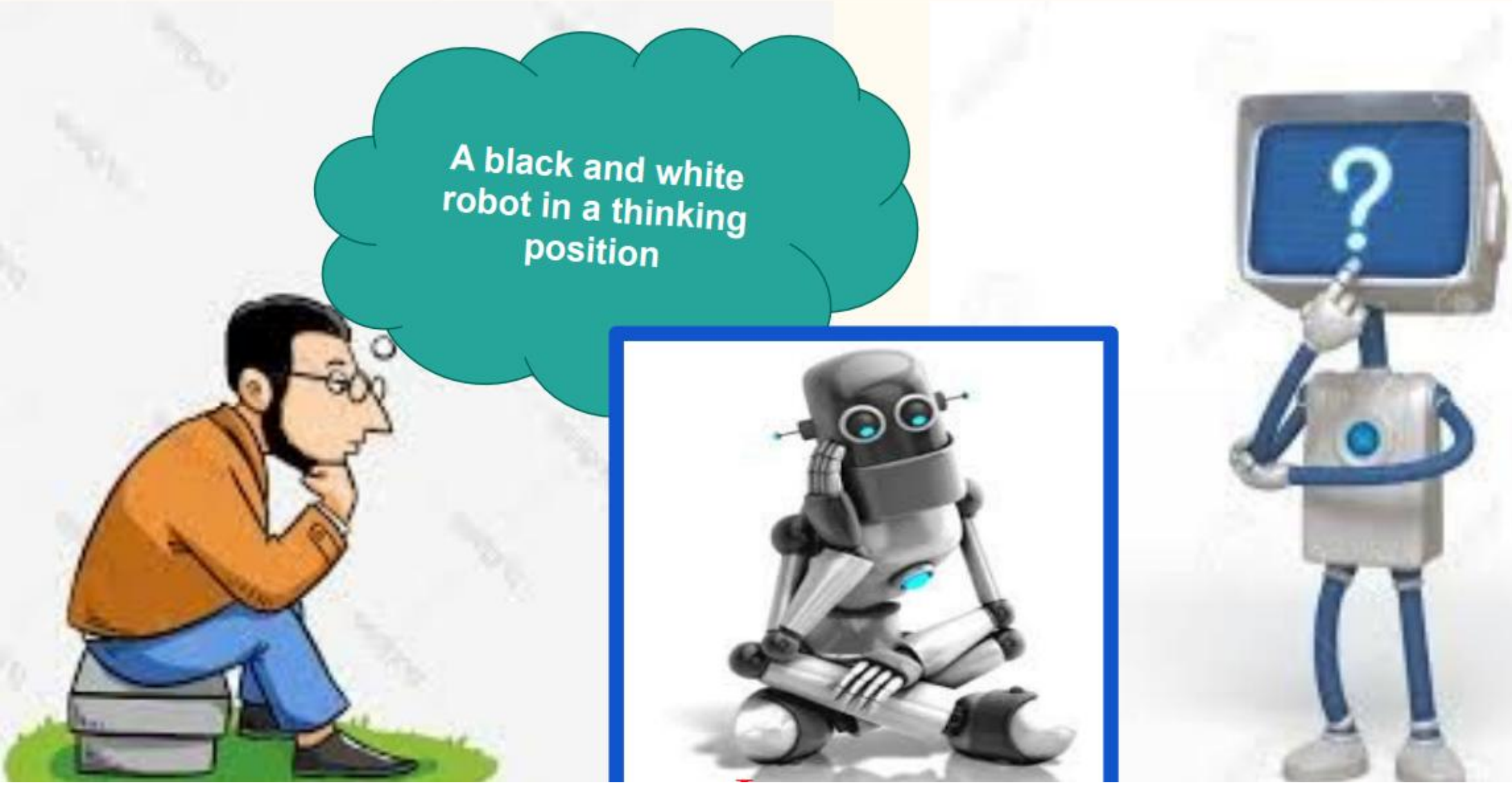
Image Caption Generator with CNN and LSTM

Submitted by:
Ashish Bisht (MDS202313)
Bhuvnesh Kumar (MDS202316)
Vikas Chaudhary (MDS202353)
Vishal Maurya (MDS202354)

1.What is Image Captioning ?



🖼️ Why Image Captioning is Needed

E-commerce
Social Media & Automation
Healthcare Imaging
Surveillance and Security

2. How Human do it !!

Find key things

Dog- object
Brown dog- about object
Running -action
Grass - background
Green grass - about background

Play with the words and Create a meaningful sentence

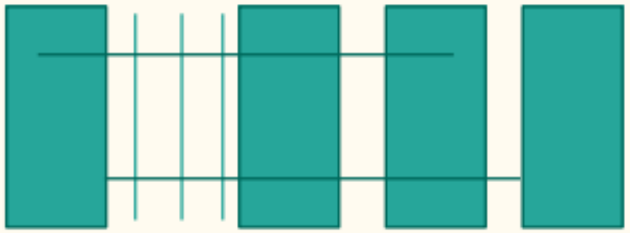
- A brown dog run
- A brown dog run over grass .
- A brown dog with its front paw off the ground on a grassy surface near red and purple flower .
- A dog run across a grassy lawn near some flower
- A yellow dog be play in a grassy area near flower .

