Image Captioning using Visual Attention

Final Project Group1

Contributors: Adel Hassen, Lydia Teinfalt, Alice Yang

Machine Learning 2 (George Washington University, Spring 2022)

Content

Introduction

Model

Attention Method

Implementation

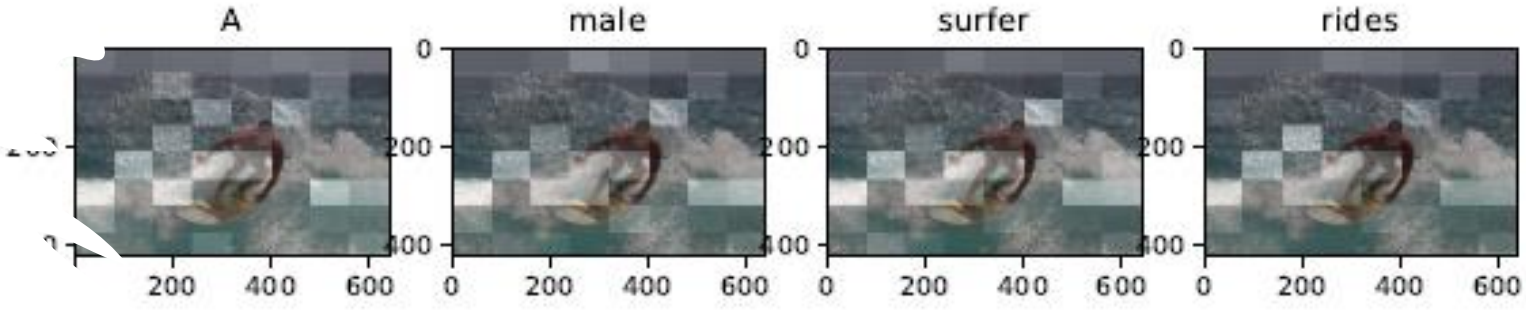
Results

Conclusions

Introduction

Image Caption Generation

- Problem Statement
  - Automatically generating caption to pictures
  - Detecting objects within an image
  - Translating an image to text
  - Creating the caption in natural language
- Why is it important?
- How can Neural Network help in solving this problem?
  - Transfer Learning
  - CNN
  - RNN



[Image Source](); *License: Public Domain*

COCO 2014 Dataset

- COCO: Common Objects in Context
- Large scale dataset published by Microsoft. Dataset organized by year of published date
- 2014 Dataset
  - Original dataset has over 120,000 images for training, validation and testing
  - 91 object categories
  - 82 categories with 5,000 labeled images
  - Our project used subset of the COCO dataset. Each image has five captions
  - COCO dataset was used for training and testing
  - New Yorker cartoons for evaluation

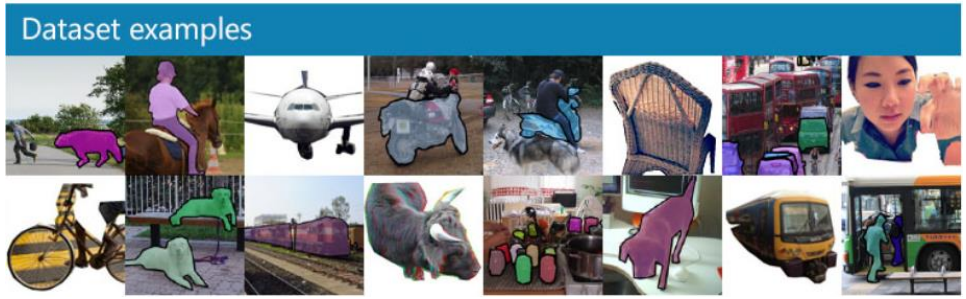
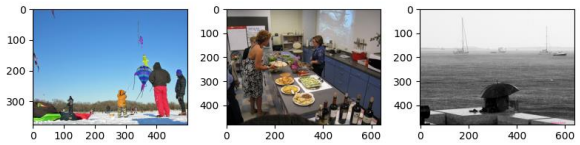
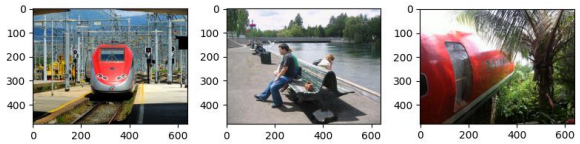


Image Source: https://cocodataset.org/#home





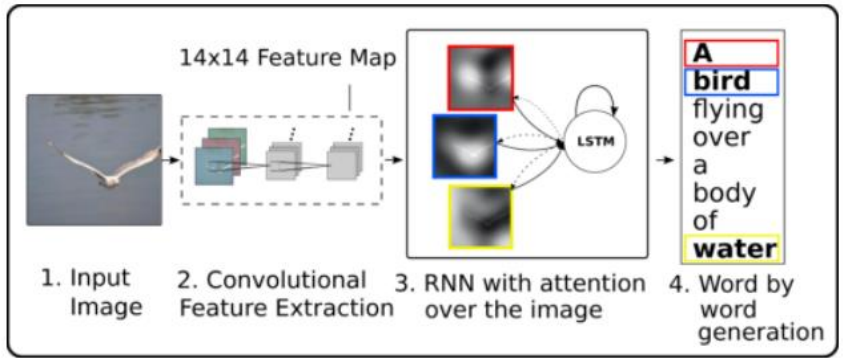
<start> a group of people in a snowy field with kites above <end>
<start> A table with lots of plates of food with wines and a slideshow in the background. <end>
<start> A couple with an umbrella overlooking boats in the water. <end>
<start> a red bullet train is coming towards us <end>
<start> People sitting on a bench near the water. <end>
<start> A red and white plane in heavy forest area. <end>

