UEMF
الجامعة الأورومتوسطية بفاس
EUROMED UNIVERSITY OF FES
UNIVERSITÉ EUROMED DE FÈS

EIDIA
École d'Ingénierie Digitale et d'Intelligence Artificielle

End of the Year Project

Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

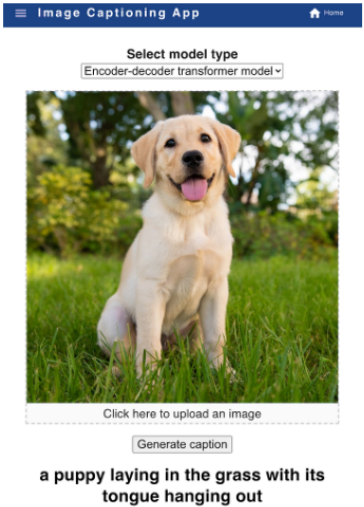
Firdawse Guerbouzi

Supervisor: Mr.Alae Ammour

Outlines

Objectives

- Develop a deep learning model for accurate and descriptive image caption generation.
- Improve visual understanding within the model.
- Enhance caption fluency and coherence.
- Evaluate model performance using appropriate metrics.
- Create a user interface for image captioning.



A 5W2H Analysis

WHY ?
Improve Accessibility and Information Extraction from images .

WHAT?
Develop an Image Captioning System .

WHO?
Project Team

HOW?
Using Deep Learning Techniques

WHERE?
During THE PFA module

WHEN?
In 3 months

HOW MUCH?
Personal Computer: CPU - Intel Core i7-10750H, RAM - 24GB, GPU - NVIDIA GeForce RTX 3060 Ti

Ressources allocated

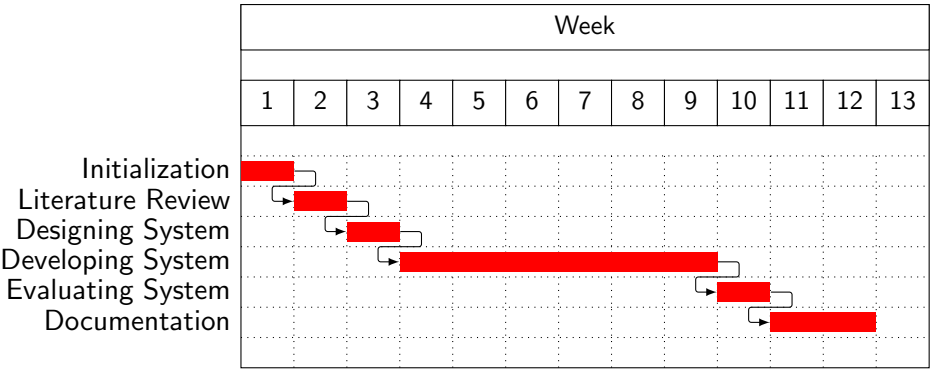
Hardware:

- CPU: Intel Core i7-10750H
- RAM: 24GB
- GPU: NVIDIA GeForce RTX 3060 Ti

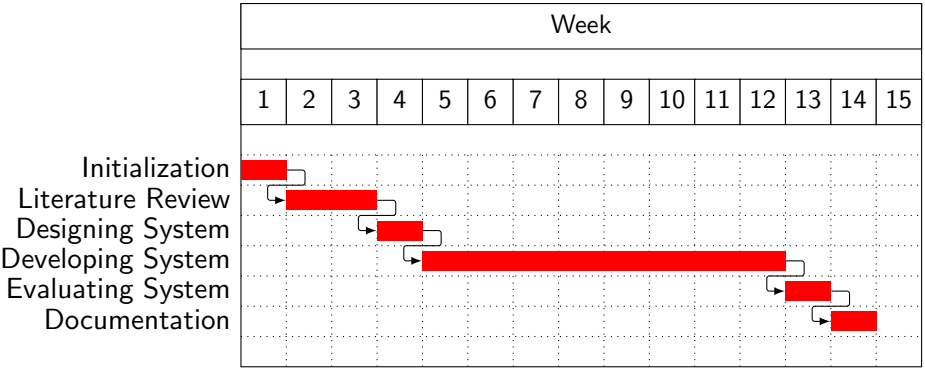
Software:

- Operating System: Windows 10
- Programming Language: Python 3.10
- Deep Learning Frameworks: TensorFlow 2.10, Keras 2.10
- Development Environment: Jupyter Notebook