How it can be done by a machine

Feature extraction

Features for the whole image

LSTM model

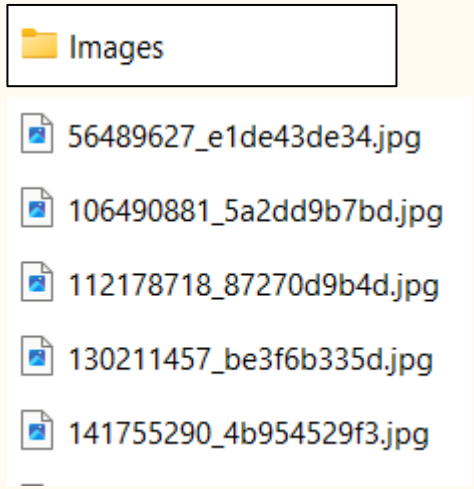


a brown dog is running over green grass.

Dataset

Flickr8k Dataset

- 8000 images and 40000 captions
- 5 captions for each image
- Source: Kaggle




```
image,caption
1000268201_693b08cb0e.jpg,A child in a pink dress is climbing up a set of stairs in an entry way .
1000268201_693b08cb0e.jpg,A girl going into a wooden building .
1000268201_693b08cb0e.jpg,A little girl climbing into a wooden playhouse .
1000268201_693b08cb0e.jpg,A little girl climbing the stairs to her playhouse .
1000268201_693b08cb0e.jpg,A little girl in a pink dress going into a wooden cabin .
1001773457_577c3a7d70.jpg,A black dog and a spotted dog are fighting
1001773457_577c3a7d70.jpg,A black dog and a tri-colored dog playing with each other on the road .
1001773457_577c3a7d70.jpg,A black dog and a white dog with brown spots are staring at each other in the street .
1001773457_577c3a7d70.jpg,Two dogs of different breeds looking at each other on the road .
1001773457_577c3a7d70.jpg,Two dogs on pavement moving toward each other .
```

Dataset

Images

Text

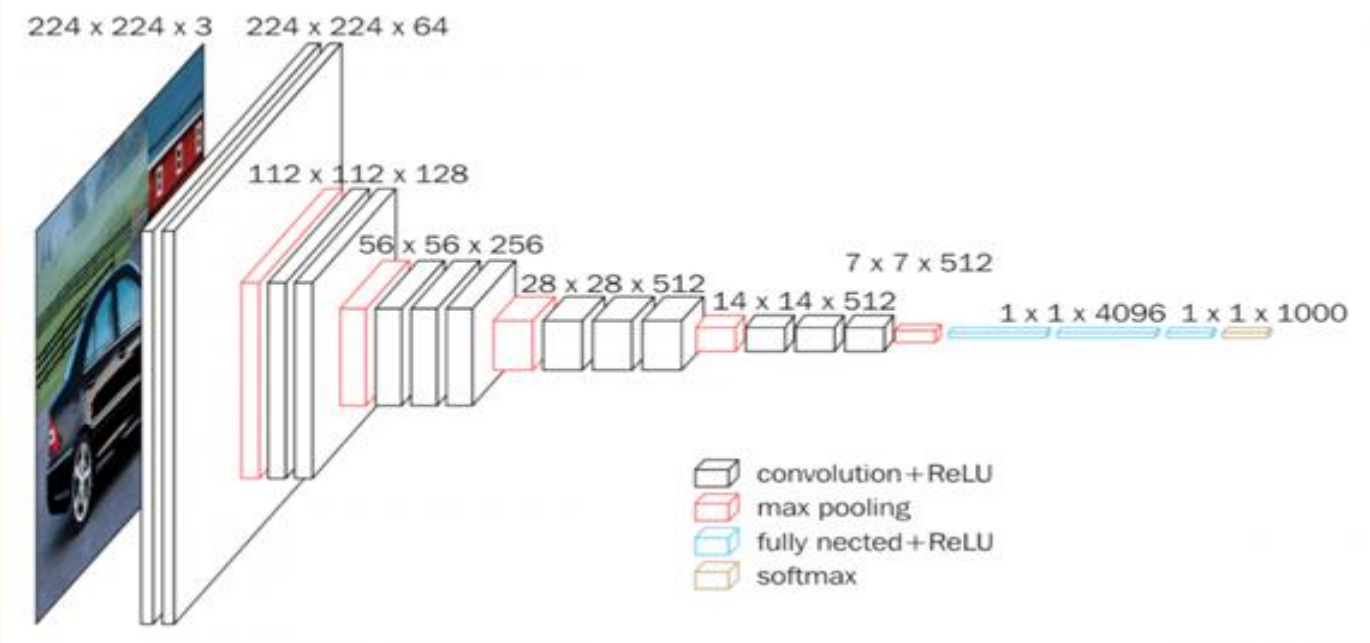
A young boy in an orange hoodie is sliding down a playground slide.
The child smiles while enjoying a ride on the blue slide.
A toddler plays happily at an outdoor playground.
A boy in jeans and a hoodie comes down a slide at the park.
A cheerful child is captured mid-slide on a colorful playground structure.

