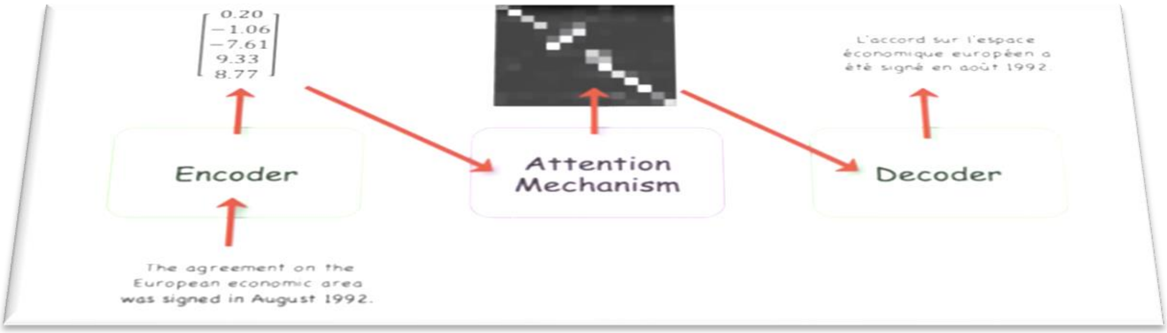
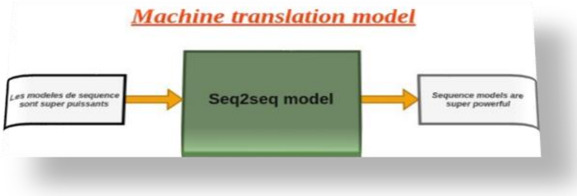

0.20
−1.06
−7.61
9.33
8.77
Encoder
Attention Mechanism
Decoder
The agreement on the European economic area was signed in August 1992.
L'accord sur l'espace économique européen a été signé en août 1992.

How the term of Attention Mechanism was introduced and What and Why of Attention Mechanism

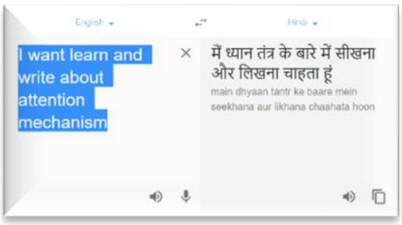
Introduction of seq2seq models and application:


Machine translation model
Les modeles de sequence sont super puissants
Seq2seq model
Sequence models are super powerful

Sequence to sequence was first introduced by Google in 2014. So let's go through our question what is seq2 seq model? Sequence to sequence model tries to map input text with fixed length to output text fixed-length where the length of input and output to the model may differ. As we know variants of Recurrent neural networks like Long short-term memory or Gated Recurrent Neural Network (GRU) are the method we mostly used since they overcome the problem of vanishing gradient.

From the example shown in the image is of language conversion from French to English.

Another example of English to Hindi Translation. Which is nothing but google translation.



I want learn and write about attention mechanism

मैं ध्यान तंत्र के बारे में सीखना और लिखना चाहता हूं

main dhyaan tantr ke baare mein seekhana aur likhana chaahata hoon

Sequence to Sequence Learning with Neural Networks was introduced by

Ilya Sutskever Google ilyasu@google.com
Oriol Vinyals Google vinyals@google.com
Quoc V. Le Google qvl@google.com

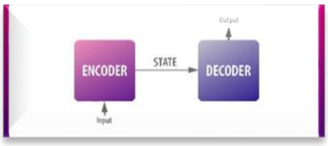
Paper Reference:

- https://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf
- https://arxiv.org/abs/1409.3215

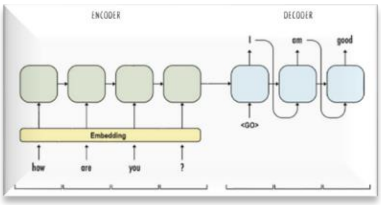
Application of Seq2seq models

- Speech Recognition
- Machine Language Translation
- Name entity/Subject extraction
- Relation Classification
- Path Query Answering
- Speech Generation
- Chatbot

- Text Summarization
- Product Sales Forecasting



Seq2Seq is encoder and decoder.



