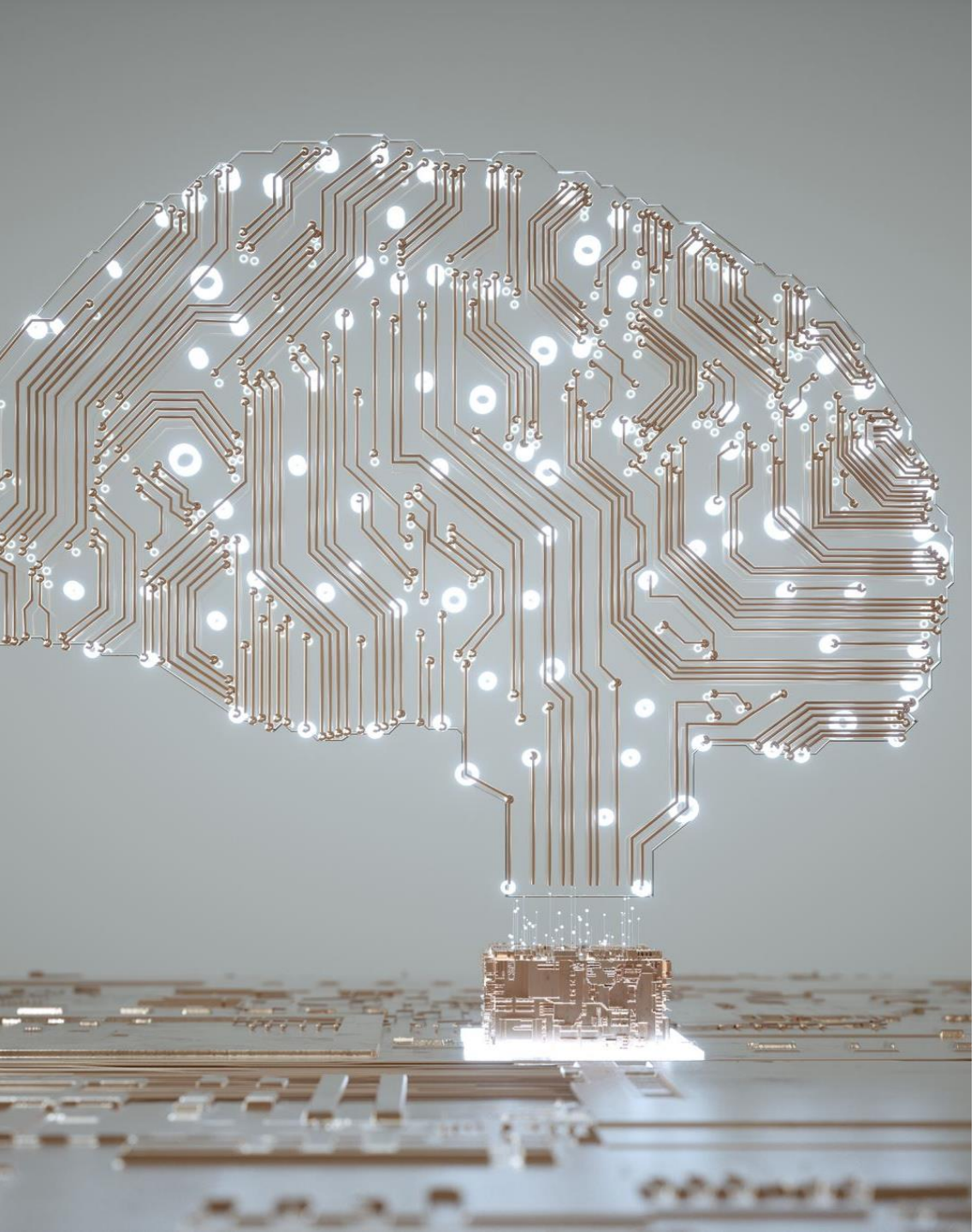
IMAGE-TEXT MATCHING USING STACKED CROSS ATTENTION FRAMEWORKS

SAKETH GAJAWADA - IMT2020531

AISHWARYA V KOUSHIK - MT2023506

UNDERSTANDING THE PROBLEM STATEMENT

- **Text-Image matching**
  - A technique used to align textual descriptions with the corresponding images
  - Text Representation : textual descriptions associated with the images are encoded into fixed-length vector representations using tools like word-embeddings
  - Image-Representation : images are transformed into fixed-length representations using CNNs or other image-encoding techniques
- **Attention Frameworks/Models**
  - A class of neural network architectures inspired by human vision
  - Allow models to focus on specific regions of the input (here, text and image) deemed most relevant to the task at hand
  - The models dynamically weigh the importance of different input elements while processing rather than treating all the features equally



Stacked Cross Attention

Stacked cross attention models consist of multiple layers of attention mechanisms to focus on relevant parts of text and image representations iteratively

- **Cross Attention** :
  - In every layer, the model computes the attention weights which indicates the importance of different parts of one modality with the other
- **Stacking** :
  - the increasingly complex relationships and correspondence between the text and images is achieved by stacking multiple layers of cross-attention
- **Alignment and Fusion**:
  - Features from both modalities are aligned using the computed attention weights and fused together which gives a joint representation of the correspondence
- **Prediction** :
  - The joint representation is fed to a classifier/regressor model that performs image-text retrieval

MOTIVATION

- **Importance of grounding language to vision:**
  - There is a need to ground natural language expressions to their corresponding visual representations for true understanding of multimodal data.
- **Interpretability of multimodal models:**
  - Existing image-text matching models act like black boxes, without explaining the reasoning behind their predictions. Inferring explicit alignments between words and visual regions can make these models more interpretable.
- **Limitations of previous approaches:**
  - Prior work either ignored differential importance of words/regions or could only capture one semantic alignment at a time through multi-step attention, limiting its applicability.
- **Varying alignments across data:**
  - The number of semantic correspondences between words and visual concepts varies across different image-text pairs. A flexible model is needed to handle this data variation
- **Leveraging attention for alignments:**
  - The success of attention mechanisms in vision and language domains motivates leveraging attentional reasoning to infer the latent alignments between the two modalities.