A black and white dog leaps to catch a red frisbee in a grassy field.
The energetic Border Collie jumps high in the air with a frisbee in its mouth.
A dog performs a trick while catching a frisbee.
A playful dog runs across a field chasing a flying disc.
A dog is captured mid-air as it catches a frisbee during play.

A group of people are smiling and holding wine glasses indoors.
Friends pose together during a gathering with drinks.
A happy group enjoys wine and conversation at a social event.
Four adults and an older man stand side by side holding wine glasses.
People smile for the camera while celebrating together at home.

Data Preparation (Images)

- CNN models to extract features
  - Preprocess images
    - 224*224*3
  - Pretrained model used - VGG16
    - 134,260,544 parameters
    - Input: 224 x 224 x 3 images
    - Output: 4096 feature vector
  - Remove last layers used for predicting



Data Preparation (Text)

- Load all image captions from the dataset.

```
captions_dict["1000268201_693b08cb0e"]=
['A child in a pink dress is climbing up a set of stairs in an
entry way .',
  'A girl going into a wooden building .',
  'A little girl climbing into a wooden playhouse .',
  'A little girl climbing the stairs to her playhouse .',
  'A little girl in a pink dress going into a wooden cabin .']
```

- Create a dictionary mapping
  - Image:[c1, c2, c3, c4]
- Clean Captions
  - Convert all words to lowercase
  - Remove punctuation, numbers, and extra spaces, one-character words
- Encode text Data
  - Add <startseq> and <endseq> tokens to each description
  - Use a tokenizer to map words to integers
  - Create a word-to-index mapping from the vocabulary

```
vocab_size, max_len

(8485, 35)
```

- Train test split
  - 90% train

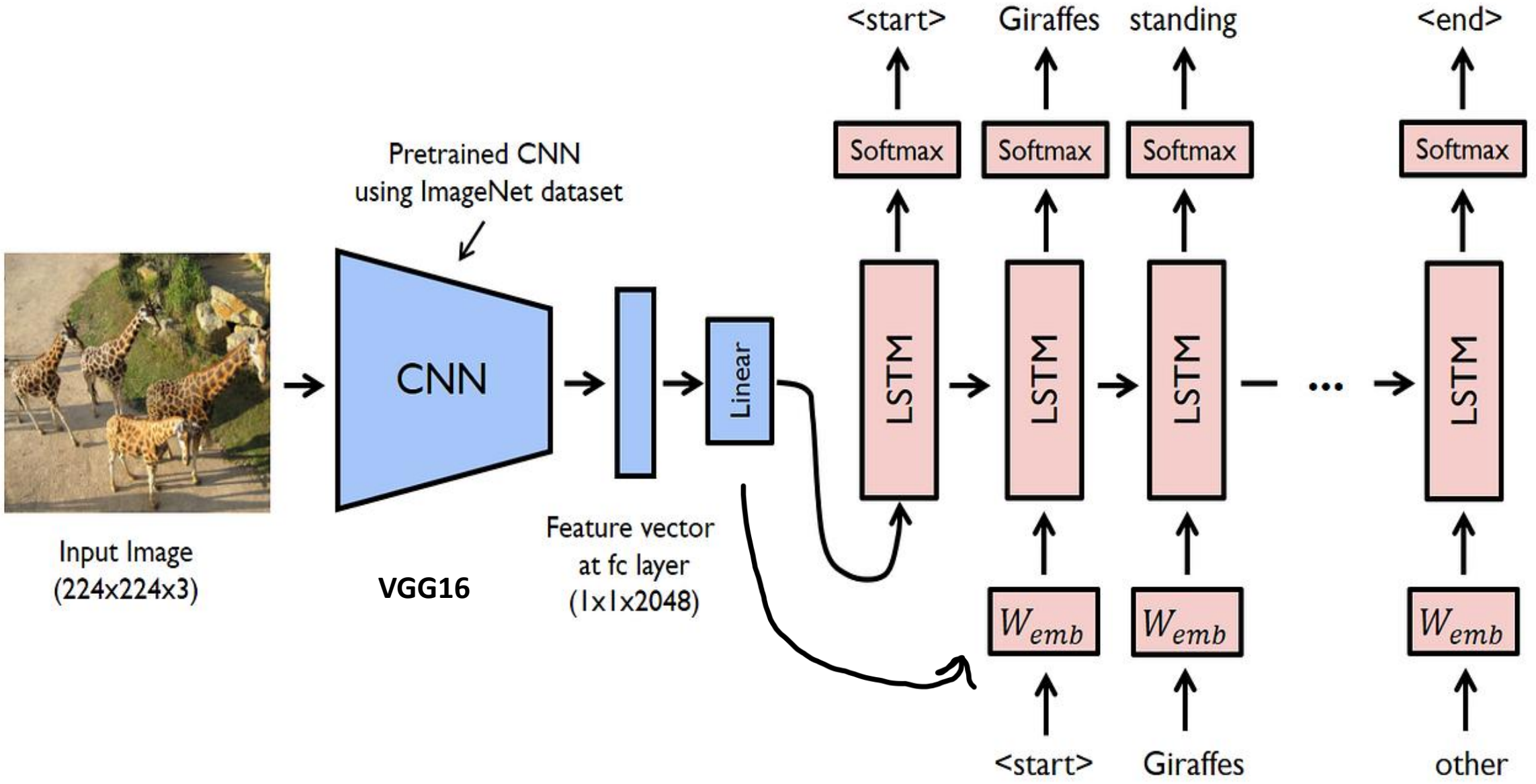
Before Cleaning

```
['A child in a pink dress is climbing up a set of stairs in an entry way .',
 'A girl going into a wooden building .',
 'A little girl climbing into a wooden playhouse .',
 'A little girl climbing the stairs to her playhouse .',
 'A little girl in a pink dress going into a wooden cabin .']
```

