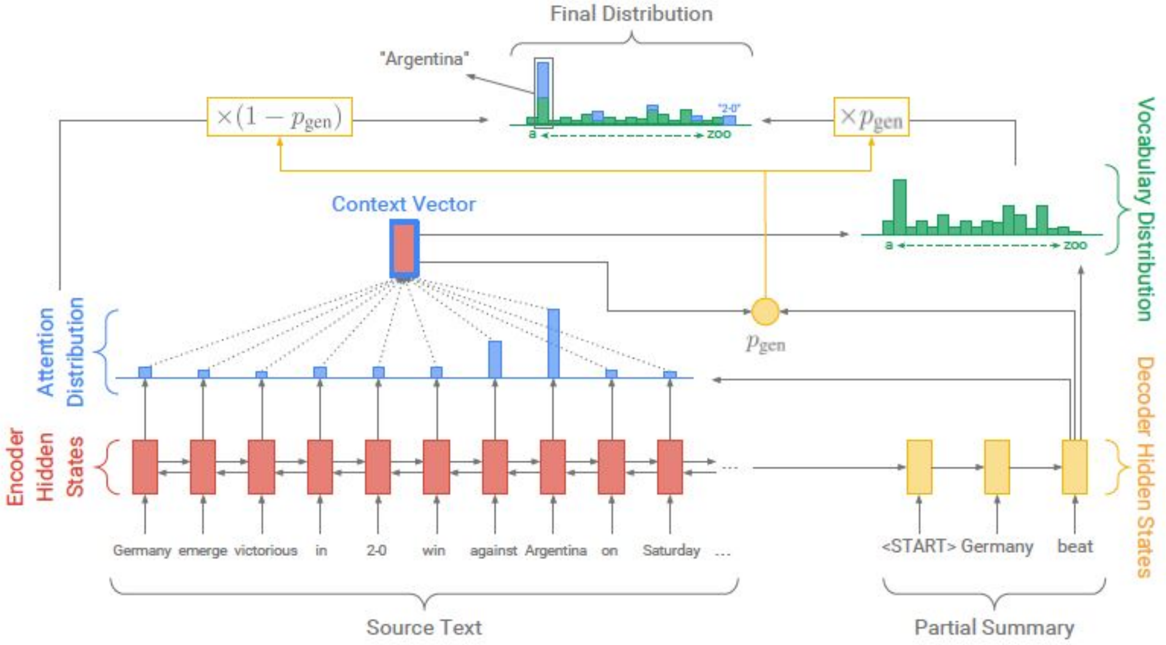
Instead of using attention to blend hidden units of an encoder to a context vector at each docoder step, Ptr-Nets use attention as a pointer to select a member of the input sequence as the output.

Here $\mathcal{P} = \{P_1, \ldots, P_n\}$ is a sequence of n vectors and $\mathcal{C}^{\mathcal{P}} = \{C_1, \ldots, C_{m(\mathcal{P})}\}$ is a sequence of $m(\mathcal{P})$ indices, each between 1 and n.

$$\begin{aligned} u_j^i &= v^T \tanh(W_1 e_j + W_2 d_i) \quad && j \in (1, \ldots, n) \\ a_j^i &= \operatorname{softmax}(u_j^i) && j \in (1, \ldots, n) \\ d_i' &= \sum_{j=1}^{n} a_j^i e_j \end{aligned}$$

$$\begin{aligned} u_j^i &= v^T \tanh(W_1 e_j + W_2 d_i) \quad j \in (1, \ldots, n) \\ p(C_i | C_1, \ldots, C_{i-1}, \mathcal{P}) &= \operatorname{softmax}(u^i) \end{aligned}$$

Summarization with Pointer-Generator Networks



P_{gen} from [0, 1] is used as a soft switch to choose between generating or copying.

Get to the point!

Pointer-generator network:

Encoder hidden state: h_t

Decoder hidden state: s_t

Decoder input: y_t

Generation probability:

$p_{gen} = \sigma(w_{h^*}^T h_t^* + w_s^T s_t + w_y^T y_t + b_{ptr})$

learnable parameters: $w_{h^*}, w_s, w_y, {}_{ptr}$

probability of word w over extended vocabulary: $P(w) = p_{gen} P_{vocab}(w) + (1 - p_{gen}) \sum_{i:w_i=w} a_i^t$

Coverage

Originally from NMT:
vector that indicates whether a **source word is translated or not**

It should help with over-translation and under-translation.