So why does the seq2seq model fails?

Details of the architecture and function already explained by my colleagues so now we see where this model lags.

- As we saw encoder takes input and converts it into a fixed-size vector and then the decoder makes a prediction and gives output sequence. It works fine for short sequence but it fails when we have a long sequence because it becomes difficult for the encoder to memorize the entire sequence into a fixed-sized vector and to compress all the contextual information from the sequence. As we observed that as the sequence size increases model performance starts getting degrading.

How can we overcome the problem of long sentences and performance of the model?

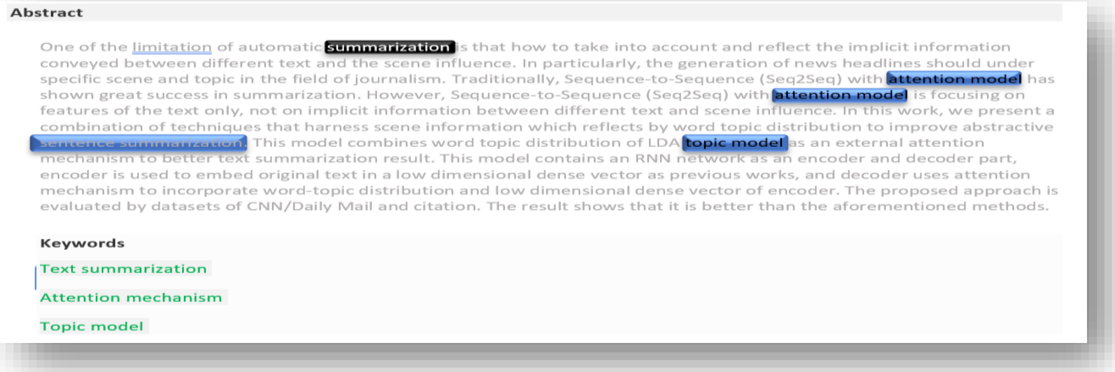
Here comes the solution with Attention Mechanism

As the word 'attention' suggest importance is given to specific part of context while so as to increase the performance and output interpretation is starts to make sense. In simple terms we give importance to specific parts of the sequence instead of the entire sequence predict that word. Basically, in the attention, we don't throw away the intermediate from the encoder state but we utilize this to generate context vector from all states so that the decoder gives output result.

- The attention mechanism has changed the way we work with deep learning algorithms
- Fields like Natural Language Processing (NLP) and even Computer Vision have been revolutionized by the attention mechanism

For Example: For Deep learning we have to read an article and get the inference out it. Or a whole book. Like the human brain attention is given to specific words which mind interprets and grasps others are just a blurry information.

Text Attention



Abstract

One of the limitation of automatic summarization is that how to take into account and reflect the implicit information conveyed between different text and the scene influence. In particularly, the generation of news headlines should under specific scene and topic in the field of journalism. Traditionally, Sequence-to-Sequence (Seq2Seq) with attention model has shown great success in summarization. However, Sequence-to-Sequence (Seq2Seq) with attention model is focusing on features of the text only, not on implicit information between different text and scene influence. In this work, we present a combination of techniques that harness scene information which reflects by word topic distribution to improve abstractive sentence summarization. This model combines word topic distribution of LDA topic model as an external attention mechanism to better text summarization result. This model contains an RNN network as an encoder and decoder part, encoder is used to embed original text in a low dimensional dense vector as previous works, and decoder uses attention mechanism to incorporate word-topic distribution and low dimensional dense vector of encoder. The proposed approach is evaluated by datasets of CNN/Daily Mail and citation. The result shows that it is better than the aforementioned methods.

