National Institute of Technology, Jamshedpur

Department of Computer Science & Engineering

Major Project on

Multimodal Image Retrieval System using Deep Semantic Common Embedding Space

Under the guidance of:
Dr. Vinay Kumar
Assistant Professor
CSE, NIT Jamshedpur

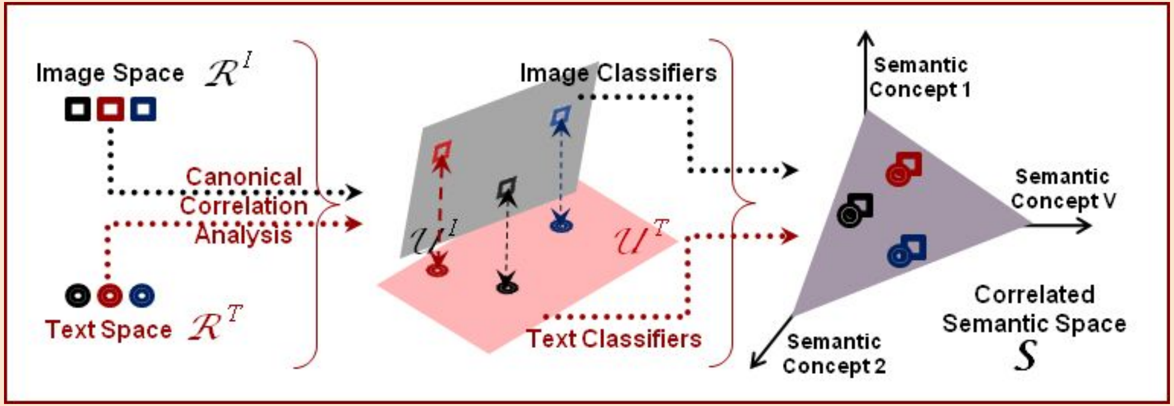
By Group 21:
Amit Bahir (2017UGCS044)
Purusharth Verma (2017UGCS066)
Harshal Desai (2017UGCS086)

Background

- Retrieval of Images based on their visual attributes is a difficult task.
- Traditional approaches based on literal string matching are inefficient due to:
  - Human intervention and proper annotation with correct textual metadata
  - Limited search optimization based on single modality (text)
- Metadata based approaches are limited by annotations given to images and thus unable to capture the visual semantic aspects.