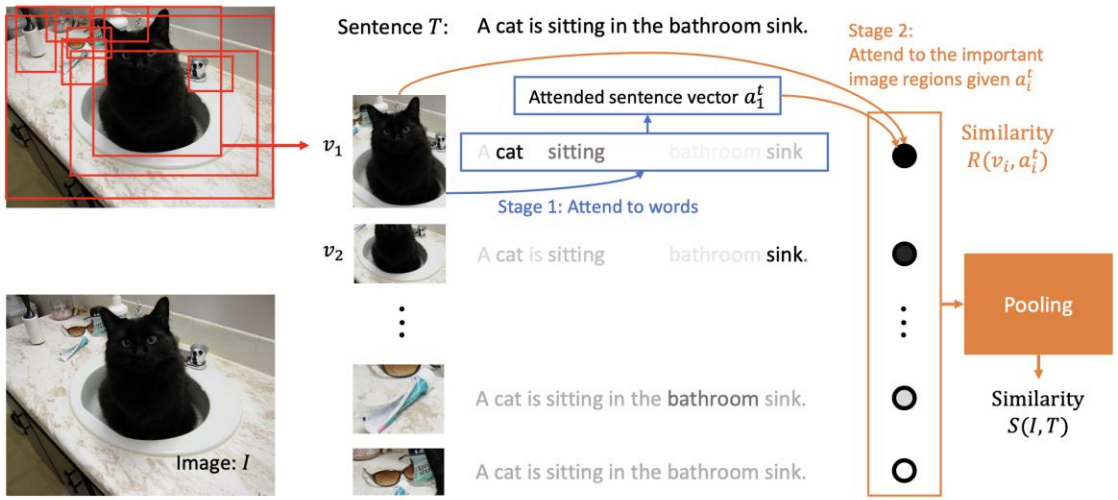
LITERATURE REVIEW

- **Image-Text Matching with Bottom-Up Attention:**
  - Karpathy and Fei-Fei (2015) detected image regions using R-CNN and aggregated region-word similarity scores without attention.
  - Niu et al. (2017) mapped noun phrases and objects to a shared embedding space on top of whole image/sentence embeddings.
- **Whole Image/Sentence Embedding Methods:**
  - Early work like Kiros et al. (2014), Faghri et al. (2017) mapped whole images and sentences to a common space without considering region-word alignments explicitly.
  - The paper positions its region-word alignment approach as being more interpretable compared to these embedding-based models.
- **Conventional Attention-based Methods:**
