Image Caption Generation using Transformers

Luka Cvetko

11/27/2023

University of Notre Dame

CSE 40657/60657

Image captioning

- Generating a description of an image
- Automate the labeling process in digital image libraries
- CBIR(Content Based Image Retrieval) – limited search
- Freedom to search different queries
- Goal – caption the image with proper relations between the objects
- Factors to consider
    1. Semantic understanding
    2. Natural Language processing

Transformers

SEQUENTIAL DATA

ATTENTION MECHANISM

PARALLELIZABLE AND EFFICIENT

ENCODER – DECODER ARCHITECTURE

Method

Dataset: Flickr8k – image + caption

1. Feature vectors – Inception V3 – no SoftMax
2. Transformer Encoder – generate sequence of words
3. Attention mechanism - refer to features which are relevant to current word
4. Token generation – until <end> token
5. Caption generated via token sequence
6. Evaluation – BLEU score

Results

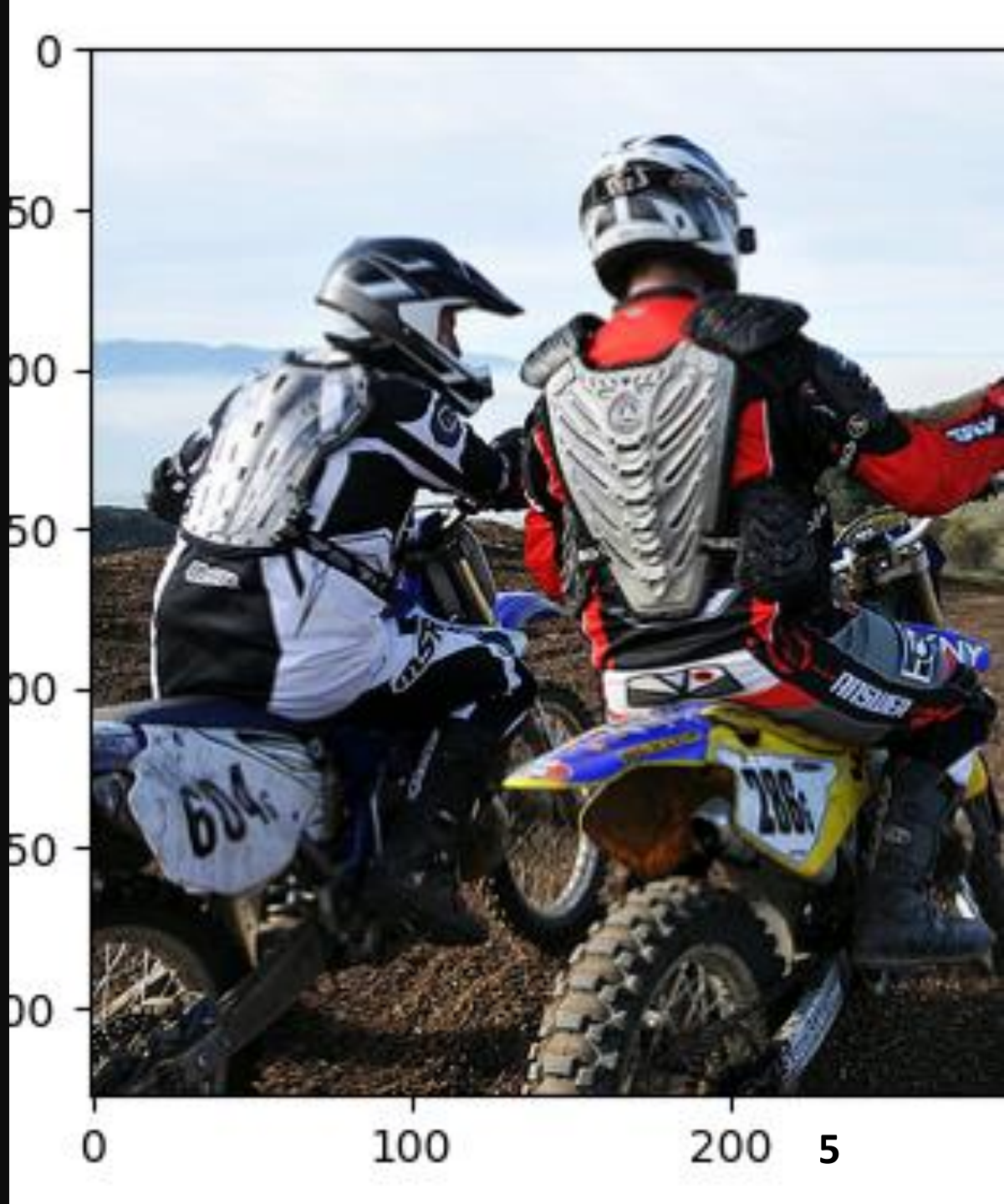
- Image dependent
- BLEU score:

1. Baseline -7.45%
2. Model - 25%

Actual caption - dirt bikers on trail

Model caption - two people on motorbikes

Baseline caption - mountain bike crash helmet alp motor scooter moped



Conclusions and improvements

- More complex scene – lower BLEU score
- Preprocessing the data to proper format
- Learning hyperparameters
- Epochs – only 30
- Bigger dataset – COCO – 12000 images

Q/A