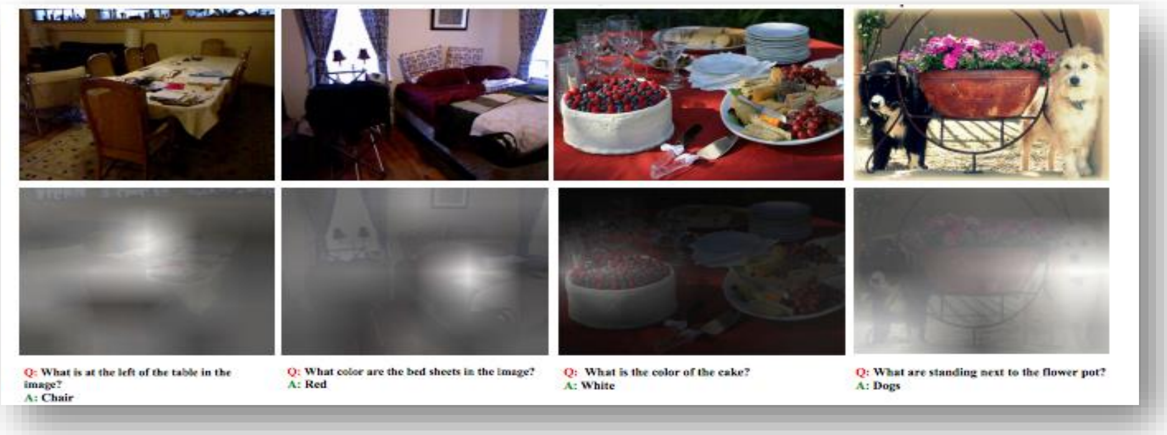
Keywords

Text summarization

Attention mechanism

Topic model

Image Attention



Q: What is at the left of the table in the image?
A: Chair

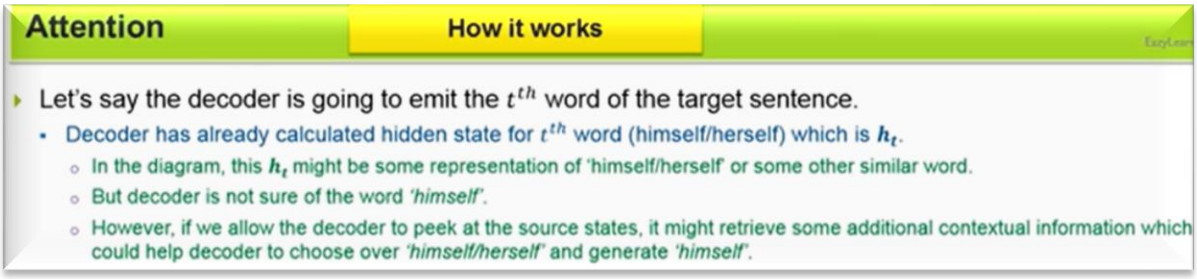
Q: What color are the bed sheets in the image?
A: Red

Q: What is the color of the cake?
A: White

Q: What are standing next to the flower pot?
A: Dogs

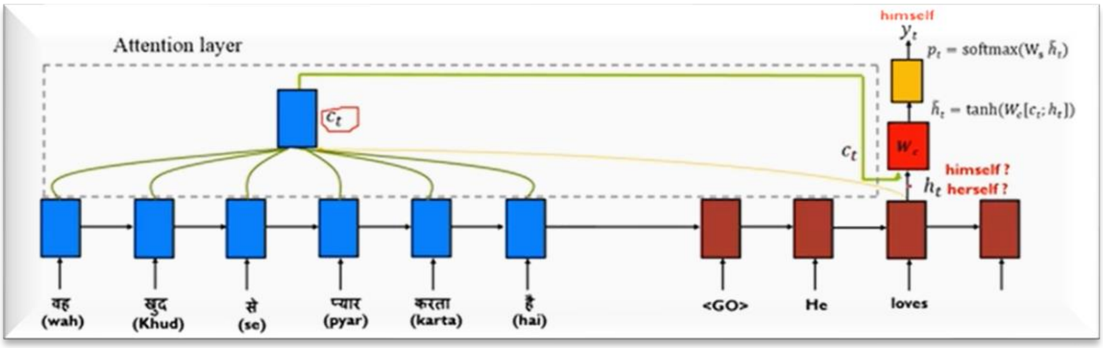
"So, whenever the proposed model generates a sentence, it searches for a set of positions in the encoder hidden states where the most relevant information is available. This idea is called 'Attention'."