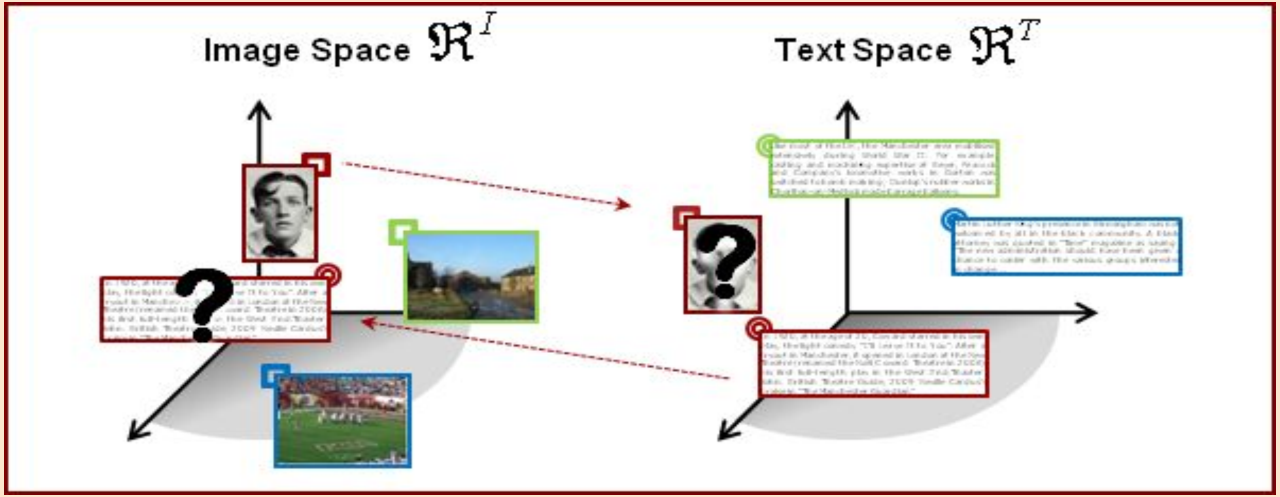
Motivation

- ❖ Limitations of Correlation Analysis
  - ➢ Obtaining maximum correlation between pairs may not necessarily make sense
  - ➢ Pairings are independent of each other
  - ➢ Only linear relations are possible
  - ➢ Semantic aspects not captured with efficiency



Objectives

- ❖ Achieving the semantic aspect of mapping
- ❖ Creating a common embedding space for both modalities simultaneously
- ❖ Making the semantic space more appropriate by using common embedding space
- ❖ Leveraging state-of-the-art techniques like BERT to capture bidirectional information
- ❖ Building a robust model to deal with semantic aspects



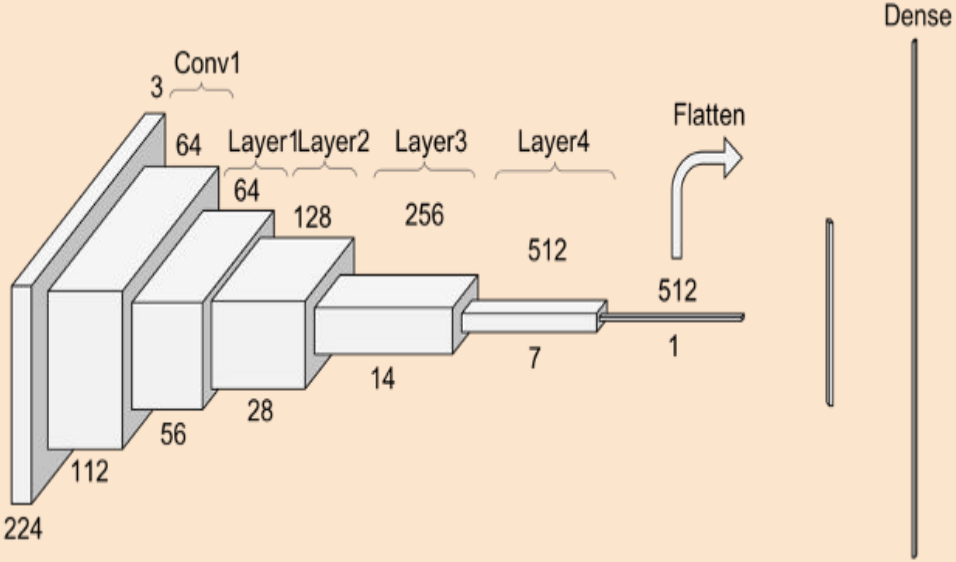
Methodology

- Training model using Flickr 8k dataset which contains 8091 images and each image has 5 textual captions describing the image
- Passing captions through sentence encoders of RoBERTa, a successor of BERT to calculate mean semantic embedding corresponding to each image
- Sentence encoders use siamese based triplet network structures to derive semantically meaningful sentence embeddings that can be compared using cosine-similarity
- Reducing effort for finding the most similar pairs