Sample Data and Caption from COCO 2014

- <start> a tan horse standing next to a fence with a head over it <end>
- <start> A man holding a frisbee while standing on a beach. <end>
- <start> The small kitchen has a bar across from the stove. <end>
- <start> Television sitting outside on the sidewalk on a sunny day. <end>
- <start> A very pretty long red train by a platform. <end>
- <start> a couple of birds standing by some water <end>

Model Architecture

- [Show, Attend and Tell: Neural Image Caption Generation with Visual Attention]
- Paper published in 2016 by Xu, Kelvin. Ba, Jimmy. Kiros, Ryan. Cho, Kyunghyum. Courville, Aaron.
  - Salakhutidnov, Ruslan. Zemel, Richard. Bengio, Yoshua.
  - Encoder-Decoder Model
  - Encoder: Convolutional Neural Network (CNN) extracts feature vectors from raw image
  - Decoder: RNN using Long Short-Term Memory Network (LSTM) architecture
    - Attention mechanism: The RNN output depends on the previously generated word
  - Google Translate API

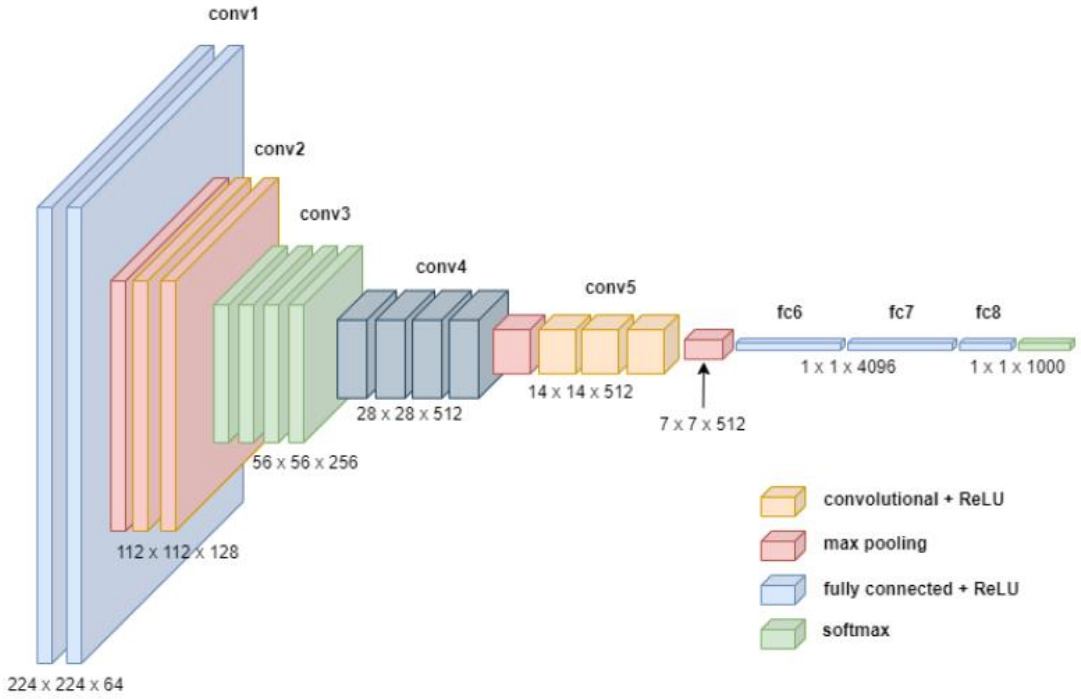


Image

Model Architecture

CNN – Transfer Learning

- Input: raw image
- Output: annotation vectors
- VGG16 – idea proposed by Karen Simonyan and Andrew Zisserman in 2013
- Extract features from lower convolution layer instead of fully connected layer
  - 14 x 14 x 512
  - Acts an encoder in an encoder-decoder model



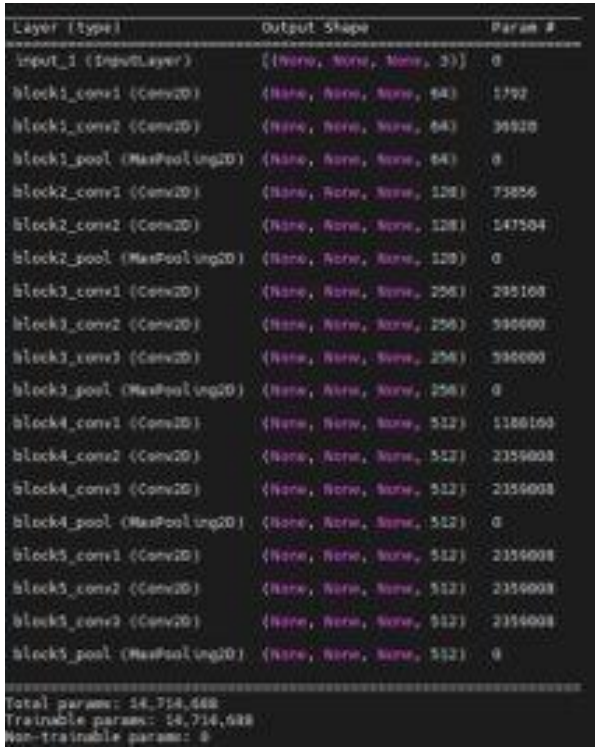
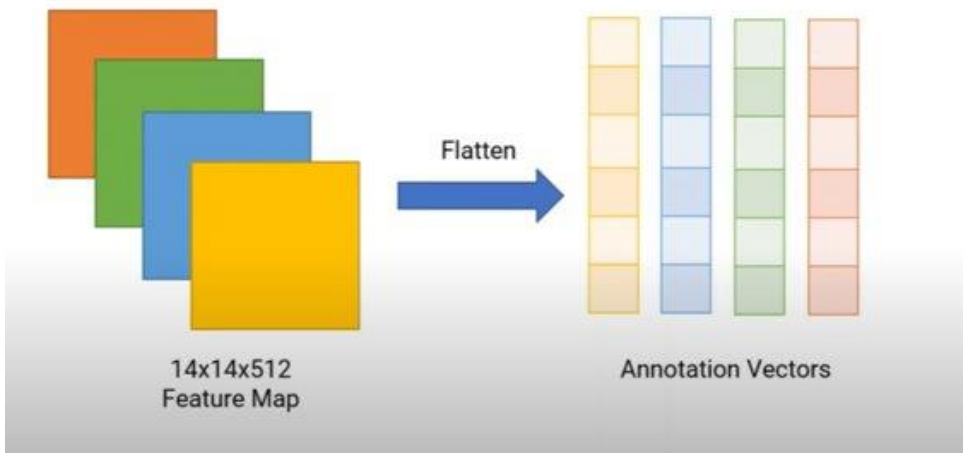