Planned Gantt Chart

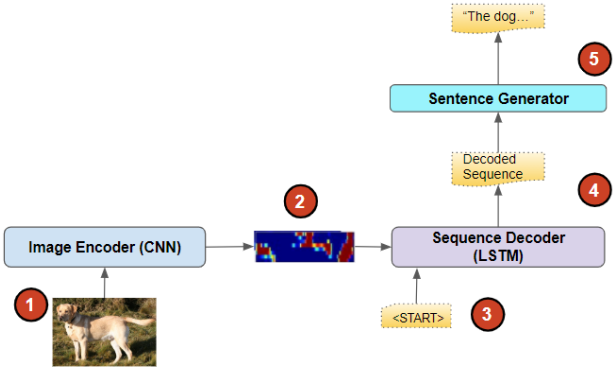


Actual Gantt Chart

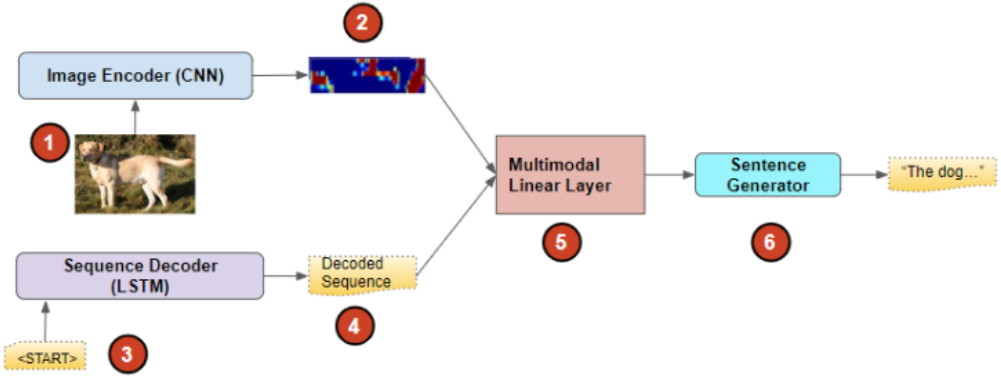


State of the art

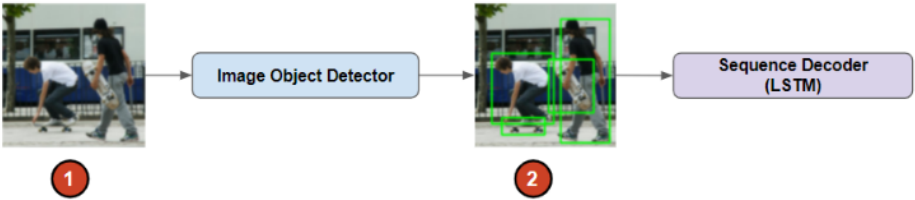
Architecture: Encoder-Decoder



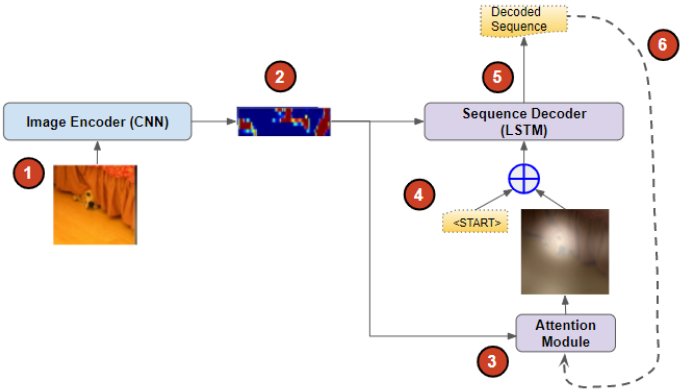
Architecture: Multi-Modal



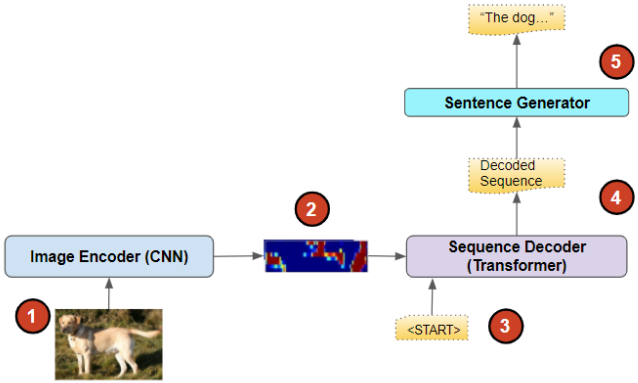
Architecture: Object Detection backbone



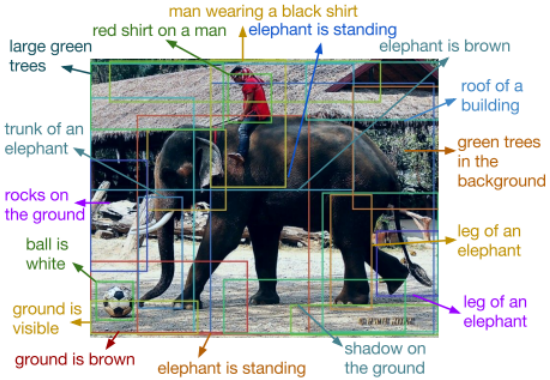
Architecture: Encoder-Decoder with Attention