  - Nam et al. (2017) proposed a dual attention network capturing interplay between vision and language through multiple steps. But these models focus on one semantic alignment at a time with a predefined number of steps.

SCAN : PAPER COMPREHENSION

Paper comprehension involves understanding the salient features of SCAN, some of them being -

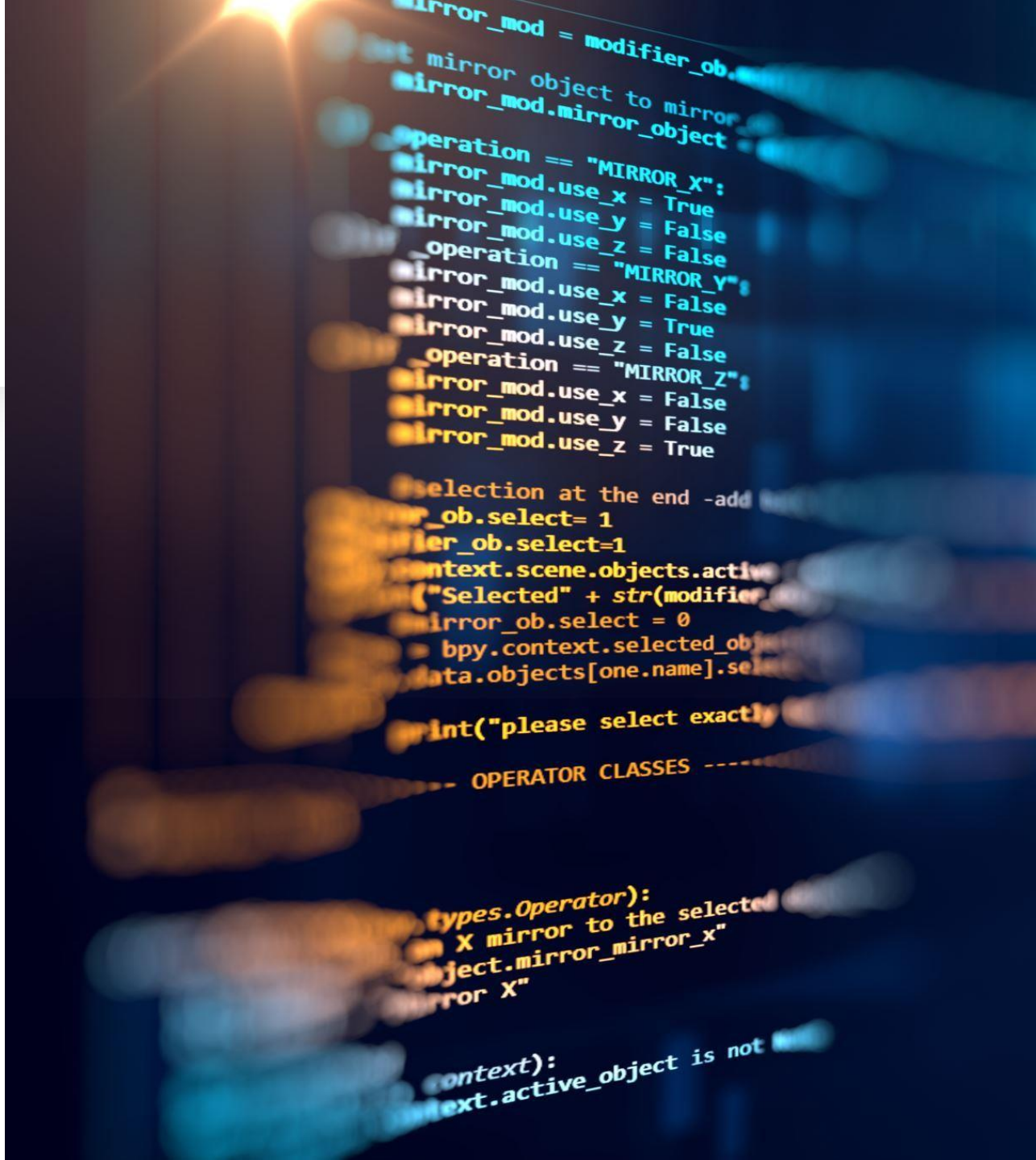
- Introduction to the Novel Stacked Cross Attention Mechanism and Two complementary Attention Flows which includes Text-Image and Image-Text attention
- A deep dive into Bottom-Up Visual attention
  - Uses a bottom-up approach based on faster R-CNN to detect important image regions at the object/stuff level.
- Got key insights on Joint Embedding Space:
  - Maps both image regions and words into a common joint embedding space.
  - Allows scoring cross-modal similarity between vision and language representations directly.