Image source: VGG-16 Architecture rendered using diagrams.net

Annotation Vectors Encoded into Matrix



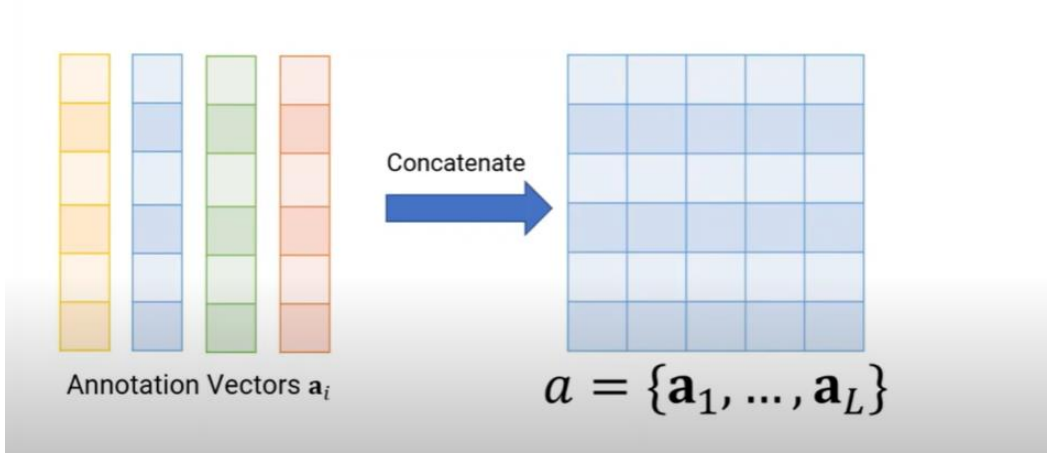
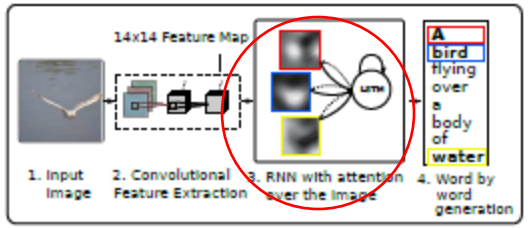


Image source: Show, Attend and Tell Explained

- Matrix *a* is the output of our CNN encoder
- Decoder will use matrix to identify sections of the image most important for predicting the next word



- Eyt-1 : Previously generated word multiplied by embedding matrix
- Ht-1: Previous decoder hidden state
- zt: Context vector representing relevant part of the input image at time t