After Cleaning

```
['<startseq> child in pink dress is climbing up set of stairs in an entry way <endseq>',
 '<startseq> girl going into wooden building <endseq>',
 '<startseq> little girl climbing into wooden playhouse <endseq>',
 '<startseq> little girl climbing the stairs to her playhouse <endseq>',
 '<startseq> little girl in pink dress going into wooden cabin <endseq>']
```

Solution Architecture & Workflow



Solution Architecture & Workflow

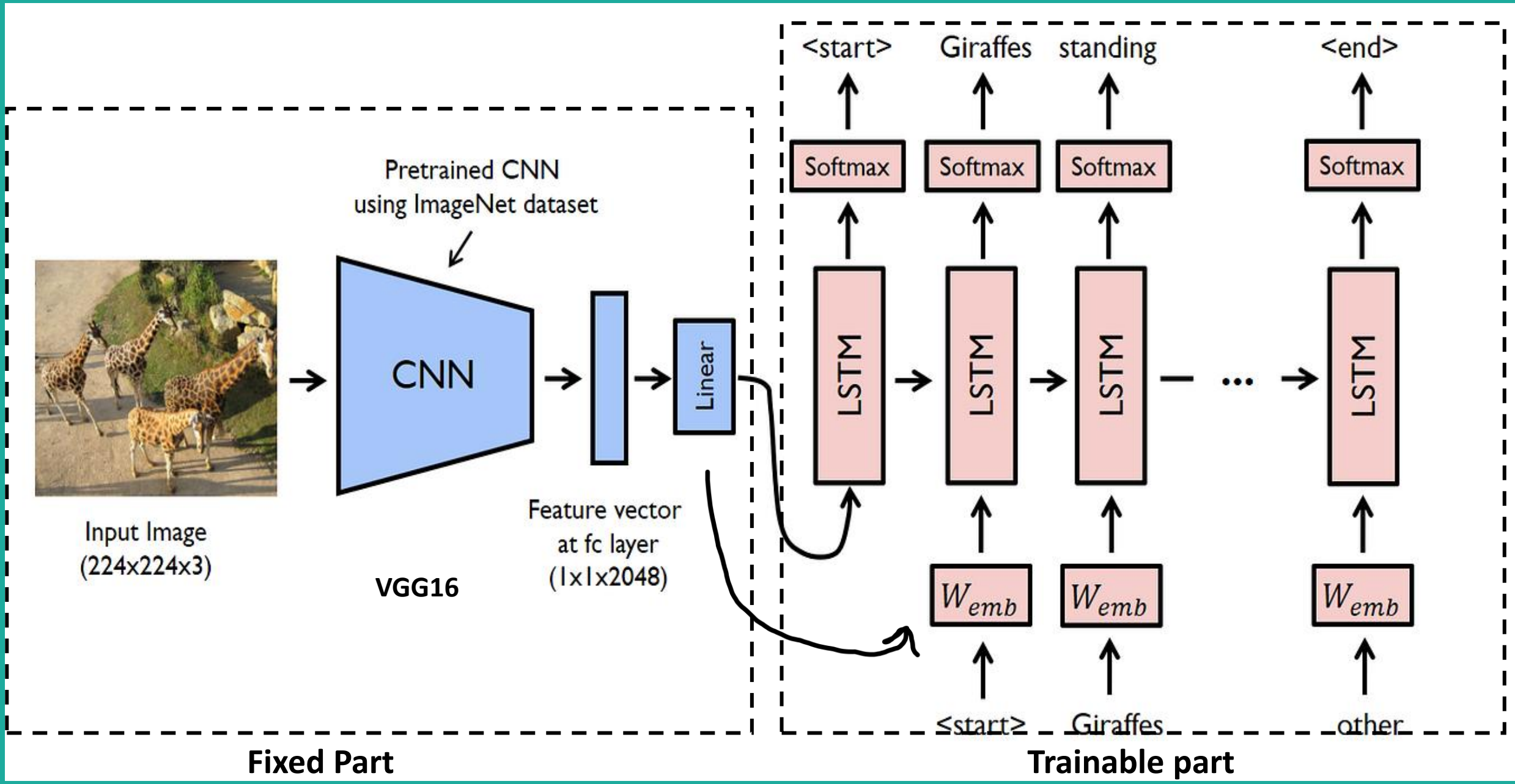
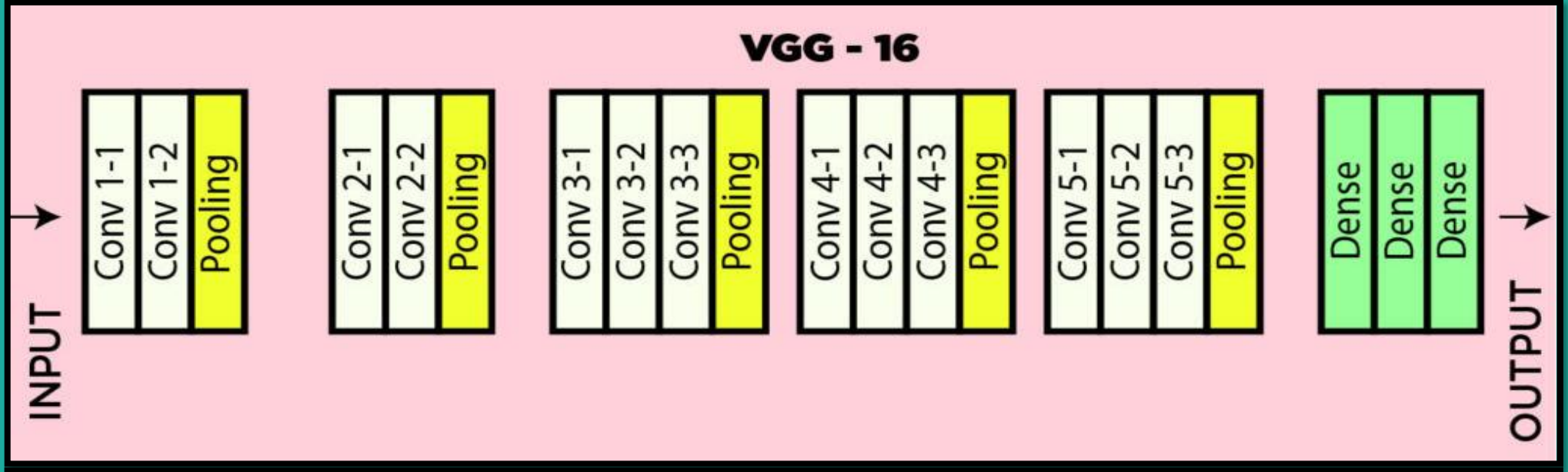
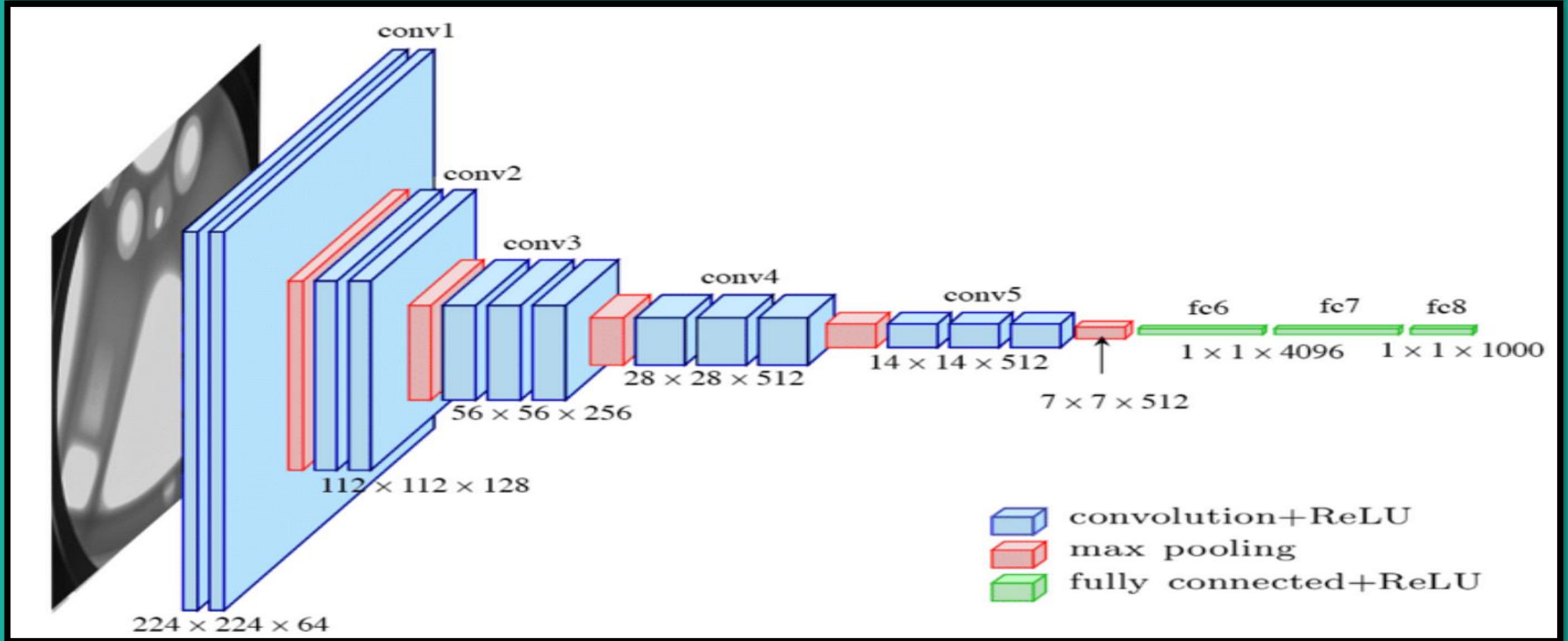