Architecture: Encoder-Decoder with Transformers



Architecture: Dense Captioning



Regarding the inspirational research paper

Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

Kelvin Xu	KELVIN.XU@UMONTREAL.CA
Jimmy Lei Ba	JIMMY@PSI.UTORONTO.CA
Ryan Kiros	RKIROS@CS.TORONTO.EDU
Kyunghyun Cho	KYUNGHYUN.CHO@UMONTREAL.CA
Aaron Courville	AARON.COURVILLE@UMONTREAL.CA
Ruslan Salakhutdinov	RSALAKHU@CS.TORONTO.EDU
Richard S. Zemel	ZEMEL@CS.TORONTO.EDU
Yoshua Bengio	FIND-ME@THE.WEB

Methodology

Data Collection

the black dog jumped the tree stump .
a mottled black and grey dog in a blue collar jumping over a fallen tree .
a large black dog leaps a fallen log .
a grey dog is leaping over a fallen tree .
a black dog leaps over a log .

the white and brown dog is running over the surface of the snow .
a white and brown dog is running through a snow covered field .
a dog running through snow .
a dog is running in the snow
a brown and white dog is running through the snow .

man on skis looking at artwork for sale in the snow
a skier looks at framed pictures in the snow next to trees .
a person wearing skis looking at framed pictures set up in the snow .
a man skis past another man displaying paintings in the snow .
a man in a hat is displaying pictures next to a skier in a blue hat .

several climbers in a row are climbing the rock while the man in red watches and holds the line .
seven climbers are ascending a rock face whilst another man stands holding the rope .
a group of people climbing a rock while one man belays
a group of people are rock climbing on a rock climbing wall .
a collage of one person climbing a cliff .

Figure: Flickr dataset

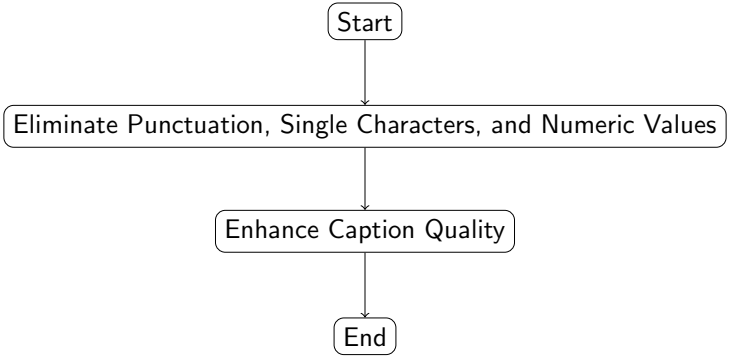
- Popular benchmark dataset for image captioning.
- Contains 8091 images sourced from Flickr.
- Each image has five human-generated captions.
- Provides diversity in subjects, scenes, and objects.

Data Collection

Figure: COCO dataset

- Popular benchmark dataset for image captioning.
- contains approximately 80000 images with rich annotations.
- The dataset covers various categories, including people, animals, vehicles, and indoor scenes.
- Each image is accompanied by multiple human-generated captions.

Caption Transformation



Caption Embedding

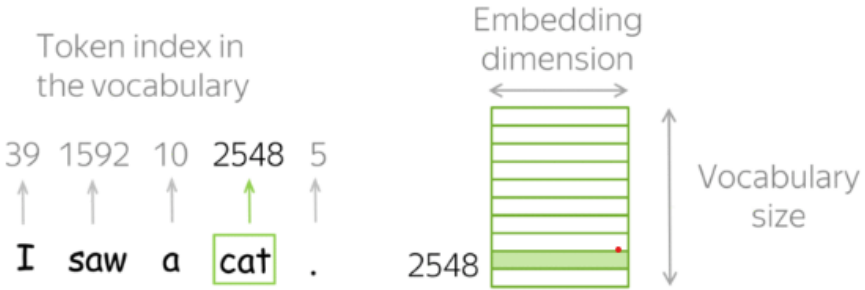


Image Resizing

- The images are resized to a fixed size of 224x224 pixels.
- This resizing step ensures compatibility with the VGG16 model, which expects inputs of this specific dimension.
- Consistency in input size is maintained across all images in the dataset.