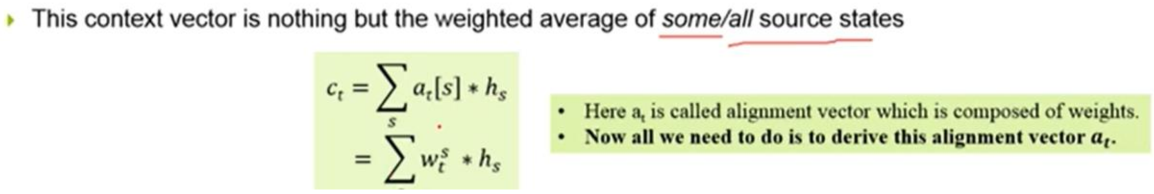
How it Works



Attention **How it works**

- Let's say the decoder is going to emit the t^{th} word of the target sentence.
  - Decoder has already calculated hidden state for t^{th} word (himself/herself) which is $\boldsymbol{h_t}$.
    - In the diagram, this $\boldsymbol{h_t}$ might be some representation of 'himself/herself' or some other similar word.
    - But decoder is not sure of the word *'himself'*.
    - However, if we allow the decoder to peek at the source states, it might retrieve some additional contextual information which could help decoder to choose over *'himself/herself'* and generate *'himself'*.



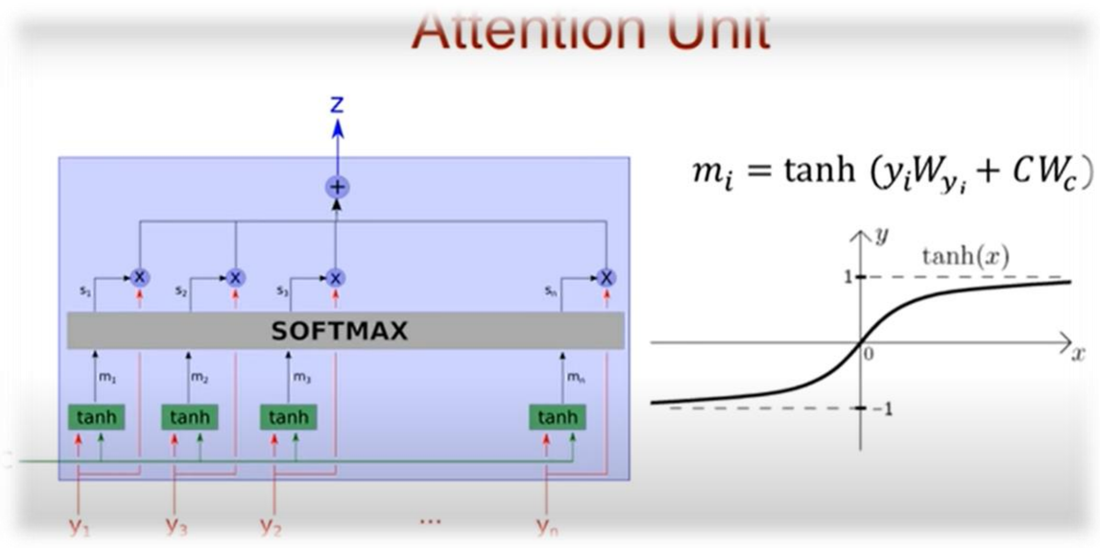


- This context vector is nothing but the weighted average of *some/all* source states

$$c_t = \sum_s a_t[s] * h_s$$
$$= \sum w_t^s * h_s$$

- Here a_t is called alignment vector which is composed of weights.
- **Now all we need to do is to derive this alignment vector $\boldsymbol{a_t}$.**