Image to Feature Map



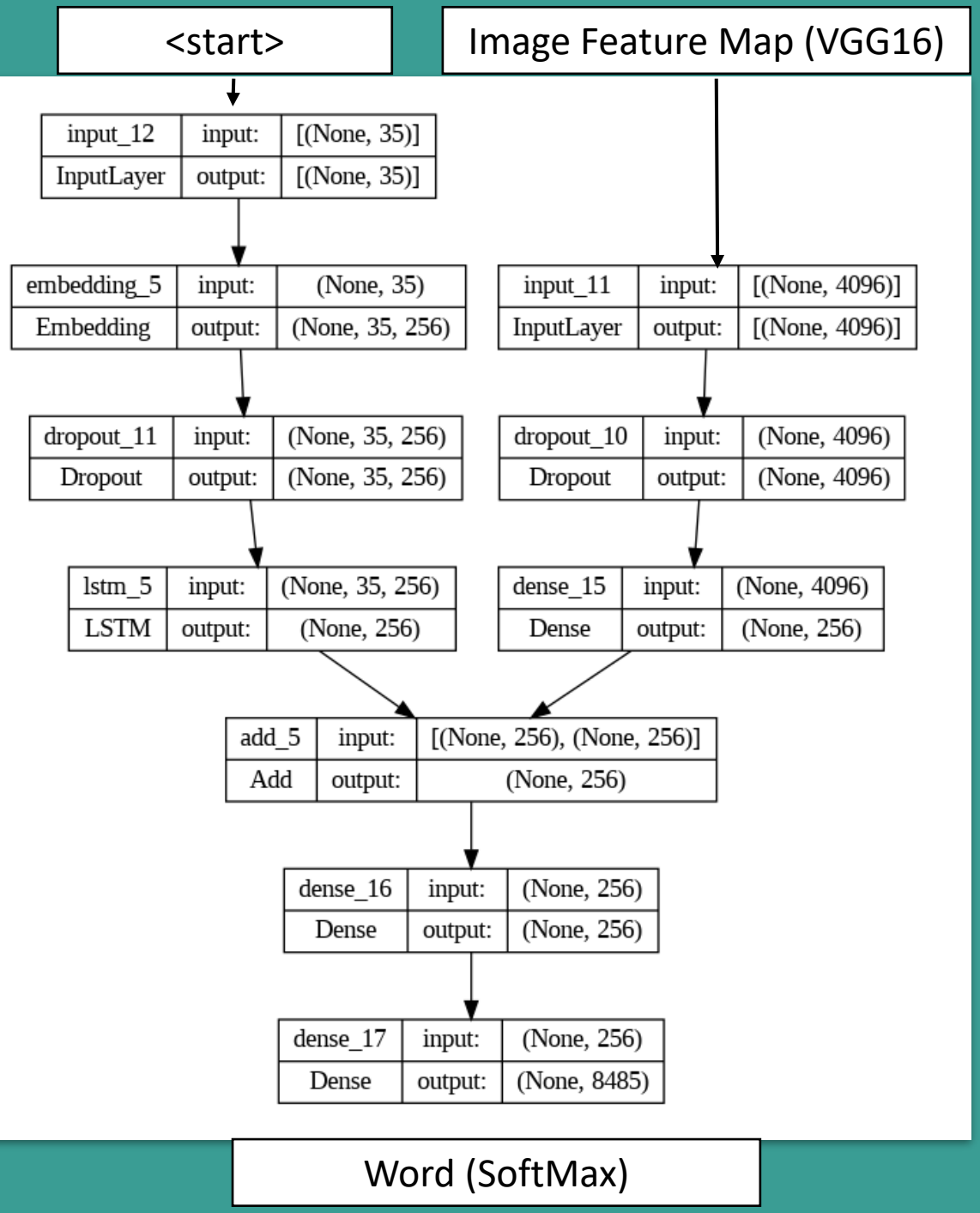


```python
def define_model(input_features, vocab_size, max_len):
    # ---- encoder model ------
    # image features layer
    inputs1 = Input(shape=(input_features,))
    fe1 = Dropout(0.4)(inputs1)
    fe2 = Dense(256, activation='relu')(fe1)

    # sequence model (text process layer)
    inputs2 = Input(shape=(max_len,))
    se1 = Embedding(input_dim=vocab_size, output_dim=256, mask_zero=False)(inputs2)
    se2 = Dropout(0.4)(se1)
    se3 = LSTM(256)(se2)

    # ------ decoder model ------
    decoder1 = Add()([fe2, se3])
    decoder2 = Dense(256, activation='relu')(decoder1)
    outputs = Dense(vocab_size, activation='softmax')(decoder2)

    # tie it together [image, seq] -> [word]
    model = Model(inputs=[inputs1, inputs2], outputs=outputs)
    model.compile(loss='categorical_crossentropy', optimizer='adam', metrics=["accuracy"])
```



- **VGG16**, a pretrained convolutional neural network trained on the **ImageNet dataset** (1.2 million images across 1000 classes), is used to **extract visual features** from input images.