Image Preprocessing

- The resized images undergo a preprocessing procedure.
- This involves operations such as mean subtraction and scaling to normalize the pixel values.
- The preprocessing steps align with those used during the training of the VGG16 model.

Deep Learning architecture

How does it work?

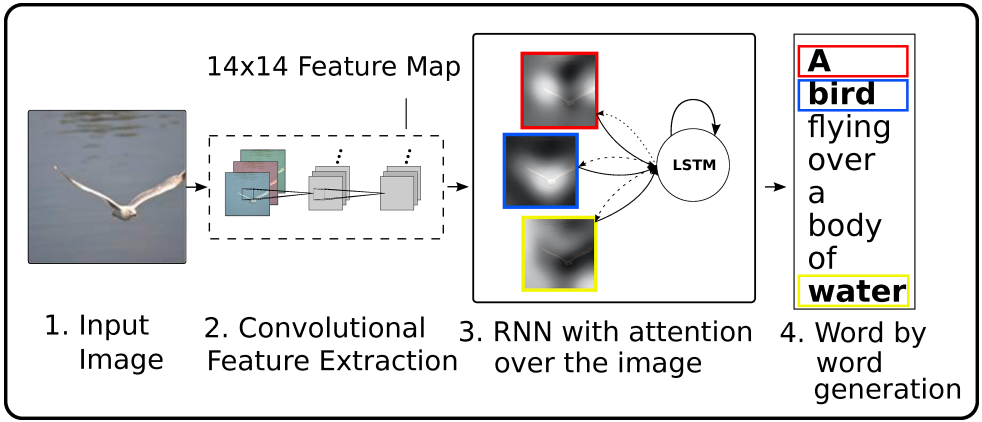