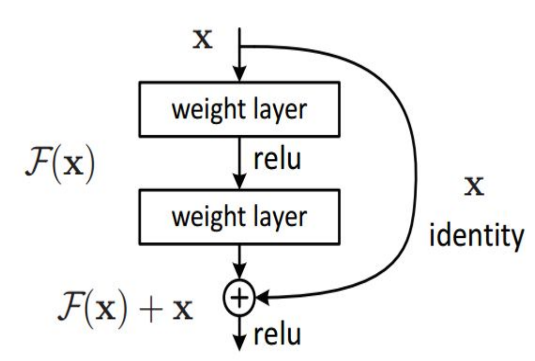
Implementation

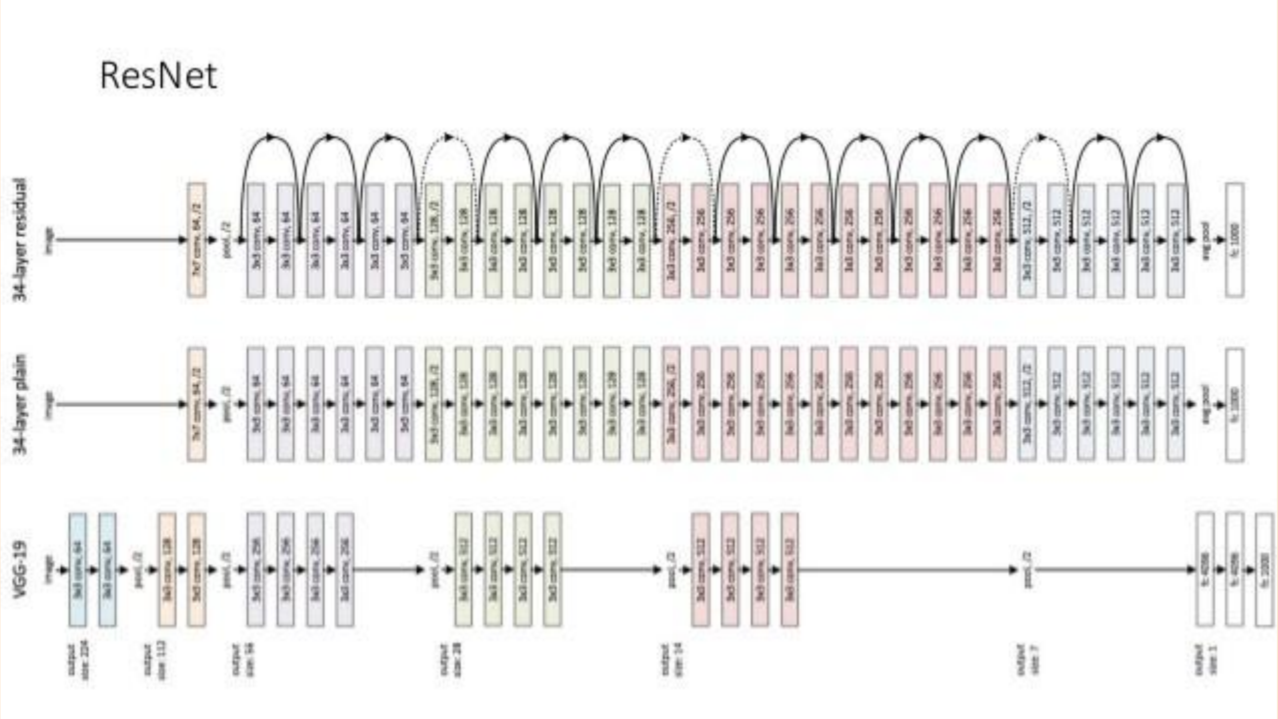
- Data Preprocessing:
  - Converting captions to target embeddings
  - Image Augmentation
- Architecture:
  - Input image was to be converted into a 1024 dimensional embedding, so a regression task was achieved using a CNN structure.
  - ResNet34 as base along with few more layers with a dense prediction layer of 1024 linear units
  - Residual network to allow training of deep neural network without moving too far from the training image