- The extracted feature maps from the CNN are **not fine-tuned**, and only the **custom-built LSTM decoder** is trained on the task-specific dataset.

- The **LSTM model** uses these visual features to **generate natural language descriptions** of the corresponding images.

Training

➢ <start> A dog run across ... lawn <end>

processing

CNN

Vector representation of a word

Long Feature vector

To calculate error

To easy and quick learn

LSTM model

A brown dog is running over green grass.

Testing

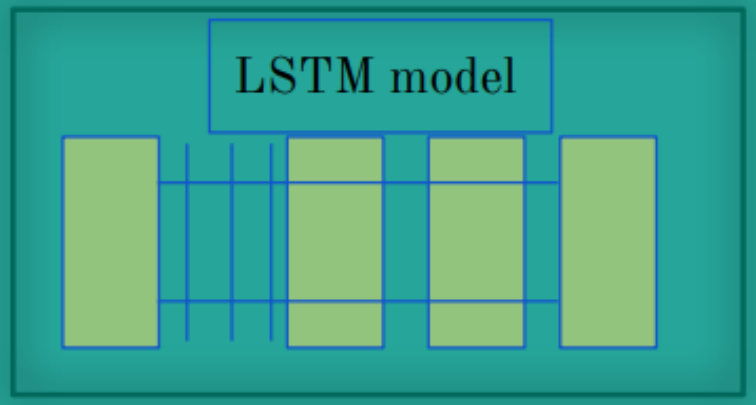
➢ <start> A dog run across ... lawn <end>

processing

CNN

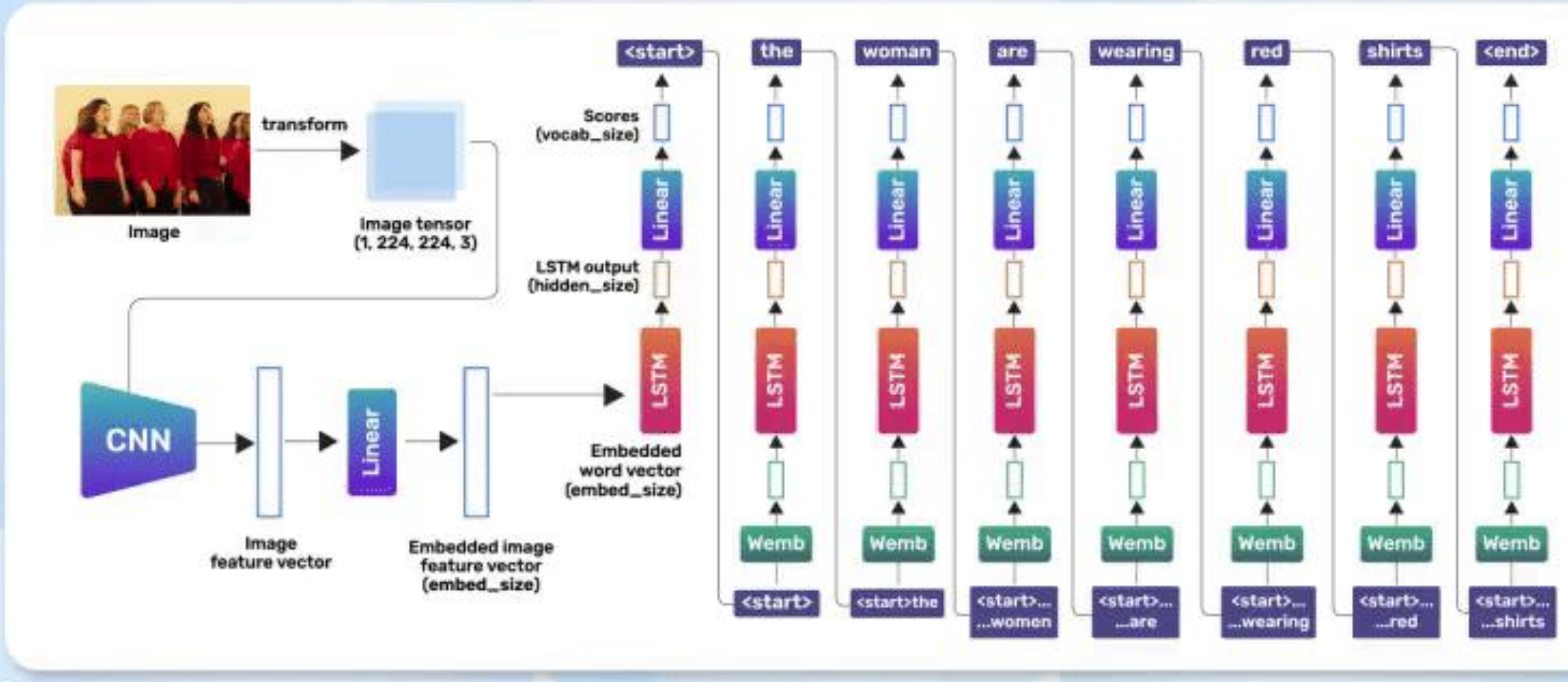
Long Feature vector

To see model performance



A brown dog is running over green grass.