Image Feature Encoder

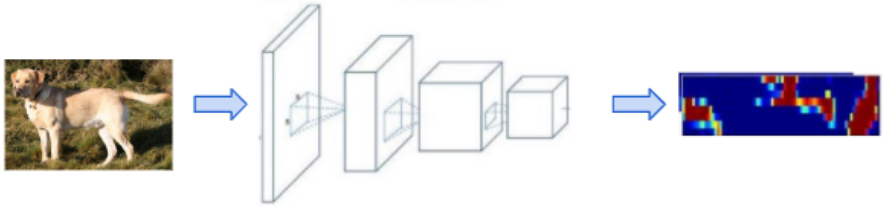


Image Feature Encoder

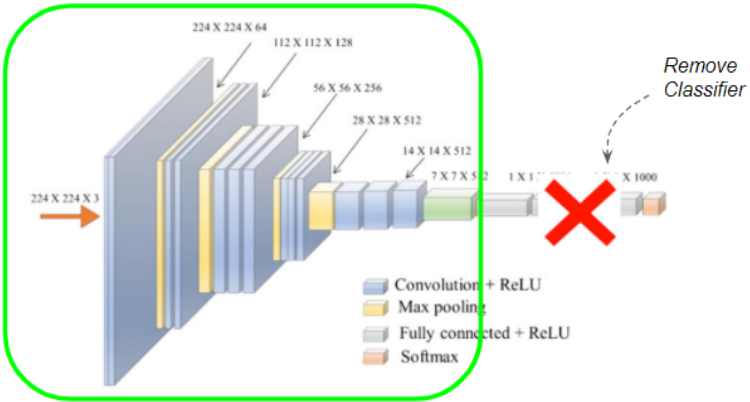


Image Feature Encoder

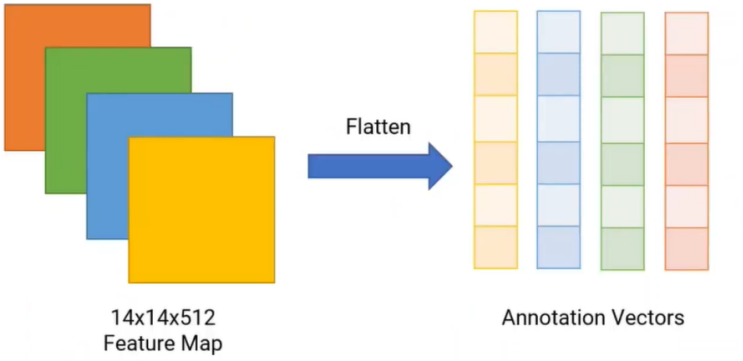


Image Feature Encoder

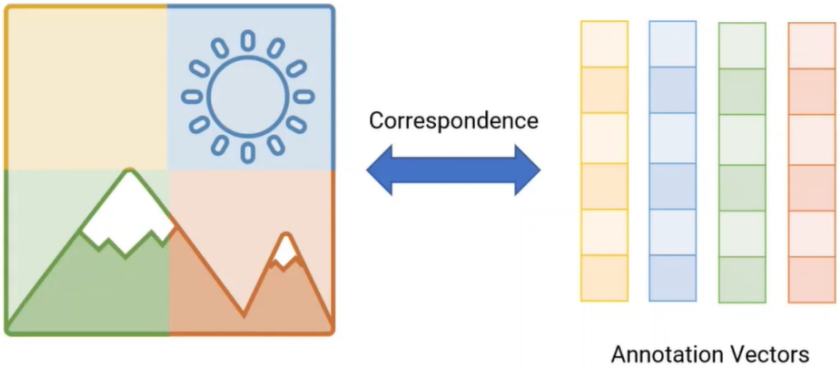
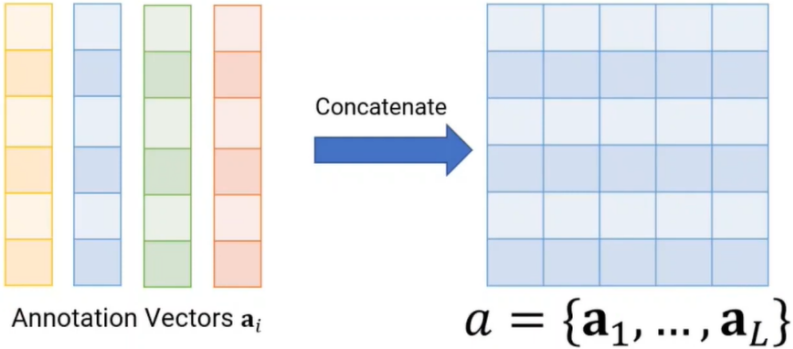
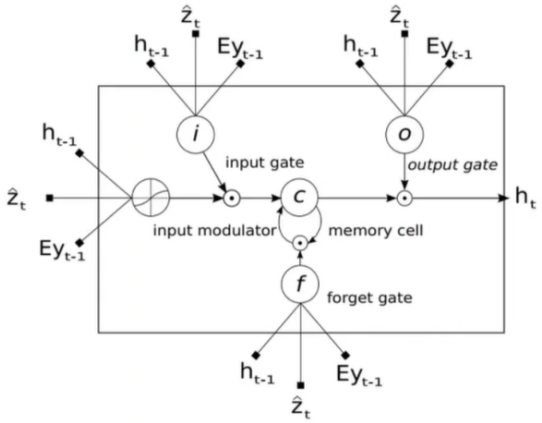


