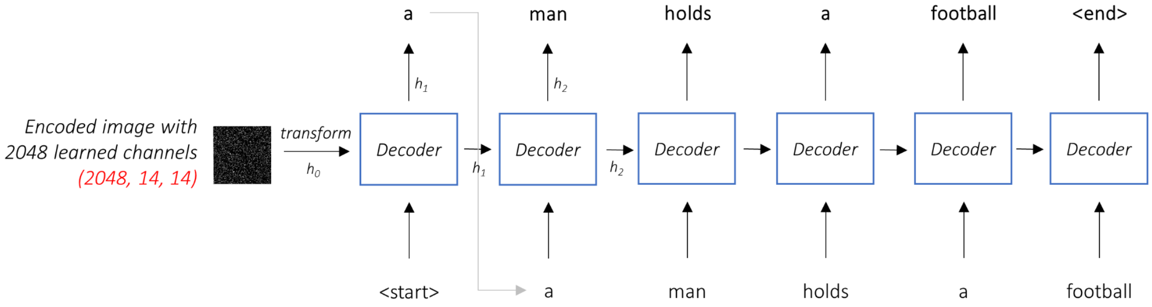
Mini-Project

TEAM ID ➖ 11
Prasad Magdum (MT2022078)
Raj Kachhadiya (MT2022513)
Gautam rizwani (MT2022142)

➢ Baseline Architecture ➖

- Here our task is image captioning. And we know that image captioning is a sequence to sequence task. So for that we will require encoder-decoder architecture.
- So here we will be using CNN as encoder and LSTM as decoder. In baseline system architecture we used pre-trained ResNet18 model as encoder and one layer LSTM as decoder.
- **Encoder (ResNet18)** ➖ It is used as a features extractor in our system. By removing the fully connected layer, we can utilize the intermediate feature maps produced by the convolutional layers of resnet18. These feature maps contain high-level semantic information about the input images. In this way we can extract the feature maps using ResNet18.And if at all required we are training only last 2 layers of ResNet.
- **Decoder (LSTM)** ➖The Decoder's job is to look at the encoded image and generate a caption word by word. Here in baseline architecture we simply average the encoded image across all pixels. And then feed this, with a linear transformation, into the Decoder as its first hidden state and generate the caption. Each predicted word is then used to generate the next word.



- Above is the baseline structure in which we only passed the averaged encoded image to the LSTM decoder.

- Word embeddings ➖ we extracted image features using CNN but along with that we should have captions embeddings so that we can pass this two embeddings as input to LSTM. so we used vocabulary to find the embeddings of the captions. For that we created one vocabulary using captions of the train dataset, So that every word in the train dataset can be represented in real numbers. Every word is going to be embedded in a higher dimensional real number space with which we can operate to handle the LSTM.

➢ Modified Baseline Architecture ➖

- In baseline architecture we only used simple averaged image feature vectors from CNN and simple learnable word representation as input to LSTM.
- But now in modified architecture we want to enhance the input embeddings. For that we will use various techniques.
- While using the various techniques we will keep the LSTM architecture as same as Baseline LSTM architecture.

- **Techniques used in modified architecture ➖**

  I. **Improved word embeddings -**

     - Instead of using simple learnable word embeddings, we will use the pre-trained word embeddings i.e. word2vec. These embeddings capture semantic relationships between words and can provide a richer representation of the words in the captioning task. We initialized the word embeddings with pre-trained vectors and fine-tuned them during training.
     - Here we used a pre-trained word2vec model instead of building vocabulary of the train dataset.

  II. **Attention mechanisms -**

     - Incorporating attention mechanisms can improve the focus of the model on relevant image regions while generating captions. Instead of using a single feature vector to represent the entire image, attention mechanisms allow the model to selectively attend to different image regions, giving more importance to relevant regions for generating each word. This can be achieved by assigning attention weights to different spatial locations of the image and combining them with the image features.
     - Instead of the simple average, we use the weighted average across all pixels, with the weights of the important pixels being greater. This weighted representation of the image can be concatenated with the previously generated word at each step to generate the next word.
     - Above is the one of modified architecture in which we pass the weighted encoded images as an input to the LSTM. and the weights that are corresponding to the encoded images are also learned with our model.
     - So in nutshell what attention mechanism does is it considers the sequence generated thus far, and attends to the part of the image that needs describing next.

➢ subjective comparison results using baseline method and modified baseline ➖

I. **Baseline Method:**
   - The baseline method without word embeddings and attention mechanism generates captions that are somewhat descriptive but lack specific details.
   - The captions tend to be generic and do not exhibit a deep understanding of the image content.
   - The model often generates captions that are repetitive or contain common phrases.
   - The generated captions may not accurately describe important objects or actions in the image.

II. **Modified Baseline (with Word2Vec and Attention Mechanism):**
   - The modified baseline method with word embeddings (Word2Vec) and attention mechanism generates more accurate and contextually grounded captions.
   - The captions exhibit a better understanding of the image content and provide specific details.
   - The model pays attention to relevant image regions while generating each word, resulting in captions that are more focused and visually grounded.
   - The generated captions show a better grasp of the relationships between objects and their attributes, leading to more coherent and meaningful descriptions.
   - The use of word embeddings (Word2Vec) helps in capturing semantic relationships between words and provides a richer representation of the vocabulary, leading to more diverse and contextually appropriate word choices in the captions.

Justification:

Justification of the differences that we observed in the context of our applied techniques.

I. **Word2Vec:** The use of pre-trained word embeddings (Word2Vec) enhances the model's understanding of the vocabulary and semantic relationships between words. The model is able to leverage the pre-trained embeddings to generate more contextually appropriate and diverse captions. This results in captions that are more specific, descriptive, and demonstrate a better grasp of the image content.

II. **Attention Mechanism:** The attention mechanism allows the model to focus on relevant image regions while generating each word of the caption. By assigning attention weights to different spatial locations, the model can selectively attend to important visual cues. This attention mechanism helps in generating captions that are visually grounded and aligned with the salient features of the image. The captions become more specific and contextually relevant, capturing the fine-grained details and relationships between objects in the image.

➢ Results

Models	BLEU SCORE
Baseline Model	0.208
Modified Baseline Model (only word2vec)	0.2085
Modified Baseline Model (only attention)	0.209
Modified Baseline Model (with word2vec & attention)	0.219

Here word2vec is used to improve the word embeddings and attention is used to improve the image embeddings.