Image Captioning using ResNet & LSTM



LearnOpenCV.com

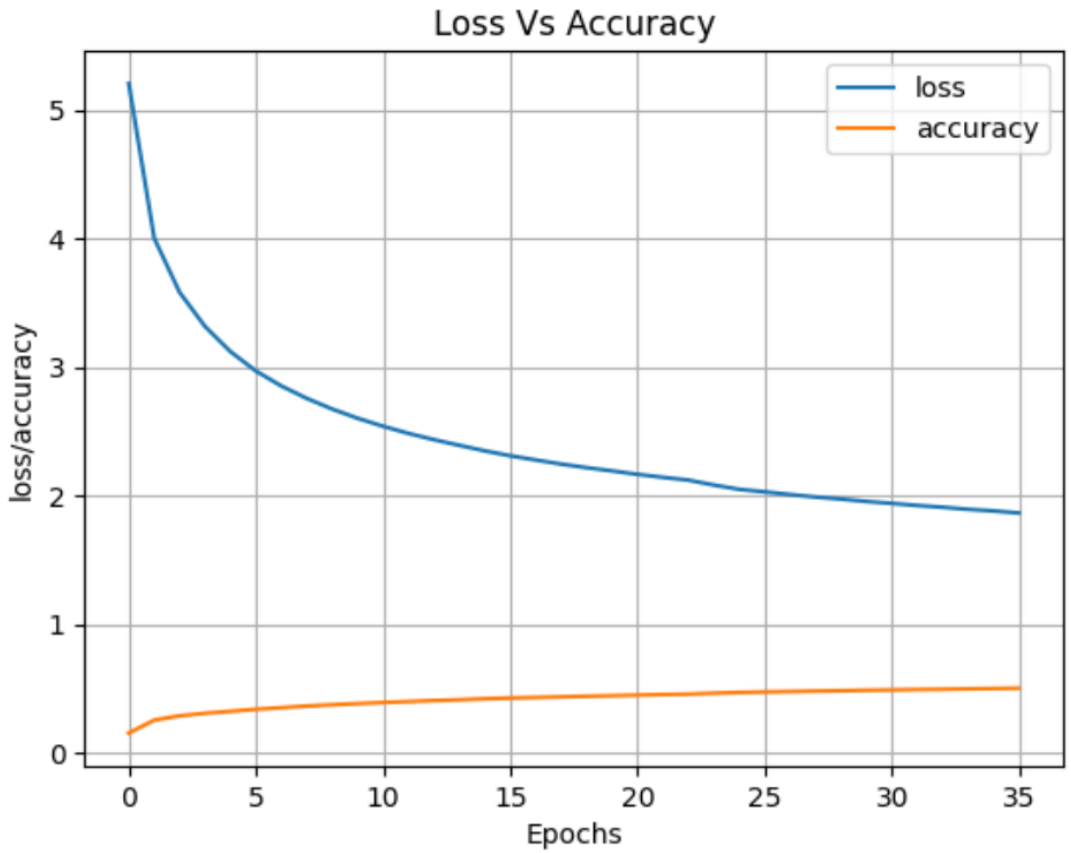
Evaluation with BLEU score

What is BLEU Score?

- **BLEU (Bilingual Evaluation Understudy)** is a metric used to evaluate the quality of **machine-generated text** (e.g., in machine translation or image captioning).
- It is **language-independent**, **fast**, and **easy to compute**.
- The score lies in the range **[0, 1]**:
  - **1** means a perfect match with the reference sentence.
  - **0** means no overlap.

Example:

reference = "I like dog"
hypothesis1 = "I like dog" -> 1.0
hypothesis2 = "i like dog i like dog" -> 0.47
hypothesis3 = "dog is liked by me" -> 0.67
hypothesis4 = "dog is liked" -> 0.76




```
# calculate BLEU score (used when using text)
print("BLEU-1: %f" % corpus_bleu(actual, predict, weights=(1.0, 0, 0, 0)))
print("BLEU-1: %f" % corpus_bleu(actual, predict, weights=(0.5, 0.5, 0, 0)))
```

```
100%|██████████| 810/810 [08:39<00:00,  1.56it/s]
BLEU-1: 0.616747
BLEU-1: 0.426457
```

Accuracy & Results

Images & Predictions

In [26]:
```
generate_image('1003163366_44323f5815.jpg')
```

<startseq> man lays on the bench while man sits on the ground by him <endseq>

```
A man lays on a bench while his dog sits by him .
A man lays on the bench to which a white dog is also tied .
a man sleeping on a bench outside with a white and black dog sitting next to him
A shirtless man lies on a park bench with his dog .
man laying on bench holding leash of dog sitting on ground
```

In [22]:
```
generate_image('1007320043_627395c3d8.jpg')
```

<startseq> little boy climbing ropes net <endseq>

```
A child playing on a rope net .
A little girl climbing on red roping .
A little girl in pink climbs a rope bridge at the park .
A small child grips onto the red ropes at the playground .
The small child climbs on a red ropes on a playground .
```

Thank You