Architecture



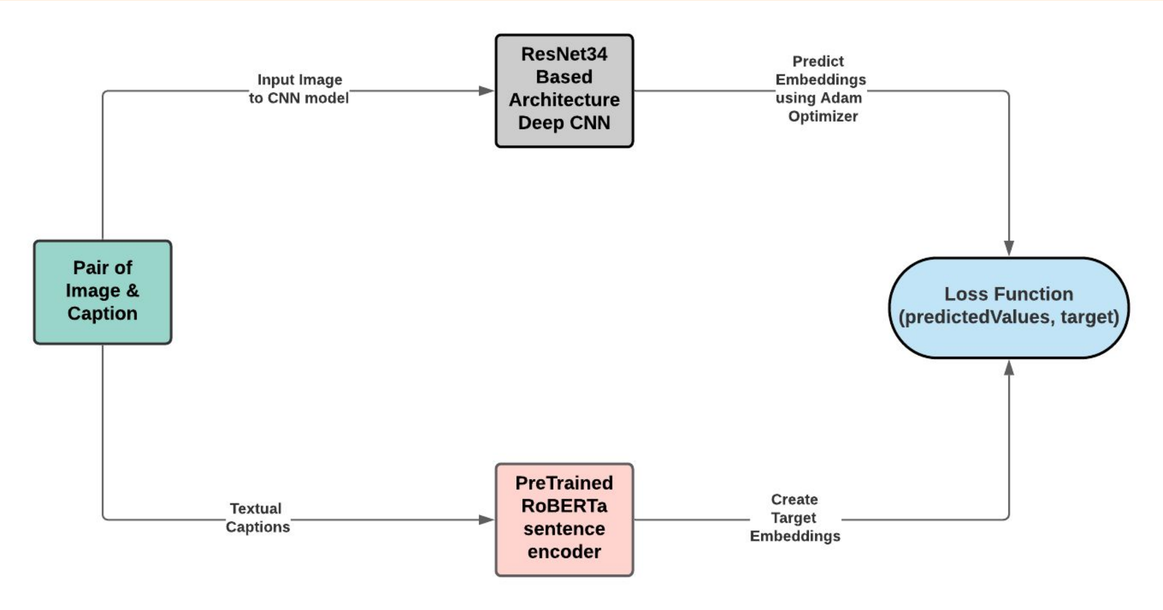
Residual Networks


x
weight layer
relu
$\mathcal{F}(\mathbf{x})$
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu



Training

Done in multiple stages using MSE as loss function, progressive image resizing and Adam optimizer with momentum


ResNet34 Based Architecture Deep CNN
Input Image to CNN model
Predict Embeddings using Adam Optimizer
Pair of Image & Caption
Loss Function (predictedValues, target)
Textual Captions
PreTrained RoBERTa sentence encoder
Create Target Embeddings