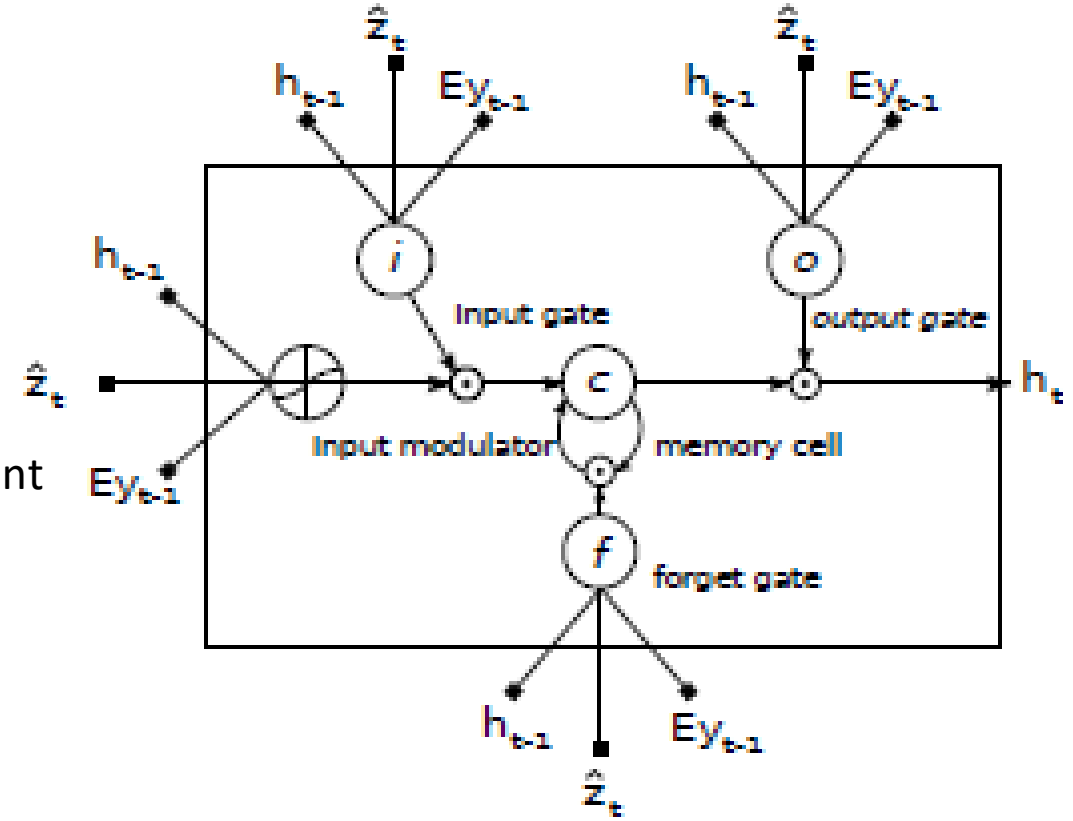


Image Source: Show, Attend and Tell Paper

RNN Decoder

Energy and Attention

Probability that location i is the right place to focus to predict the next word

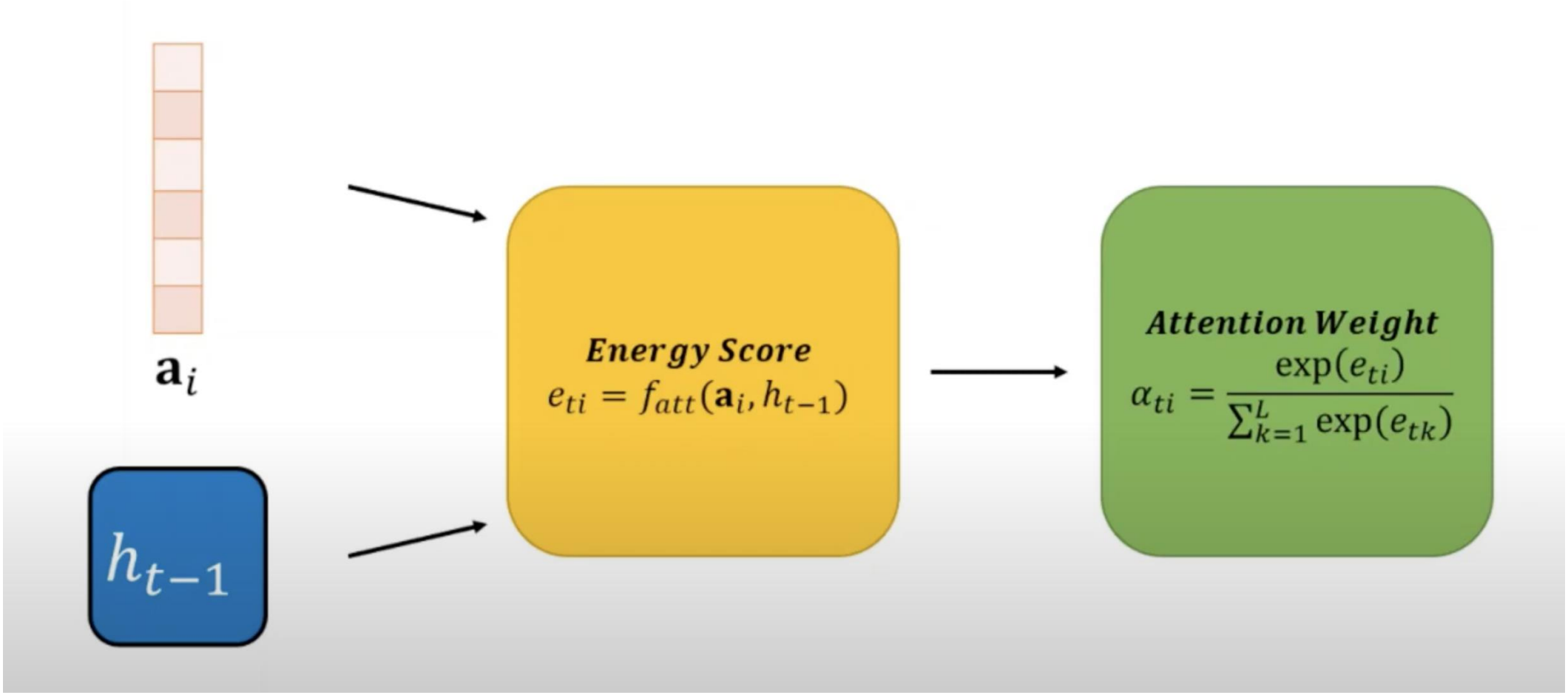


Image source: Show, Attend and Tell Explained

Hard vs. Soft Attention

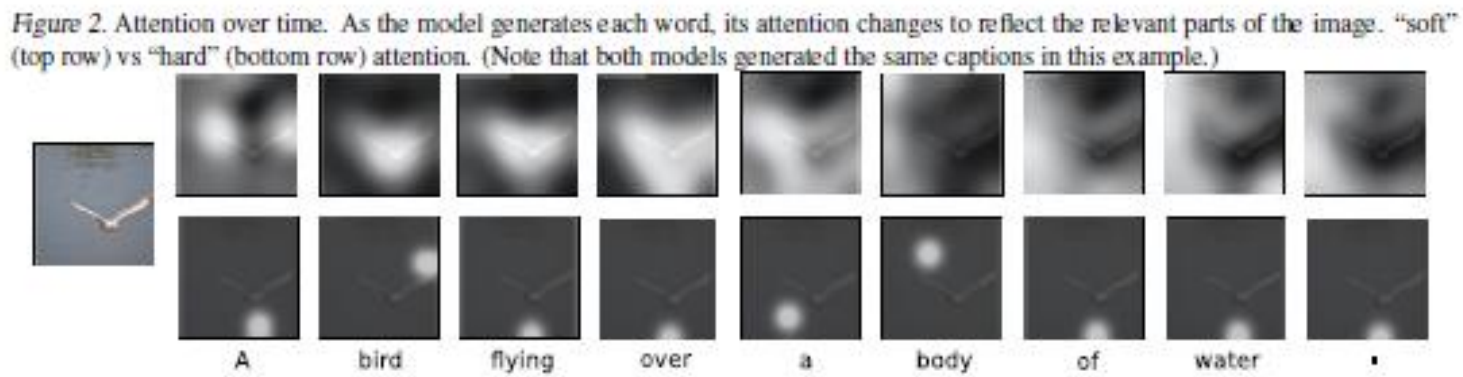


Figure 2. Attention over time. As the model generates each word, its attention changes to reflect the relevant parts of the image. "soft" (top row) vs "hard" (bottom row) attention. (Note that both models generated the same captions in this example.)

