Vision Voice: Image Captioning App with Audio

By:
Aravinda Vijayaram Kumar - G36084456
Nemi Makadia – G29362869
Nihar Domala - G34884842
Yash Doshi – G33250233

Objective

- Generate textual descriptions for images to assist visually impaired users.
- Use Pre-Trained and Custom Encoders align with LSTM based decoders for accurate captioning.
- Convert captions to audio using text-to-speech (TTS) for seamless accessibility.
- Leverage state-of-the-art AI to make visual content comprehensible for all.

Introduction of the project

What is the Project About:

- Develop an accessible image captioning system using advanced deep learning techniques.
- Generate textual descriptions for images with state-of-the-art models.
- Integrate a text-to-speech (TTS) module to convert captions into audio.
- Enhance accessibility for visually impaired users.

Importance and Uniqueness:

- Combines image captioning and text-to-speech (TTS) technologies to make visual content accessible to visually impaired users.
- Integrates advanced computer vision models, like Vision Transformers, with natural language processing and TTS modules.
- Provides a user-friendly, end-to-end solution for translating images into meaningful audio descriptions, promoting inclusivity.

Data Description

- COCO 2017 Dataset
- Contains images with 5–7 human-generated captions per image.

- Training Set: ~118,000 images
- Validation Set: ~5,000 images

Preprocessing:

- Image resizing (224x224)
- Normalization
- Caption tokenization.