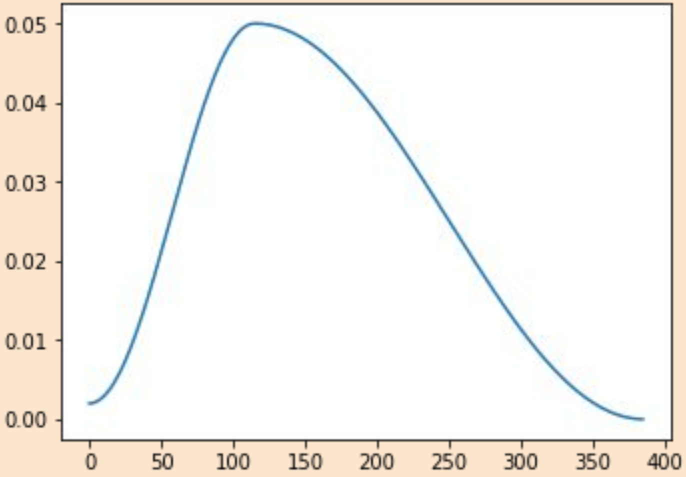
Training Hyperparameters

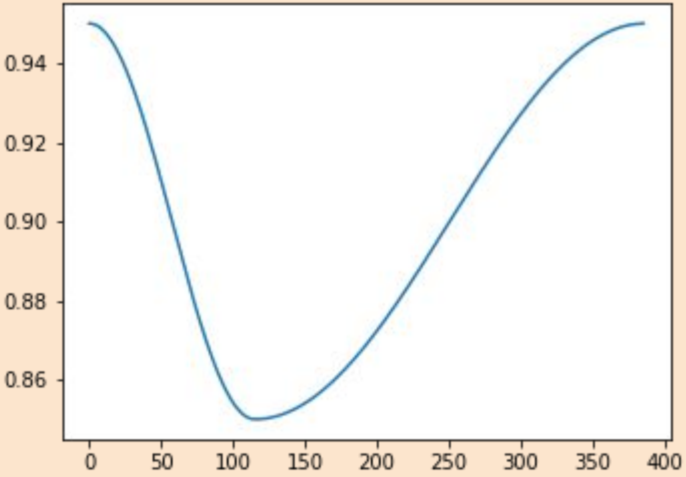
Image Size	Stage	Learning Rates	# Epochs
(224, 224, 3)	1 (Head)	3e-2	24
(224, 224, 3)	2 (Complete Network)	slice(5e-6, 3e-5)	12
(256, 256, 3)	1 (Head)	4e-4	12
(256, 256, 3)	2 (Complete Network)	slice(8e-6, 1e-4)	12

- Fine tuning in two stages
- Progressive resizing
- Discriminative Learning rates

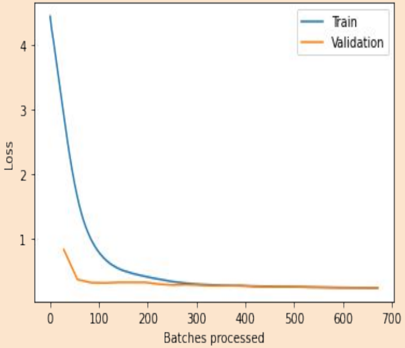
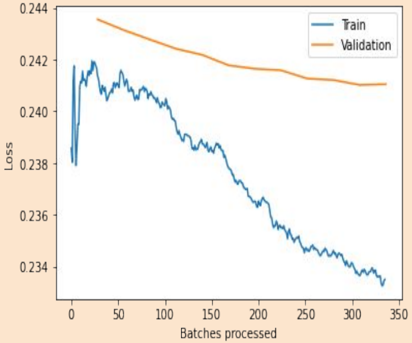
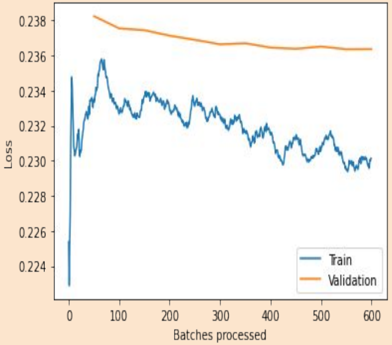
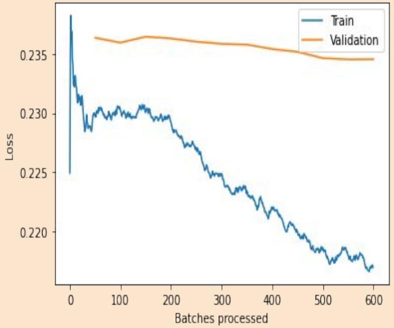
One Cycle Policy

To achieve fast convergence of a loss function by varying the learning rate over a cycle.