DECODING THE CODE

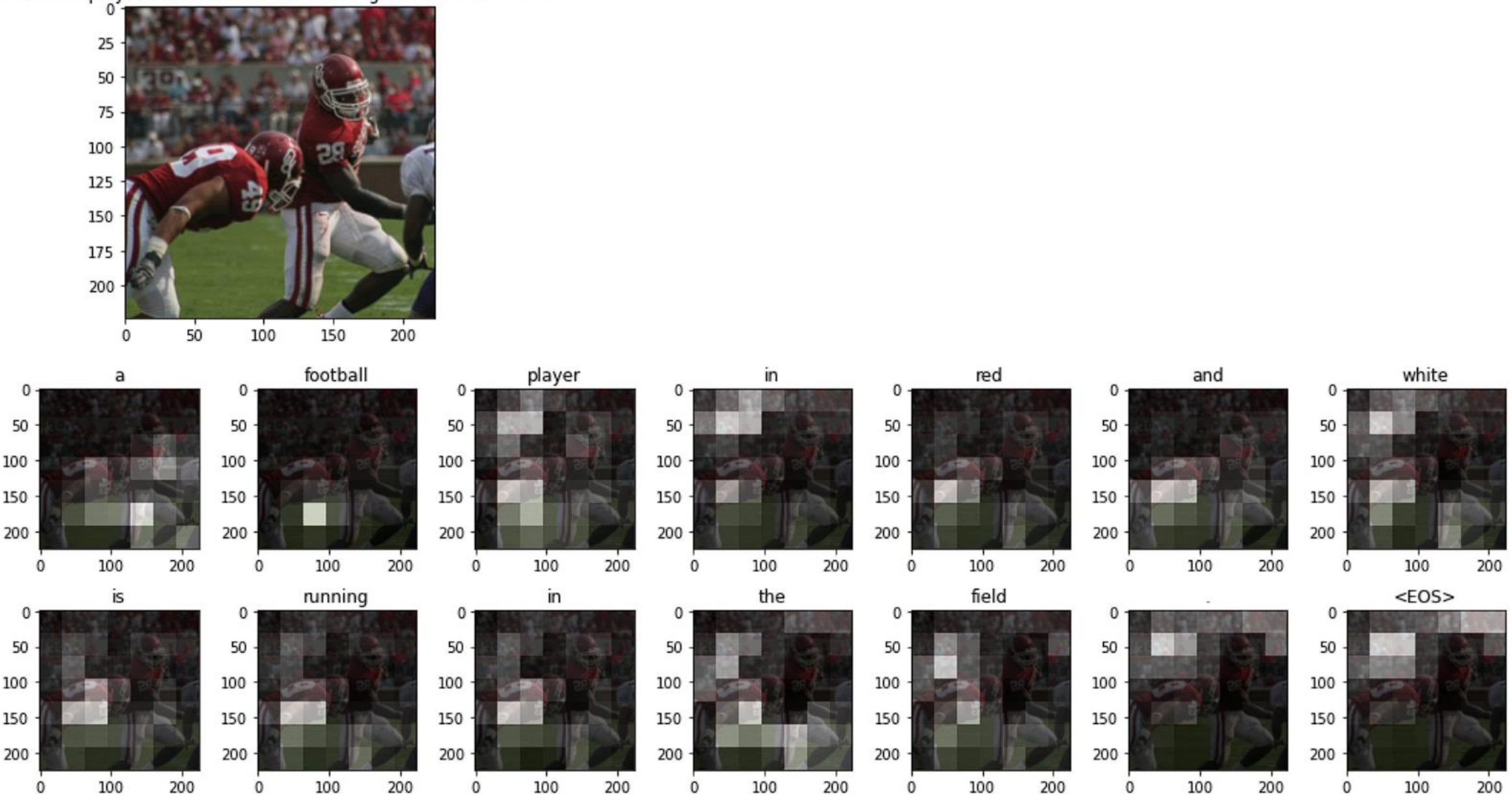
Code fragments :