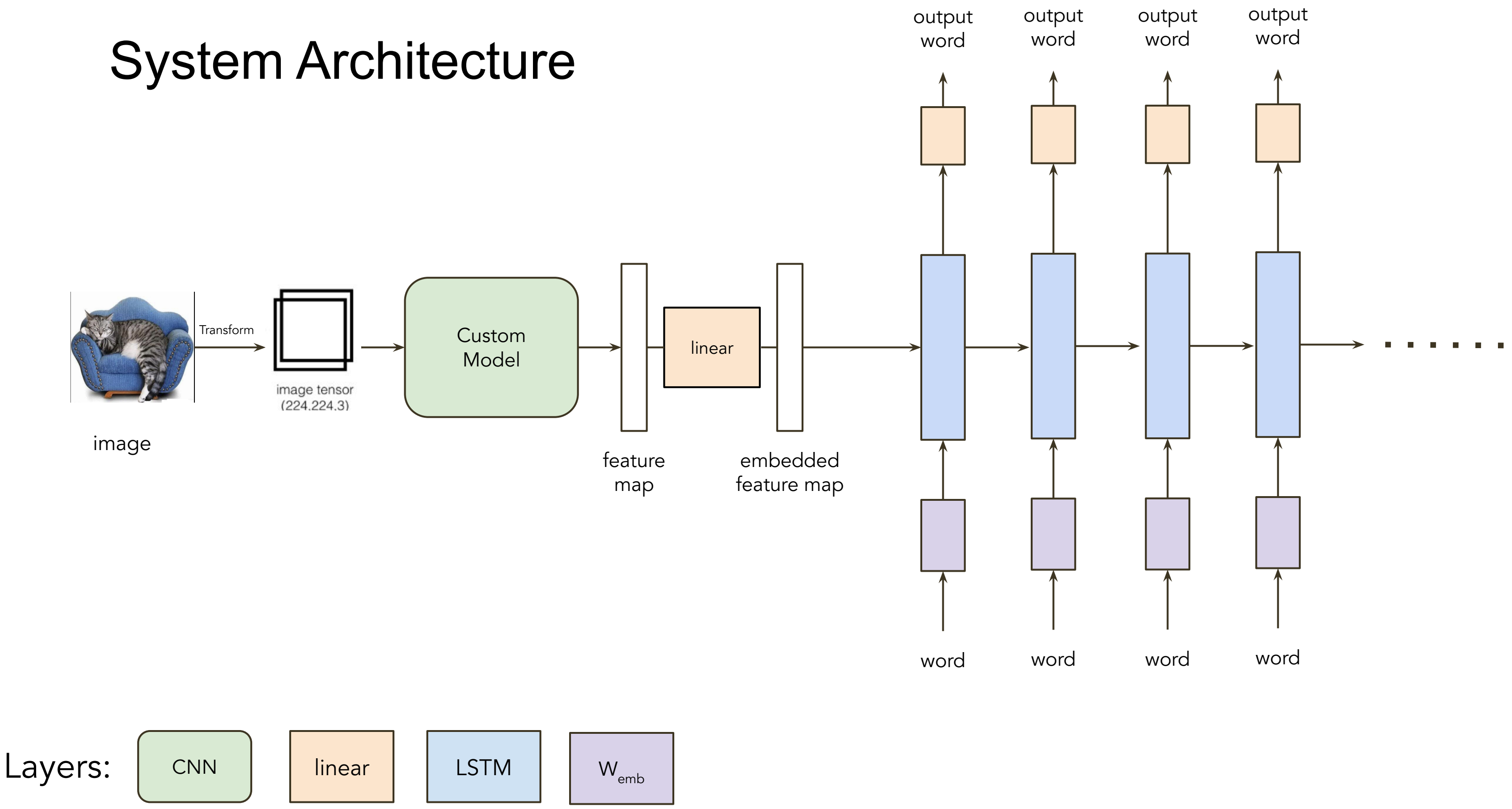
Data Loader

Vocabulary Generation: Creates word-to-index & index-to-word mapping with special tokens like <start>, <end>, and <pad> from all the captions

Image Transformation: Transforms the images with the given transformations

resize(256) - crop(224) - normalize - toTensor

System Architecture



Custom Model Architecture

CNN Encoder Structure:

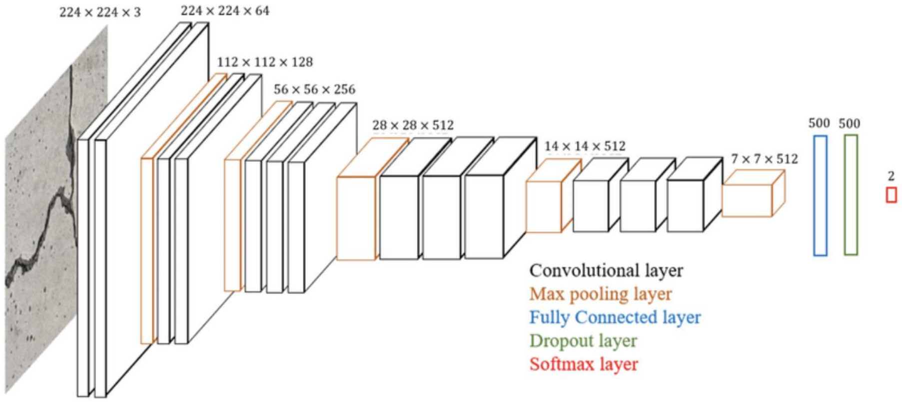
- Three Conv2D layers
- ReLU activation function for non-linearity
- Average Pooling
- FC Layer

RNN-based Decoder Structure:

- Embedded Layer
- GRU(Gated Recurrent Unit) Layer
- FC Layer

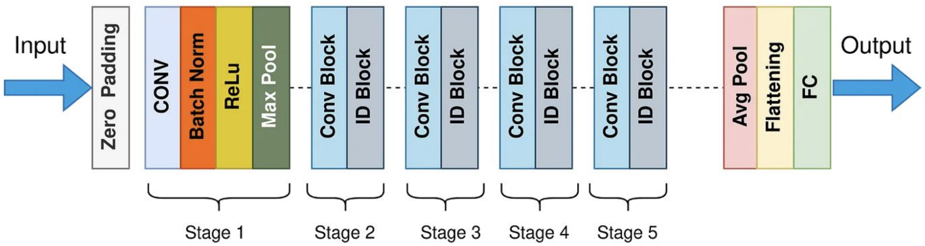
The model combines convolutional layers for visual feature extraction with recurrent layers for sequential text generation.