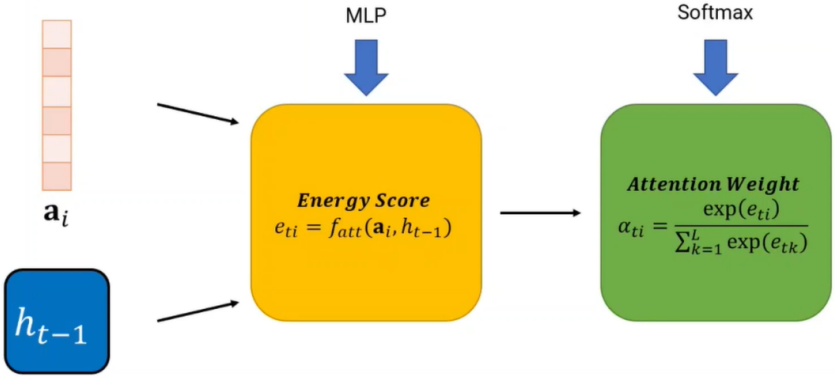
Image Feature Encoder



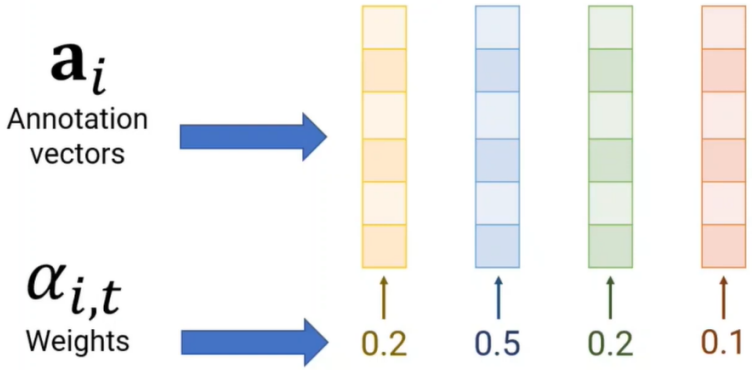
Sequence Decoder



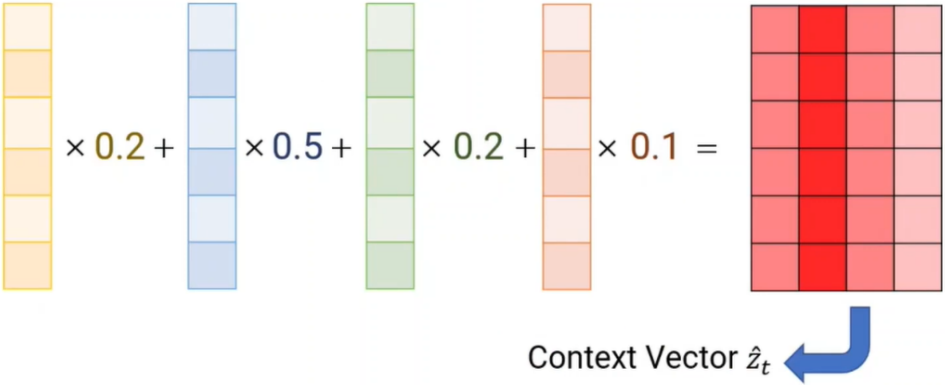
Attention mechanism



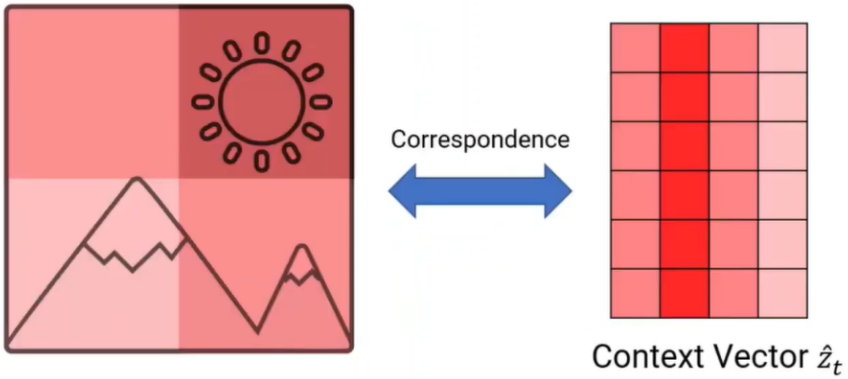
Attention mechanism



Soft attention



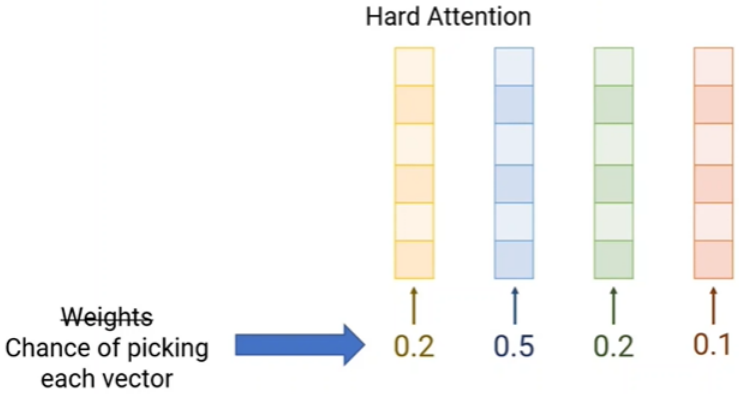
Soft attention



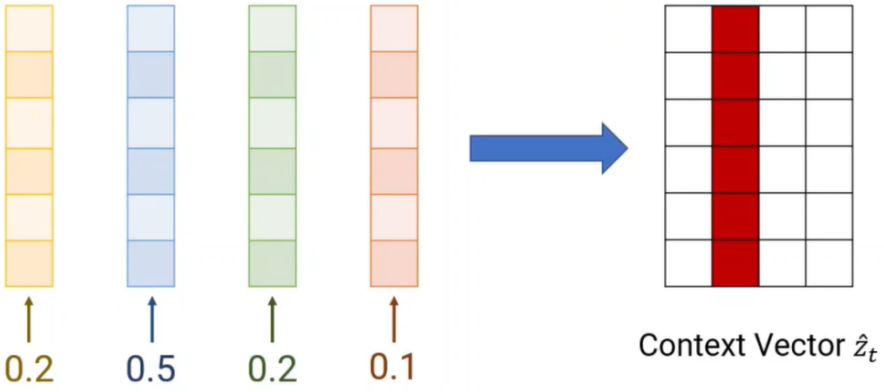
Soft attention

$$E[p(s_t, i = 1|a)] = \sum_{i=1}^{L} \alpha_{t,i} a_i$$

Hard attention



Hard attention

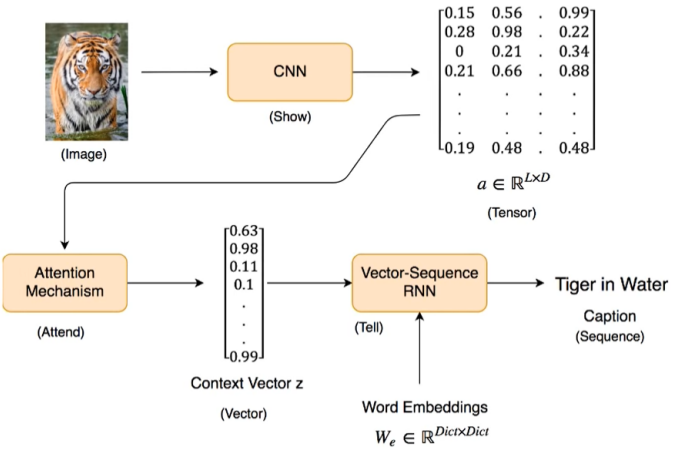


Hard attention

$$p(s_{t,i} = 1 | s_{j<t}, a) = \alpha_{t,i}$$