- To start with, all the necessary libraries have been installed onto the system
- CapsCollate
- Flickr Dataset
- Vocab and Spacy
- Modified ResNet-50 architecture to accommodate our attention model
- Data loader
- Batch Processing and Loss Function
- Logging and Evaluation

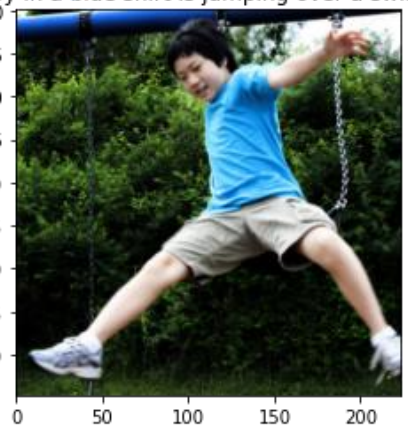


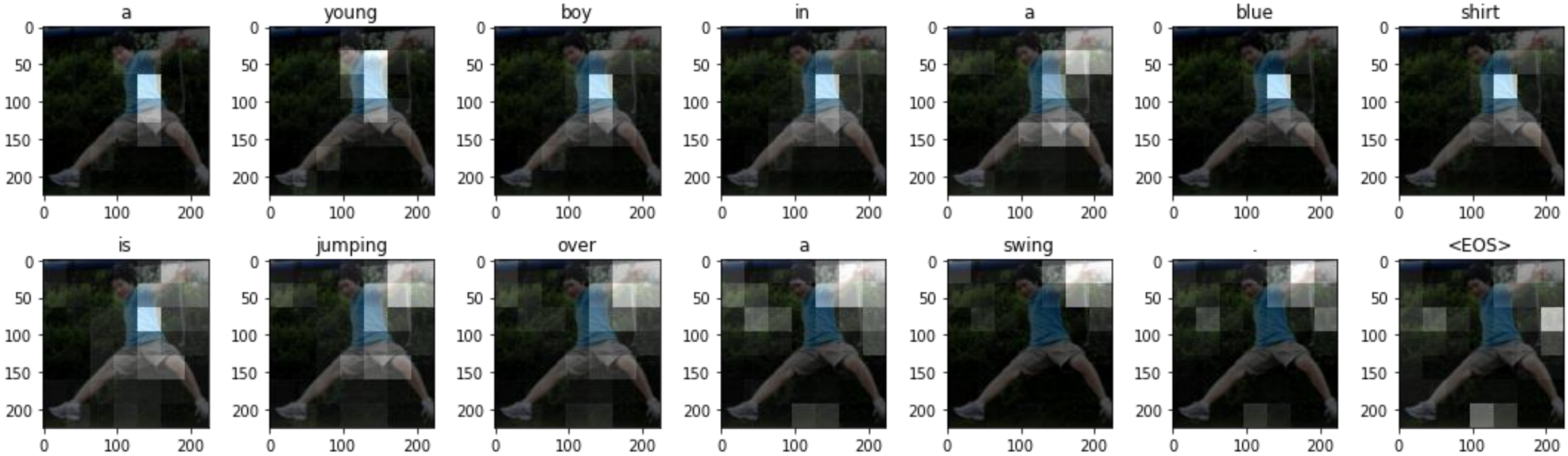
RESULTS – ATTENTION HEATMAPS FOR IMAGE-TEXT MATCH