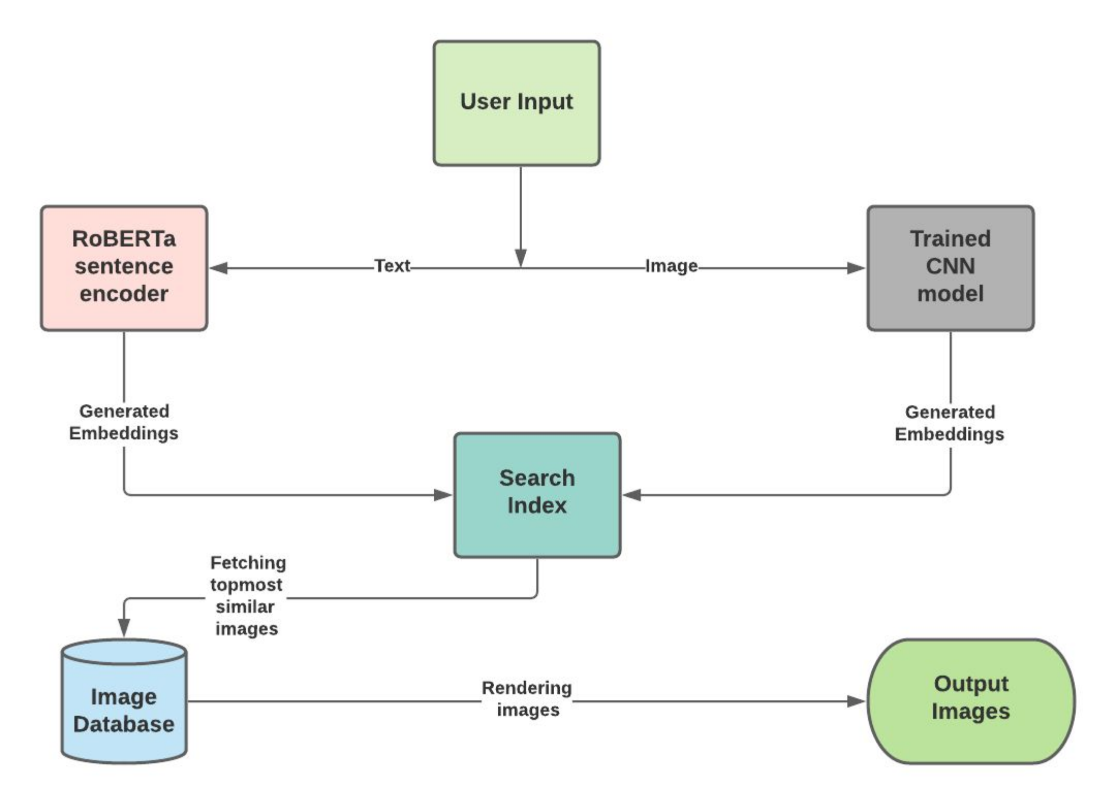
Training Progress

Image Size 224		Image Size 256	
Stage 1	**Stage 2**	**Stage 1**	**Stage 2**
			

Search System

- Using trained model, predictions of all images in database were recorded and a search index was created
- For a query embedding and value k, the system returns a list of indices of top k nearest neighbours in non-decreasing order of their angular distances to the query embedding
- Efficient search by using non-metric based space paradigm to search for approximate nearest neighbours.
- Algorithm creates hierarchical navigable small world groups
- HNSW builds a hierarchical graph incrementally
- Each node in the graph represents a point in the vector space, and nodes are linked to other nodes that are close in space
- Algorithm used shows state-of-the-art results on various retrieval tasks