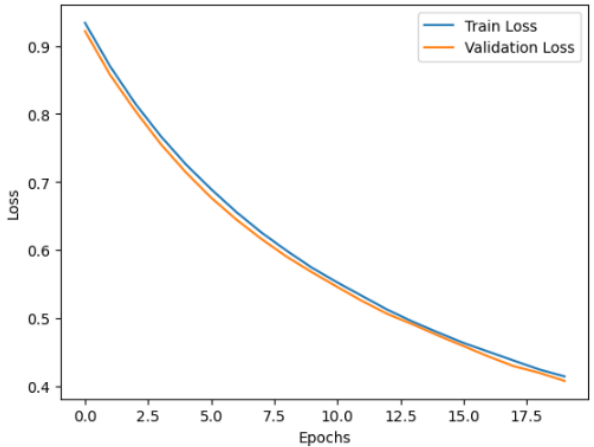
$$\hat{z}_t = \sum_{i=1}^{L} s_{t,i} \alpha_{t,i}$$

Our architecture so far

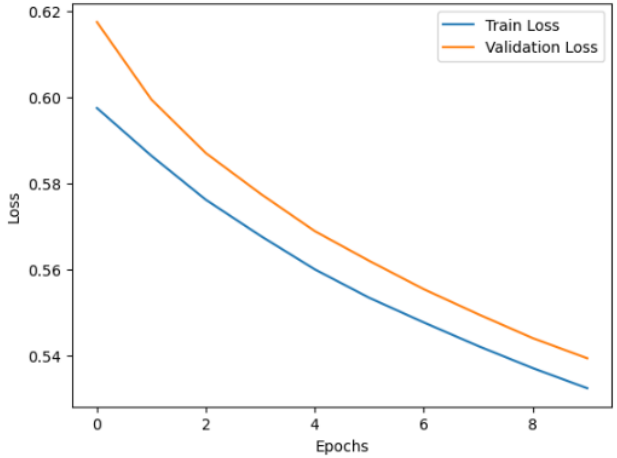


Quantitative Performance

Flickr dataset LOSS



COCO dataset LOSS



Cross-Entropy Loss Function

Cross-entropy is a loss function that is commonly used in classification problems. It measures the difference between the predicted probability distribution and the actual probability distribution.

$$L_{\text{CE}} = -\sum_{i=1}^{n} t_i \log(p_i), \text{ for n classes,}$$

where t_i is the truth label and p_i is the Softmax probability for the i^{th} class.