- Stochastic Hard Attention

$$Z \sim x_i, \alpha_i$$

  - Input x_i with a sample rate α_i into LSTM
- Deterministic Soft Attention

$$Z = \sum_i \alpha_i x_i$$

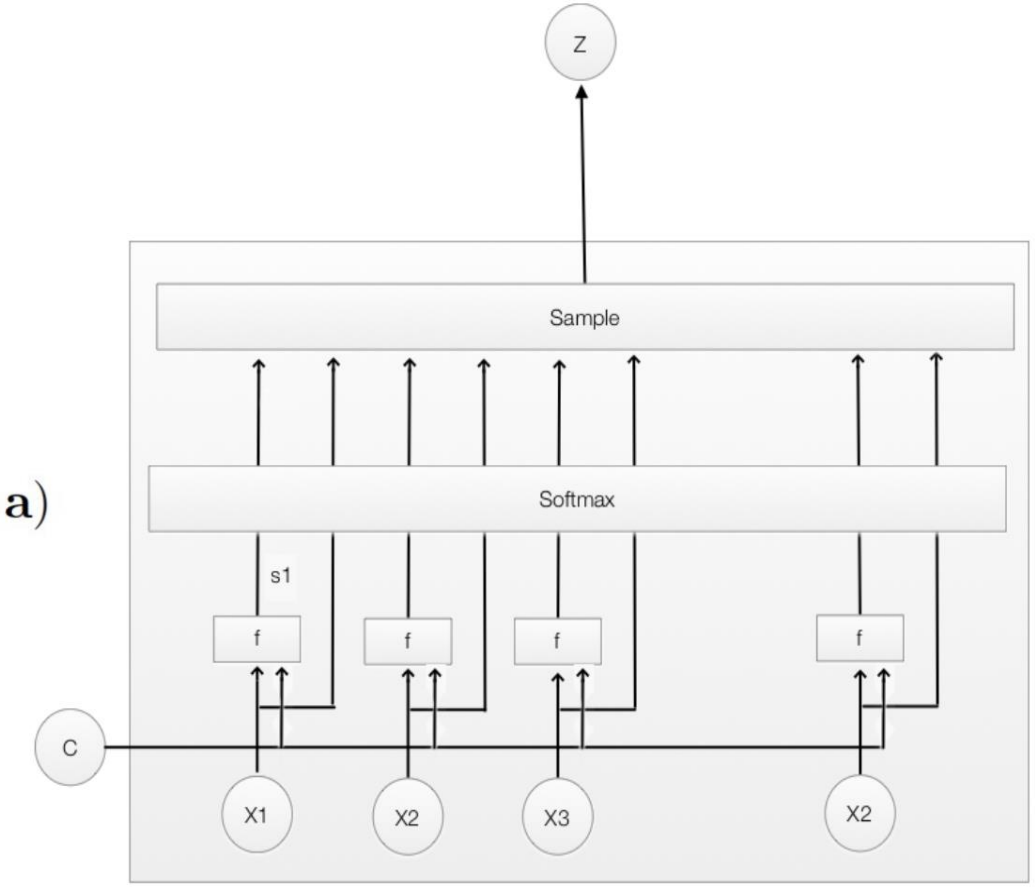
  - Input the sum of $x_i * \alpha_i$ with weighted averages α_i into LSTM

Image Source: [Show, Attend and Tell Paper]

Stochastic Hard Attention

- Contex Vector z: $\hat{\mathbf{z}}_t = \sum_i s_{t,i}\mathbf{a}_i.$
- Learning Algorithm: $L_s = \sum_s p(s \mid \mathbf{a}) \log p(\mathbf{y} \mid s, \mathbf{a})$
- Gradient of Ls: $\frac{\partial L_s}{\partial W} \approx \frac{1}{N}\sum_{n=1}^{N}\left[\frac{\partial \log p(\mathbf{y} \mid \tilde{s}^n, \mathbf{a})}{\partial W} + \log p(\mathbf{y} \mid \tilde{s}^n, \mathbf{a})\frac{\partial \log p(\tilde{s}^n \mid \mathbf{a})}{\partial W}\right]$

where $\tilde{s}^n \sim \text{Multinoulli}_L(\{\alpha_i\})$



Stochastic Soft Attention

- Score function s_i :