Complete System Flow

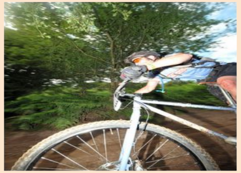
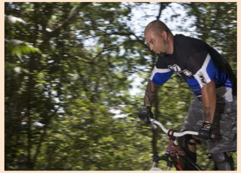
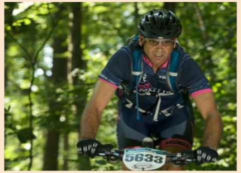
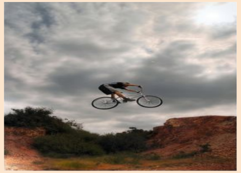
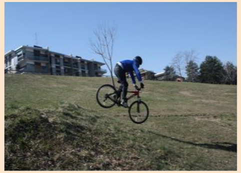
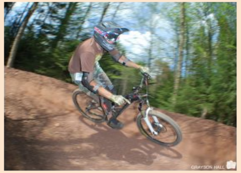
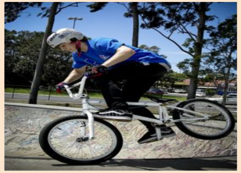
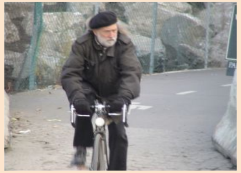
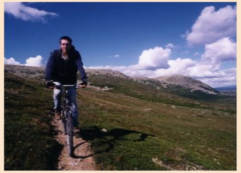
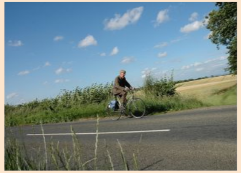
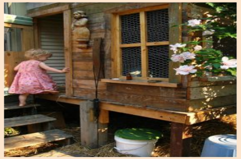
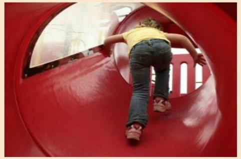
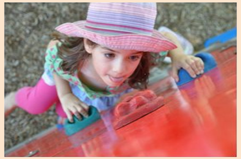
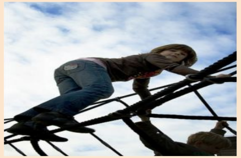
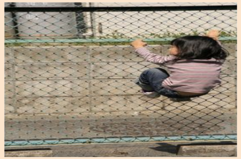
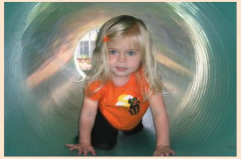
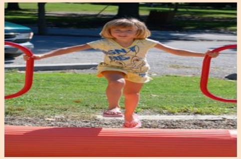
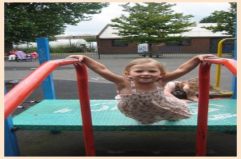
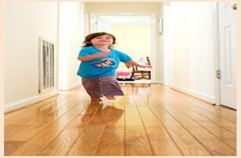
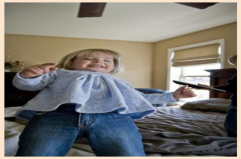