BLUE

Evaluation Metrics for Machine Translation:

- N-gram overlap between machine translation output and reference translation
- Compute precision for n-grams of size 1 to 4
- Add brevity penalty (for too short translations)

$$\text{BLEU} = \min\left(1, \frac{\textit{output-length}}{\textit{reference-length}}\right) \left(\prod_{i=1}^{4} \textit{precision}_i\right)^{\frac{1}{4}}$$

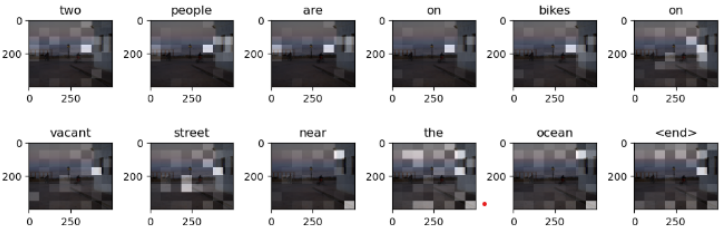
Performance Metrics

Table: Performance Metrics on Flickr8k and Microsoft COCO Test datasets

Dataset	BLEU Score	Sparse Categorical Crossentropy Loss
Flickr8k	0.75	0.407556
Microsoft COCO	0.68	0.502

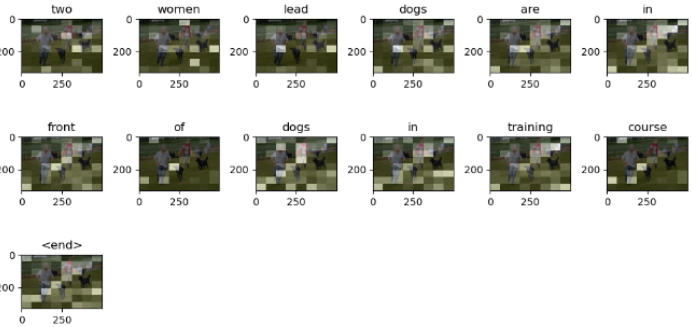
Qualitative Performance

```
BELU score: 31.702331385234306
Real Caption: two people are on bikes are next to white building
Prediction Caption: two people are on bikes on vacant street near the ocean
```



Out[215]:

```
Real Caption: two women <unk> are walking their dogs in dog show
Prediction Caption: two women lead dogs are in front of dogs in training course
```



Out[211]:

Interface

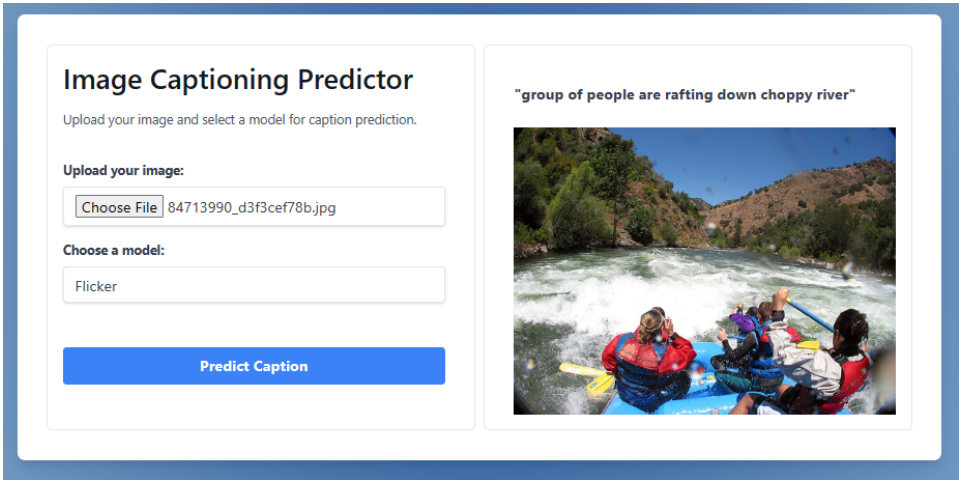
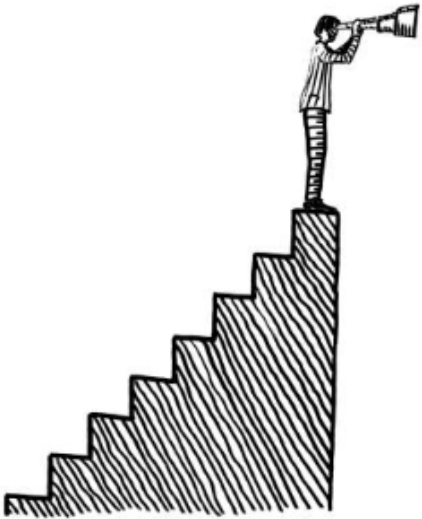


Image Captioning Predictor
Upload your image and select a model for caption prediction.
Upload your image:
Choose File
84713990_d3f3cef78b.jpg
Choose a model:
Flicker
Predict Caption
"group of people are rafting down choppy river"

Future Perspectives

- Embracing State-of-the-Art Transformers
- Focusing on a Specific Domain or Application



Thank you