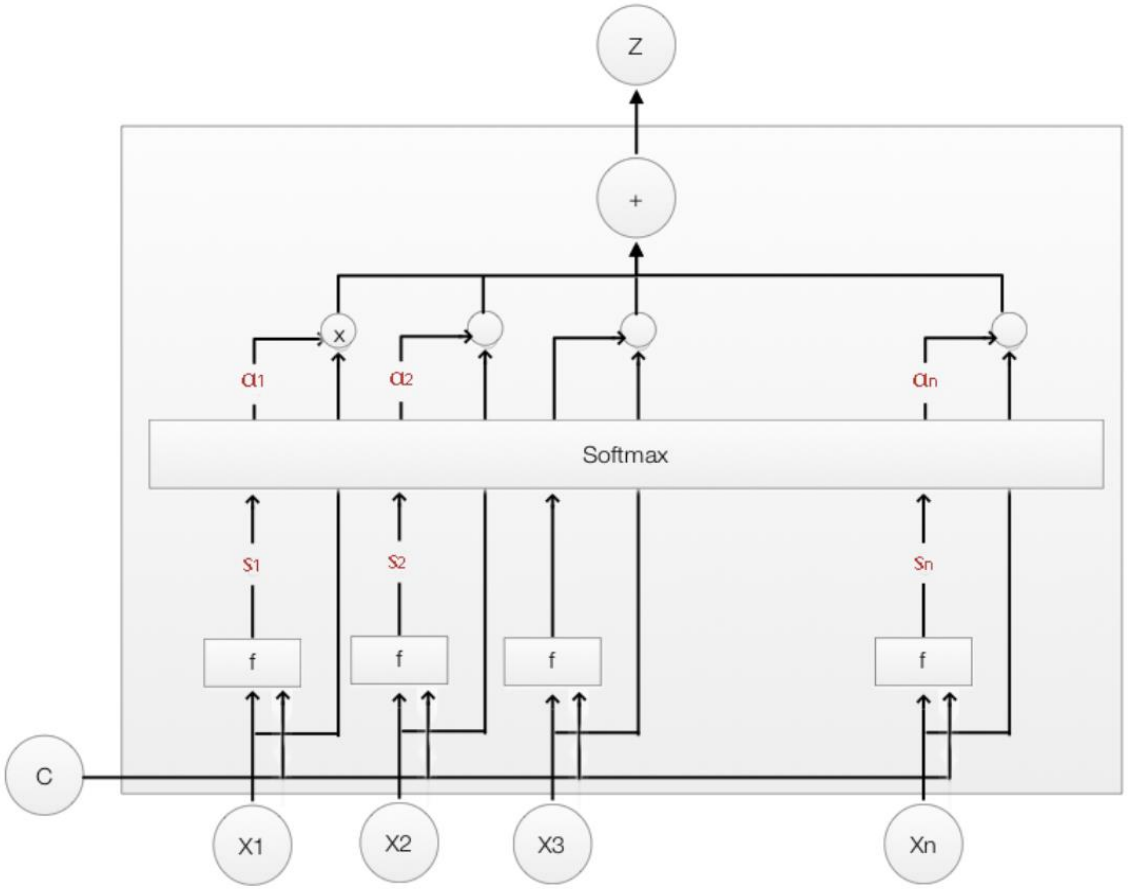
  $s_i = \tanh(W_c C + W_x X_i) = \tanh(W_c h_{t-1} + W_x x_i)$

  measures how much attention is applied to x_i

- Weight α_i: $\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{L} \exp(e_{tk})}.$

  Where α_i sum up to 1

- Context Vector z: $\mathbb{E}_{p(s_t|a)}[\hat{\mathbf{z}}_t] = \sum_{i=1}^{L} \alpha_{t,i} \mathbf{a}_i$

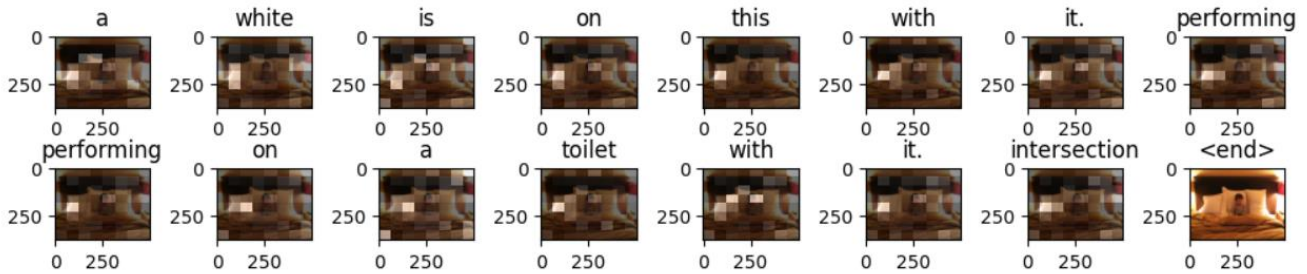


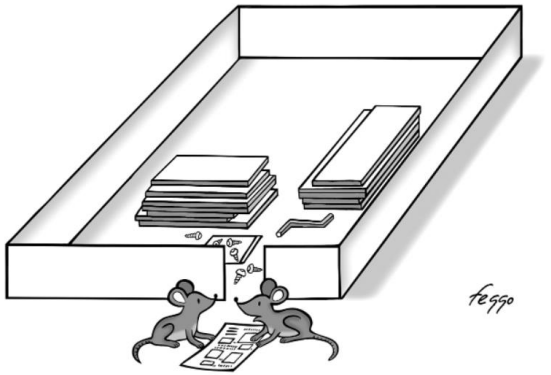
Training and Evaluation

- Number of epochs = 50,
- COCO 2014 Dataset
  - *Training* Images: 40,018
  - *Training* Captions: 200,090
  - *Test* Images: 10,003
  - *Test* Captions: 50,015

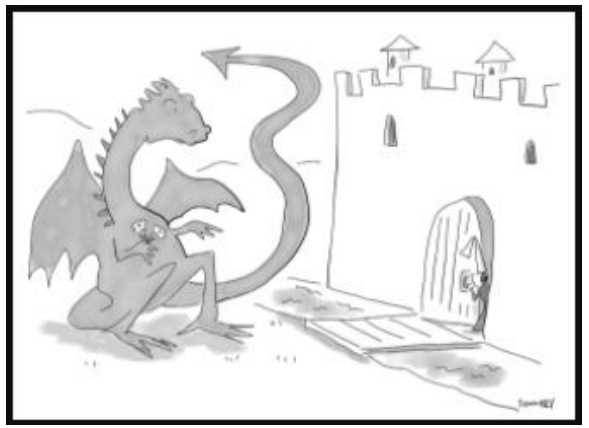
- Evaluation using New Yorker cartoons
- Translations by Google Translate API

- Image Source: https://github.com/nextml/caption-contest-data/tree/gh-pages/cartoons