In the context of document summarization, it should control repetition.

coverage vector: $c^t = \sum_{t'=0}^{t-1} a^{t'}$

$e_i^t = v^T tanh(W_h h_i + W_s s_t + w_c c_i^t + b_{atten})$

loss at time t for the target word w_t^*: $loss_t = -logP(w_t^*) + \lambda \sum_i min(a_i^t, c_i^t)$

Dataset

CNN/Daily Mail dataset of online news articles paired with multi-sentence summaries.

- articles: 781 tokens on average
- summaries: 56 tokens on average

Train set: 287, 226
Validation set: 13, 368
Test set: 11, 490

Dataset

Source Text

munster have signed new zealand international francis *saili* on a two-year deal . utility back *saili* , who made his all blacks debut against argentina in 2013 , will move to the province later this year after the completion of his 2015 contractual commitments . the 24-year-old currently plays for *auckland-based* super rugby side the blues and was part of the new zealand under-20 side that won the junior world championship in italy in 2011 . *saili* 's signature is something of a coup for munster and head coach anthony foley believes he will be a great addition to their backline . francis *saili* has signed a two-year deal to join munster and will link up with them later this year . ' we are really pleased that francis has committed his future to the province , ' foley told munster 's official website . ' he is a talented centre with an impressive *skill-set* and he possesses the physical attributes to excel in the northern hemisphere . ' i believe he will be a great addition to our backline and we look forward to welcoming him to munster . ' *saili* has been capped twice by new zealand and was part of the under 20 side that won the junior championship in 2011 . *saili* , who joins all black team-mates dan carter , *ma'a nonu* , conrad smith and charles *piutau* in agreeing to ply his trade in the northern hemisphere , is looking forward to a fresh challenge . he said : ' i believe this is a fantastic opportunity for me and i am fortunate to move to a club held in such high regard , with values and traditions i can relate to from my time here in the blues . ' this experience will stand to me as a player and i believe i can continue to improve and grow within the munster set-up . ' as difficult as it is to leave the blues i look forward to the exciting challenge ahead . '

Reference summary

utility back francis *saili* will join up with munster later this year .
the new zealand international has signed a two-year contract .
saili made his debut for the all blacks against argentina in 2013 .

Summarization Evaluation

ROUGE: Recall-Oriented Understudy for Gisting Evaluation

$$ROUGE-n = \frac{\sum_{C \in RSS} \sum_{gram_n \in C} Count_{match}(gram_n)}{\sum_{C \in RSS} \sum_{gram_n \in C} Count(gram_n)}$$

Standard measures are ROUGE-1, ROUGE-2, ROUGE-L (longest common sequence)

Experiment - in numbers

Word representations: 128-dimensional word embeddings
Source and target vocabulary size: 50, 000 words/150 000 words
Truncated article size: 400 tokens
Maximal summary length: 100
Hidden state: 256-dimensional vector
Total number of network parameters: 21499600+1153+512 = 21 501 265

Adagard with learning rate 0.15 and an initial accumulator value 0.1
Gradient clipping with a maximum gradient norm of 2
Early stopping
Batch size: 16

Experiment - in numbers

Baseline Model:

Training on Single Tesla K40m GPU 600 000 iterations (33 epochs)

Training time for baseline model: 4 days 14 hours / 8 days 21 hours

Pointer-Generator Model:

Training on Single Tesla K40m GPU 230 000 iterations (13 epochs)

Training time for baseline model: 3 days 4 hours

Final model:

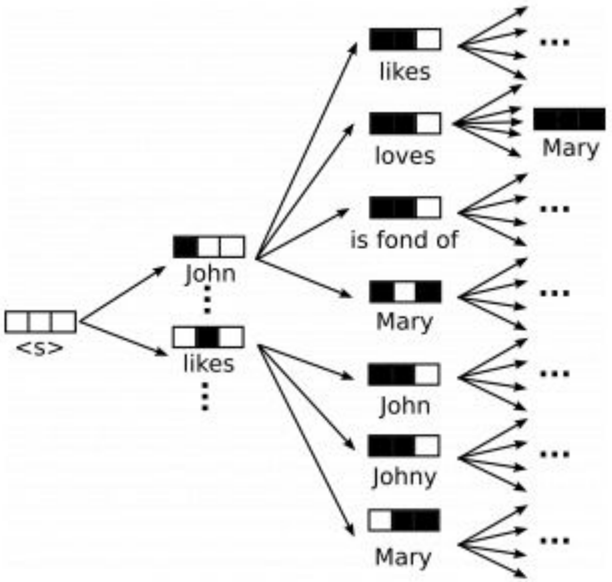
+ additional 3000 iterations with coverage (2 hours)

Experiment - in numbers

At test time:

Maximal summary length: 120

Beam search with beam size: 4



Thank you!