Attention Unit



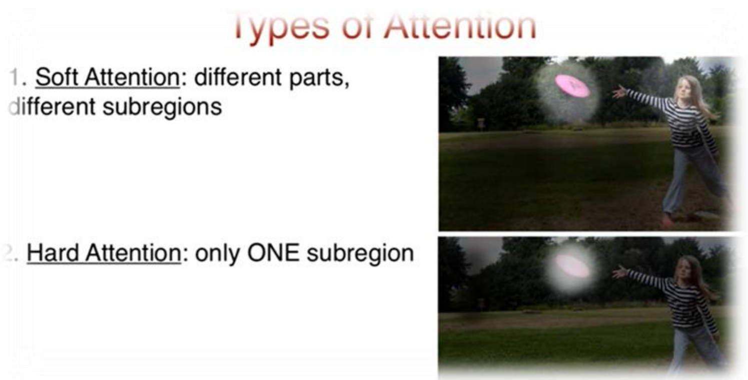
Attention Unit

$$m_i = \tanh\left(y_i W_{y_i} + CW_c\right)$$

a

Soft

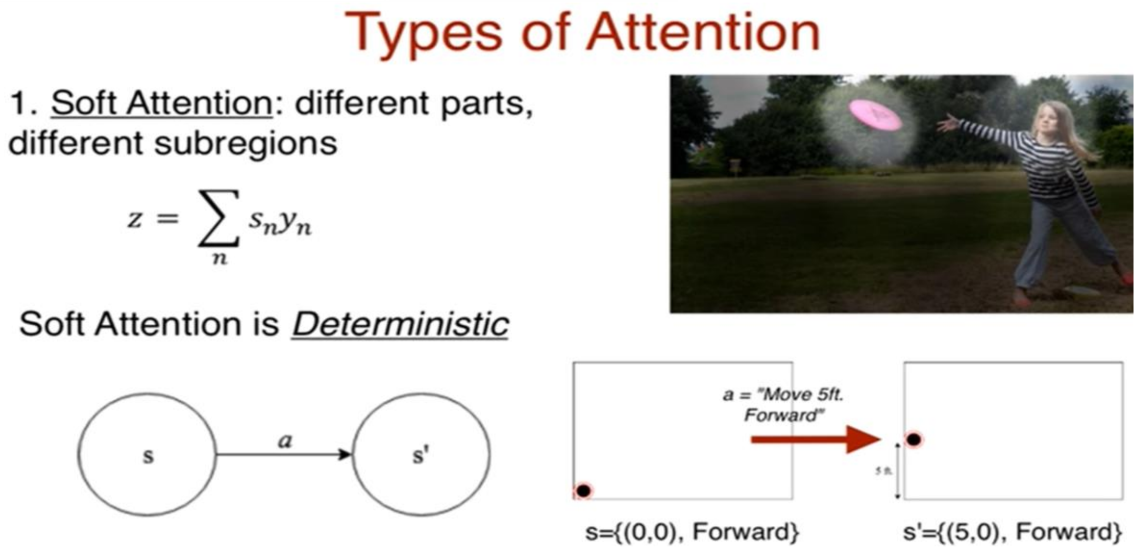
Attention



Types of Attention

1. Soft Attention: different parts, different subregions

2. Hard Attention: only ONE subregion



Types of Attention

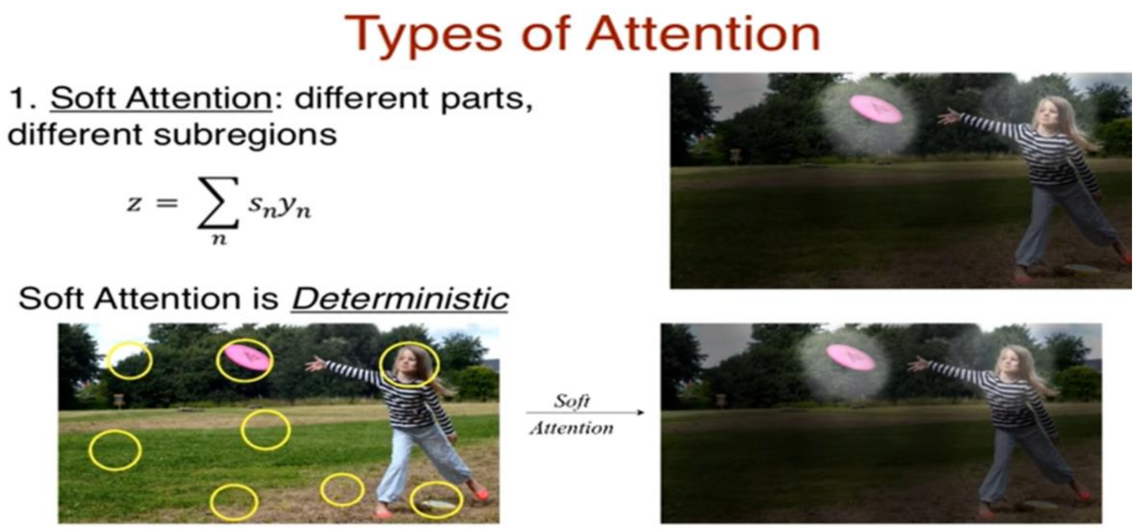
1. Soft Attention: different parts, different subregions