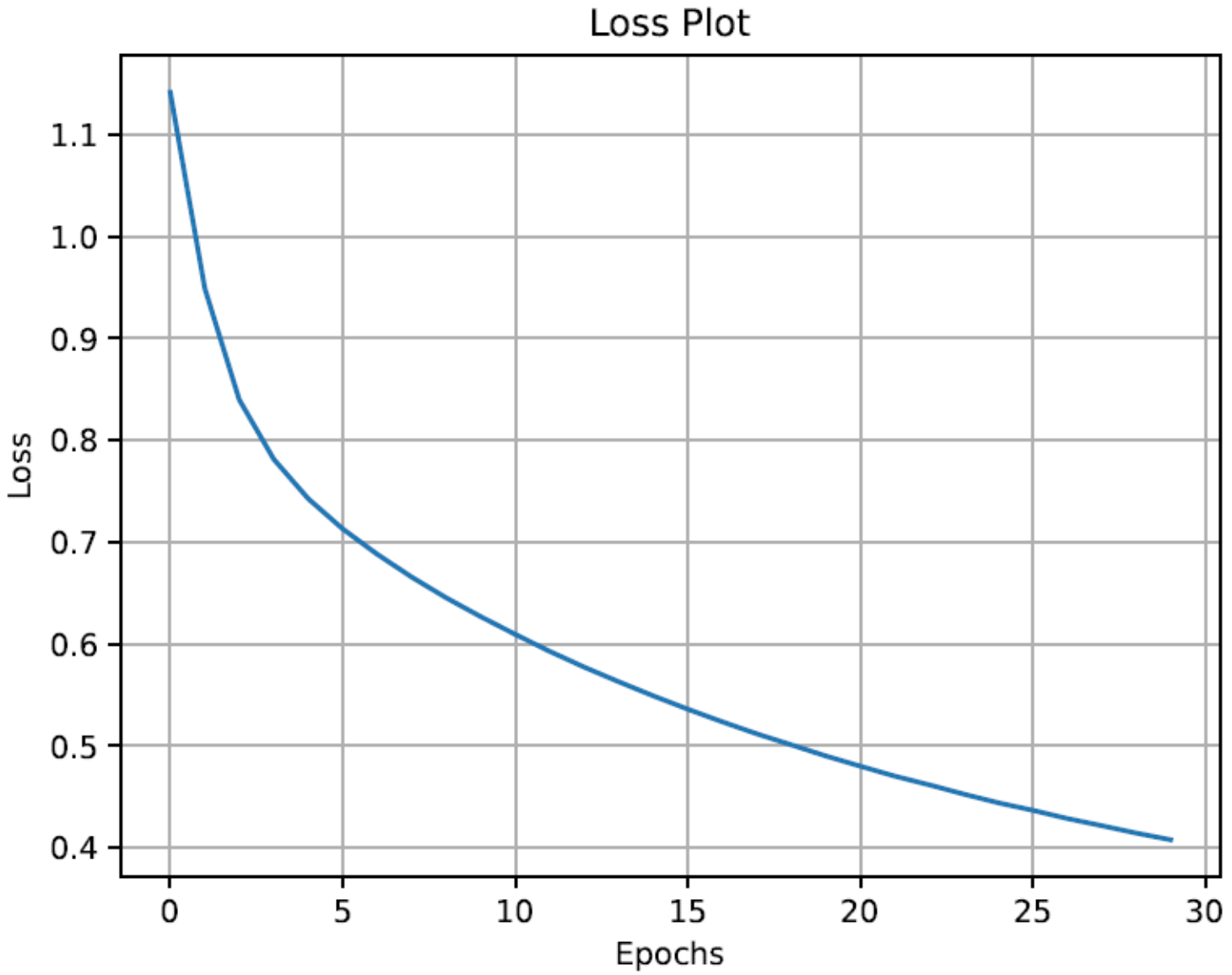


Generated Caption: A table [UNK] a table and a box placed on it. <end>
German Translation:
Ein Tisch [UNK] ein Tisch und eine darauf gestellte Kiste. <Ende>



Generated Caption: Many cows lounging at sheep goat <end>
German Translation:
Viele Kühe räkeln sich bei der Schafziege <Ende>

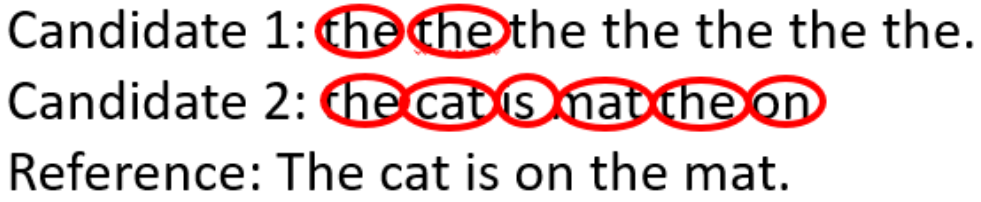
Adam Optimizer



Results

BLEU Score

- BLEU: Bilingual Evaluation Understudy
- Evaluate quality of machine-generated text against reference text
- Compares n-grams of candidate with n-grams of reference and counts matches
- Score between 0 and 1



- Source: https://aclanthology.org/P02-1040.pdf

Image source: https://towardsdatascience.com/bleu-bilingual-evaluation-understudy-2b4eab9bcfd1

BLEU Formula

- Modified n-gram precision: clip counts of n-grams to the number of occurrences in the reference
- Source: http://kv-emptypages.blogspot.com/2017/04/the-problem-with-bleu-and-neural.html