a football player in red and white is running in the field . <EOS>



a young boy in a blue shirt is jumping over a swing . <EOS>

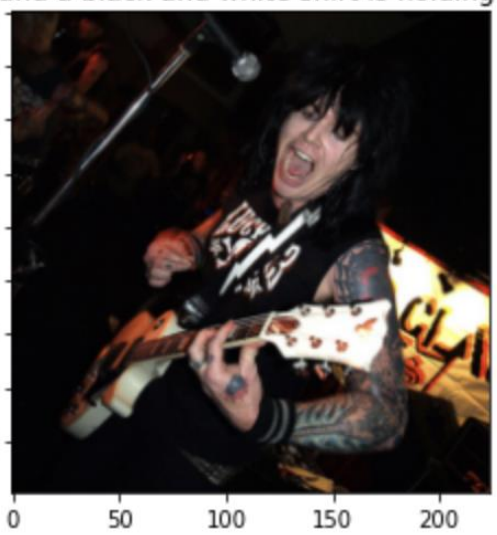




RESULTS : EPOCHS AND LOSS

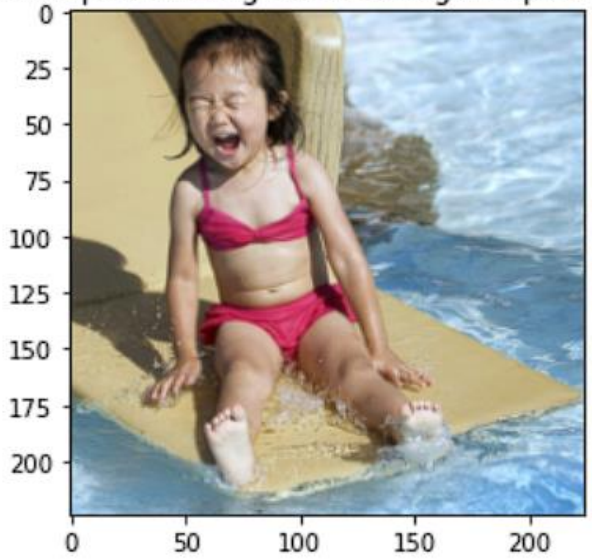
```
Epoch: 15 loss: 2.26783
```

a man in a red shirt and a black and white shirt is holding a guitar on a stage .




```
Epoch: 25 loss: 2.03723
```

a girl in a pink bathing suit is sitting in a pool . <EOS>



CONCLUSION AND FUTURE SCOPE

- The paper uses the proposed attention frameworks and achieves state-of the-art results for image retrieval given text query and vice versa.

- A few potential areas where we can expect significant advancement -
  - Exploring better ways to represent and attend to image regions beyond just object detection ie, incorporating relationships between objects
  - Models that accommodate visual question-answering, captioning etc where fine-grained alignments between vision and language could be beneficial
  - Investigating more powerful architectures like transformers in place of CNNS/RNNs
  - Scalability for larger image-caption datasets

THANK YOU