Results - Top K accuracies

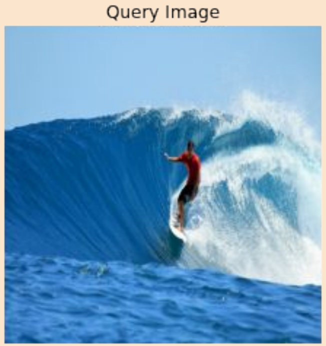
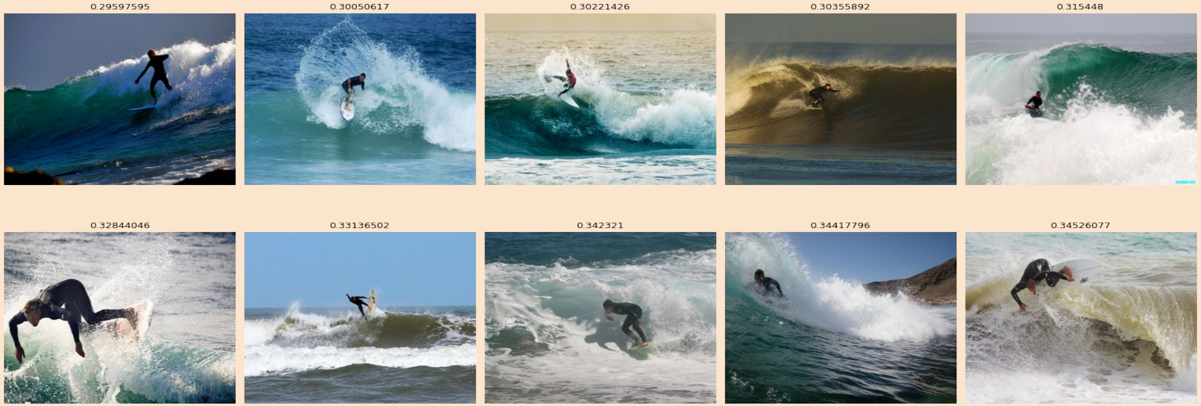
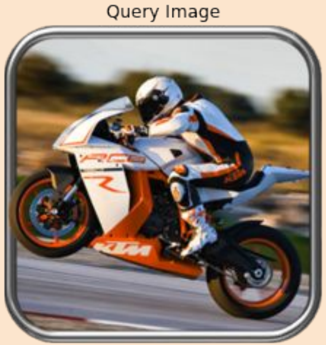
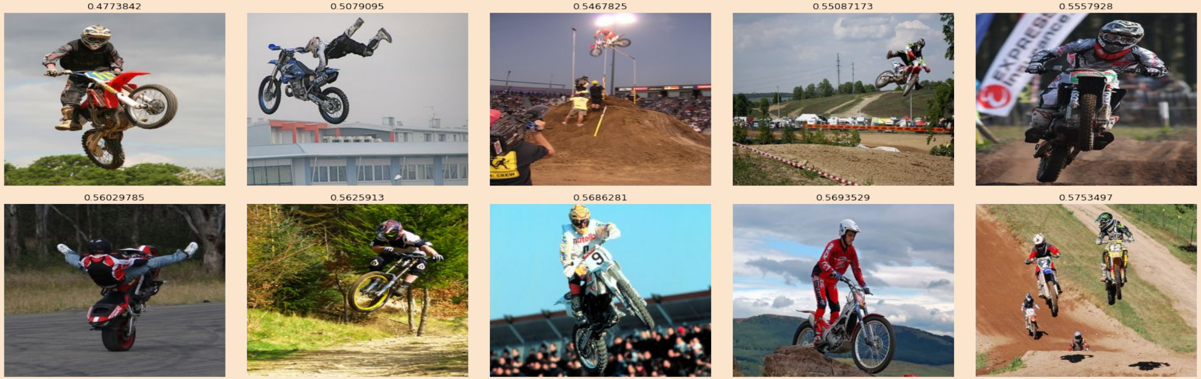
The top 1, 5 & 10 accuracies were calculated for each of the 5 captions as queries and if the image corresponding to query was present in top K results.

Caption #	**Top 1**	**Top 5**	**Top 10**
1	72.77	88.38	92.56
2	75.63	91.09	94.24
3	76.33	90.55	93.73
4	74.77	89.79	93.30
5	72.81	88.80	92.07
Mean	**74.46**	**89.72**	**93.18**

Retrieval Results - TextToImage

Query	Top 10 Results
" A man riding a bike"	0.5113842  0.5341821  0.54886234  0.5640544  0.59729457  0.5983668  0.60246336  0.6035284  0.60374665  0.61008173 
"A little girl climbing into a wooden playhouse ."	0.48381922  0.8791608  0.88085395  0.8925527  0.8937837  0.8992302  0.9007713  0.90756863  0.91956675  0.9275699 

Retrieval Results - ImageToImage

Query	Top 10 Results
	 0.29597595 0.30050617 0.30221426 0.30355892 0.315448 0.32844046 0.33136502 0.342321 0.34417796 0.34526077
	 0.4773842 0.5079095 0.5467825 0.55087173 0.5557928 0.56029785 0.5625913 0.5686281 0.5693529 0.5753497

DEMO

Thank You!