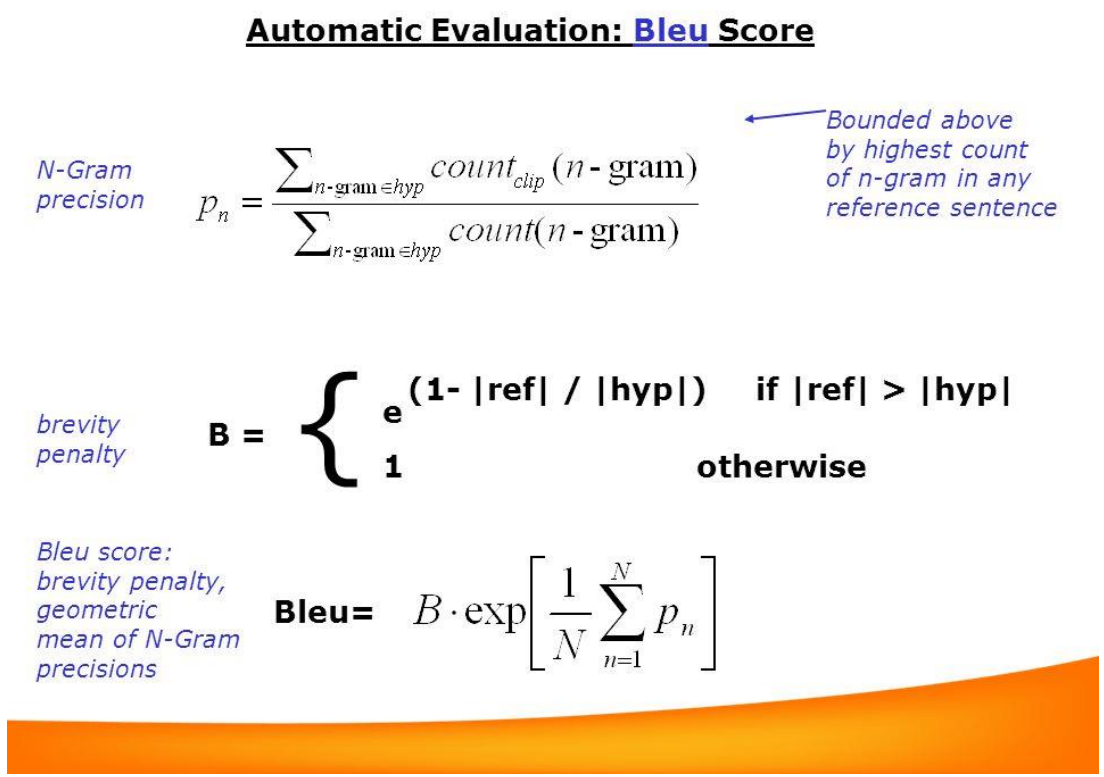


Image source: http://kv-emptypages.blogspot.com/2017/04/the-problem-with-bleu-and-neural.html

	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Our model	46.4	23.9	12.3	6.3
Soft-attention	70.7	49.2	34.4	24.3
Hard-attention	71.8	50.4	35.7	25.0

Conclusions

Summary

- We built a working model using visual attention, trained on COCO dataset, taking in a raw image and creating a caption
- The visual attention mechanism resembles what humans do on an intuitive level
- Tested our models against New Yorker cartoons with interesting captions
- BLEU score for model performed 50% below the results in Show, Attend and Tell paper
- The model is slow to train ~ over 2.5 minutes per epoch
- Google API Translate works well but not suitable for scaling large volumes of captions

References

- *COCO - Common Objects in Context*. (n.d.). COCODataset. Retrieved April 11, 2022, from https://cocodataset.org/
- Hui, Jonathan. *"Soft & Hard Attention"*, 15 Mar. 2017, https://jhui.github.io/2017/03/15/Soft-and-hard-attention/
- *Image captioning with visual attention | TensorFlow Core*. (2022, January 26). TensorFlow. Retrieved April 15, 2022, from https://www.tensorflow.org/tutorials/text/image_captioning
- Makarov, Artyom. "Image Captioning with Attention: Part 1." *Medium*, Analytics Vidhya, 14 Dec. 2020, https://medium.com/analytics-vidhya/image-captioning-with-attention-part-1-e8a5f783f6d3
- *Section 508 (Federal Electronic and Information Technology)*. (n.d.). U.S. Access Board. Retrieved April 25, 2022, from https://www.access-board.gov/law/ra.html
- *Show, Attend And Tell - Paper Explained*. (2021, May 3). YouTube Halfling Wizard. Retrieved April 16, 2022, from https://www.youtube.com/watch?v=y1S3Ri7myMg&t=6s
- Sarkar, Subham. "Image Captioning Using Attention Mechanism." *Medium*, The Startup, 15 June 2021, https://medium.com/swlh/image-captioning-using-attention-mechanism-f3d7fc96eb0e.
- Xu, Ba, Kiros, K. J. R. (2016, April 19). *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention*. Arxiv.Org. Retrieved April 4, 2022, from https://arxiv.org/pdf/1502.03044.pdf?msclkid=97863852c49311ecb15903abb2f078b2
- Khandelwal, R. (2021, December 13). *BLEU — Bilingual Evaluation Understudy - Towards Data Science*. Medium. Retrieved April 26, 2022, from https://towardsdatascience.com/bleu-bilingual-evaluation-understudy-2b4eab9bcfd1
- Brownlee, J. (2019, December 18). *A Gentle Introduction to Calculating the BLEU Score for Text in Python*. Machine Learning Mastery. Retrieved April 26, 2022, from https://machinelearningmastery.com/calculate-bleu-score-for-text-python/
- Papineni, Roukos, Ward, Zhu (2022, July 1). BLEU: a Method for Automatic Evaluation of Machine Translation. Dl.acm.org. Retrieved April 26, 2022, from https://aclanthology.org/P02-1040.pdf
- Vashee, Kirti. (2017, April 11). *The Problem with BLEU and Neural Machine Translation*. Blogger.Com. Retrieved April 26, 2022, from http://kv-emptypages.blogspot.com/2017/04/the-problem-with-bleu-and-neural.html