PROJECT 3: “See & Say”: Captioning Images with CNN + RNN on Flickr8k

Business scenario

Automatic image captioning powers alt-text for accessibility, radiology report pre-drafts, and smart photo search. You will build a small-scale captioner that learns to describe everyday images—skills that transfer to medical imaging later.

Key Tasks and Tech:

- Combine **vision** (CNN encoder) and **language** (LSTM decoder) in a single pipeline.
- Experiment with **attention mechanisms** (“Show, Attend and Tell”) vs. vanilla CNN-LSTM.
- Evaluate captions with BLEU, METEOR, ROUGE-L, and CIDEr.

Dataset

- **Flickr8k** (8 k images + 5 captions each) – Kaggle link provided.
  Sizes: 6 k train / 1 k val / 1 k test.
- https://www.kaggle.com/datasets/adityajn105/flickr8k

Reference notebook:

https://www.kaggle.com/code/quadeer15sh/flickr8k-image-captioning-using-cnns-lstms/notebook

Tasks

1. **Feature extraction**
   - Use a pre-trained **ResNet-50** (or MobileNet V2) cut at the penultimate layer; cache features to disk.
2. **Caption decoder**
   - Tokenise captions, build <SOS>/<EOS> vocabulary (min-freq=5).
   - LSTM decoder with embedding dim = 256, hidden dim = 512.
   - Teacher forcing during training; greedy decode at test time.
3. **Training**
   - Cross-entropy loss (ignore padding tokens).
   - Early-stop on **BLEU-4** over the validation set.
4. **Evaluation & demo**
   - Report BLEU-1/2/3/4, METEOR, CIDEr on test set.
   - Feed an unseen image (provided by you) and show the generated caption.

5 Enhancements / stretch goals

- **Attention module** – implement the soft-attention from “Show, Attend and Tell”; compare metrics.
- **Beam search** or **nucleus sampling** at inference; analyse diversity vs. BLEU.
- Transfer-learn the decoder with **GloVe** initial embeddings or **GPT-2** language priors.
- Streamlit web demo: drag-and-drop an image → caption appears.
- Swap CNN-encoder for **CLIP ViT** pooling to test zero-shot robustness.

6 Deliverables

- Colab notebook with: data prep, model architecture diagram, training curves, and metric table.

- At least three captioned test images (ground truth vs. model).
- Presentation deck (template provided)
- README + requirements.txt.