$$z = \sum_n s_n y_n$$

Soft Attention is *Deterministic*

s={(0,0), Forward}

s'={(5,0), Forward}



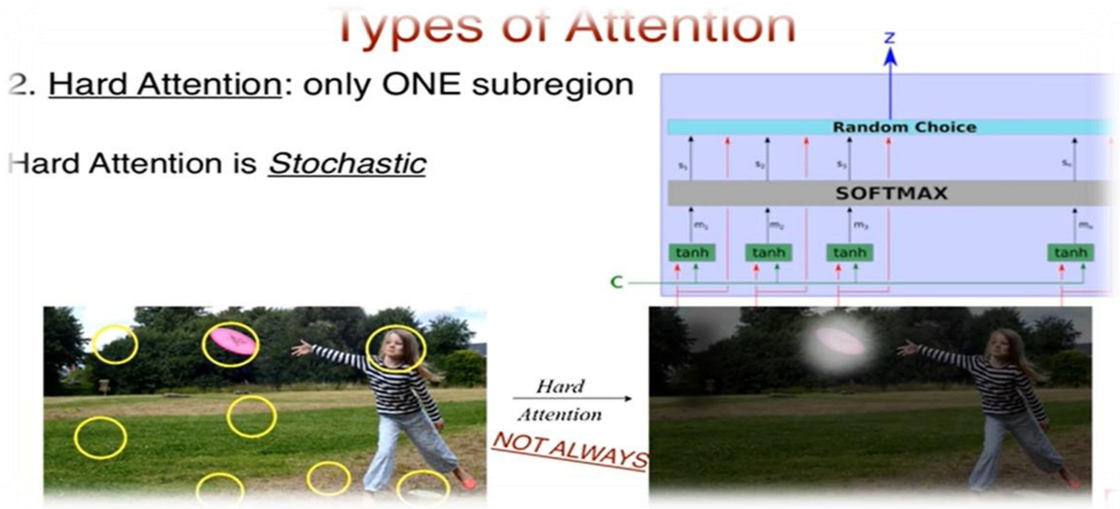
Types of Attention

1. Soft Attention: different parts, different subregions

$$z = \sum_n s_n y_n$$

Soft Attention is *Deterministic*

Hard Attention

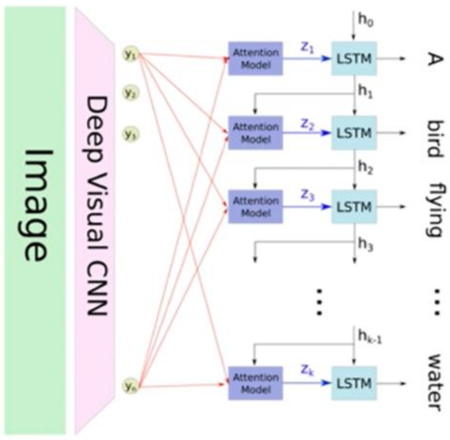


Types of Attention

2. Hard Attention: only ONE subregion

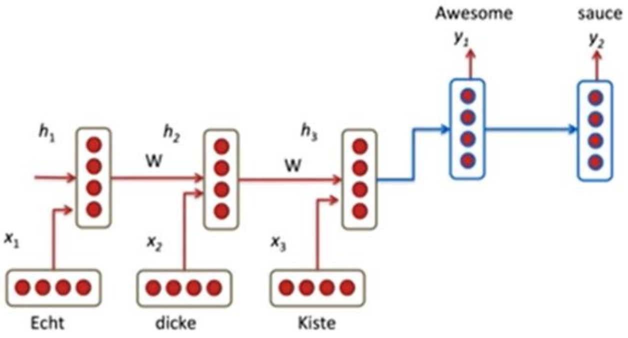
Hard Attention is *Stochastic*

Architecture



Applications

1. Neural Machine Translation (NMT)



2. Microsoft's Attention GANs

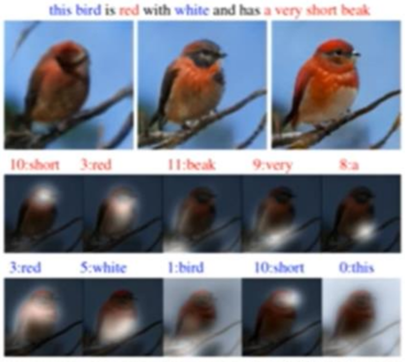
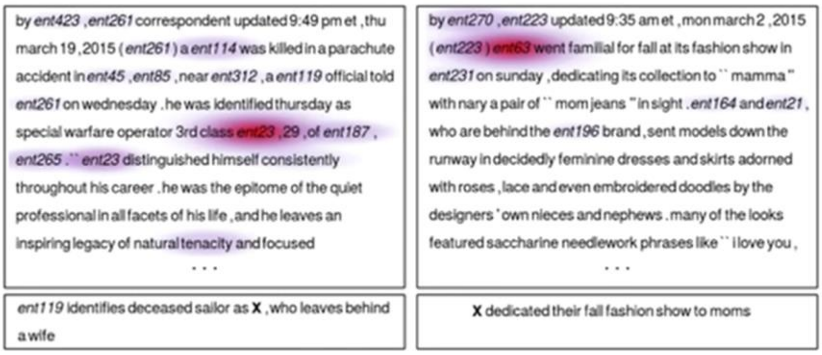


Figure 1. Example results of the proposed AttnGAN. The first row gives the low-to-high resolution images generated by G_0, $G1$ and G_2 of the AttnGAN; the second and third row shows the top-5 most attended words by F_1^{attn} and F_2^{attn} of the AttnGAN, respectively. Here, images of G_0 and G_1 are bilinearly upsampled to have the same size as that of G_2 for better visualization.

3. Teaching Machines to Read & Comprehend



by *ent423* ,*ent261* correspondent updated 9:49 pm et ,thu march 19 ,2015 (*ent261*) a *ent114* was killed in a parachute accident in *ent45* ,*ent85* ,near *ent312* ,a *ent119* official told *ent261* on wednesday .he was identified thursday as special warfare operator 3rd class *ent23* ,29 ,of *ent187* , *ent265* .`` *ent23* distinguished himself consistently throughout his career .he was the epitome of the quiet professional in all facets of his life ,and he leaves an inspiring legacy of natural tenacity and focused

. . .

ent119 identifies deceased sailor as **X** ,who leaves behind a wife

by *ent270* ,*ent223* updated 9:35 am et ,mon march 2 ,2015 (*ent223*) *ent63* went familial for fall at its fashion show in *ent231* on sunday ,dedicating its collection to `` mamma " with nary a pair of `` mom jeans "in sight .*ent164* and *ent21* , who are behind the *ent196* brand ,sent models down the runway in decidedly feminine dresses and skirts adorned with roses ,lace and even embroidered doodles by the designers ' own nieces and nephews .many of the looks featured saccharine needlework phrases like `` i love you ,

. . .

X dedicated their fall fashion show to moms

Drawbacks:

Only one drawback of attention is that it's time-consuming. To overcome this problem Google introduced "Transformer Model" .

Summary

1. Attention involves focus of *certain* *parts* of input

2. Types of Attention: Soft & Hard Attention

3. Soft Attention is Deterministic. Hard Attention is Stochastic.

4. Attention is used in NMT, AttnGAN, teaching machines to read.