VGG-16 Architecture



- VGG16 is a deep convolutional neural network with 16 layers
- small 3x3 convolution filters with a stride of 1
- Lower layers - basic features
- Deeper layers - abstract representations
- Fully connected layers - generalizes these features
- Large params

ResNet50 Model Architecture



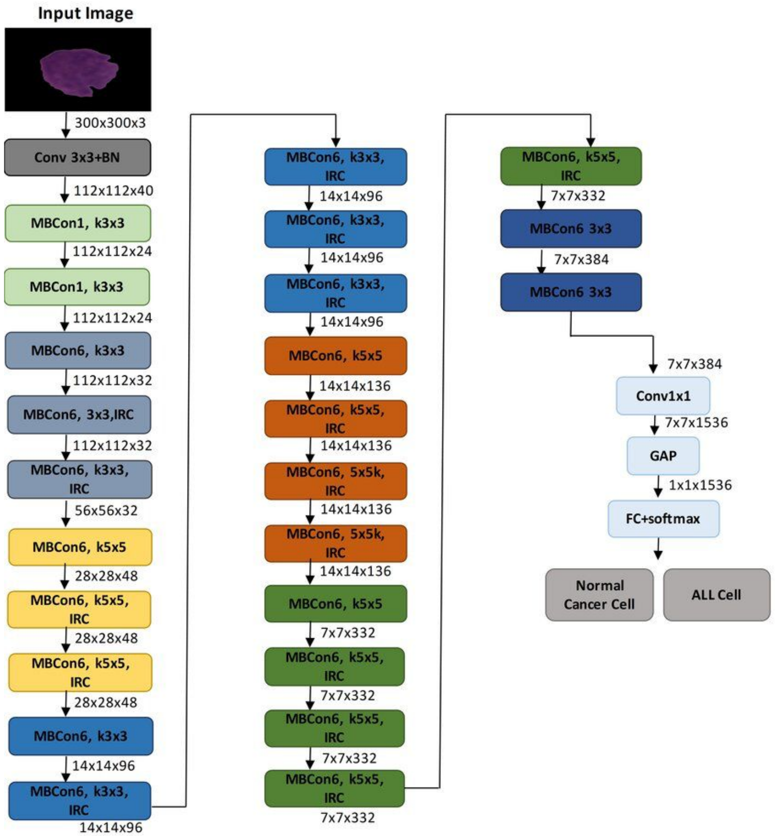
- ResNet50 is a deep convolutional neural network with 50 layers.
- Uses residual connections.
- connections allow the model to skip certain layers
- The architecture is divided into 4 stages

ResNet101 Model Architecture

- ResNet101 is a deeper convolutional neural network with 101 layers.
- Also uses residual connections.
- connections allow the model to skip certain layers
- The architecture is divided into 4 stages
- Higher depth

Aspect	VGG-16	ResNet-101	ResNet-50
Architecture	16-layer CNN with simple sequential layers	101-layer CNN with deeper residual learning	50-layer CNN with residual connections for ease of training
Number of Parameters	~138M	~44.5M	~25.6M
Inference Time	Moderate (slower due to large size)	Slower than ResNet-50 due to added depth	Fast (optimized for deeper architecture)
Feature Extraction Quality	Good, but less robust for fine-grained details	Very high quality but computationally intensive	Excellent: captures hierarchical features efficiently
Memory Usage	High (due to large parameters)	Higher than ResNet-50	Low-to-moderate

EfficientNet-B3 Model Architecture



- Lightweight CNN model with ~12M parameters
- Uses MBConv layers, reducing computation while maintaining accuracy
- Used beam search to optimize generating captions

BLIP Model

BLIP (Bootstrapped Language-Image Pretraining) has two part architecture:

- **Encoder**: Vision Transformer (ViT) for extracting visual features.
- **Decoder**: Transformer-based text decoder for generating captions.

BLIP is trained on COCO Captions, Visual Genome, Conceptual Captions, SBU Captions and Flickr30k

Evaluation

Bleu Score (Bilingual Evaluation Understudy)

- Measures the similarity between a generated caption and reference captions objectively
- Flaw: Focuses on exact matches, missing semantic understanding

Model	Bleu Score
ResNet50	0.1487
Efficient Net B3	0.2002

Demo Time…