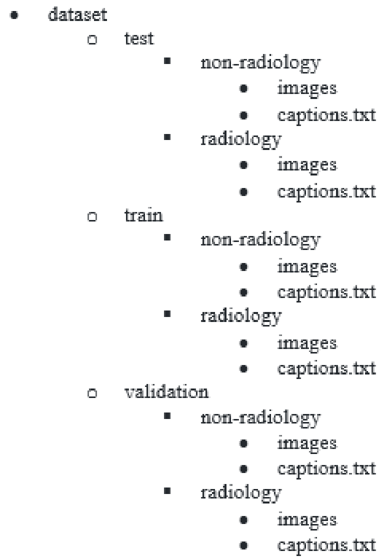
Medical image captioning

The ROCC dataset was created by Pelka et al. (2018)[1], aiming at detecting the interplay between visual elements and semantic relations present in radiology images. The dataset could be used for multi-modal image representations in classification tasks, for multi-class image classification and labeling, as well as medical image captioning. It was constructed by retrieving all image-caption pairs from the open-access biomedical literature database PubMedCentral and then, eliminating irrelevant images using a binary radiology and non-radiology classification. The dataset contains 81k radiology images with several medical imaging modalities and was used for ImageCLEF 2015 Medical Classification, and ImageCLEF 2013/2016 Medical Task. The dataset can be downloaded from https://www.kaggle.com/datasets/virajbagal/roco-dataset?select=all_data. The dataset repository includes three sets: train, validation, and test, where each of them consists of two folders: non-radiology and radiology. Restrict the data to be used for this project to the following structure (you do not need the remaining flies):



- dataset
    - test
        - non-radiology
            - images
            - captions.txt
        - radiology
            - images
            - captions.txt
    - train
        - non-radiology
            - images
            - captions.txt
        - radiology
            - images
            - captions.txt
    - validation
        - non-radiology
            - images
            - captions.txt
        - radiology
            - images
            - captions.txt

1. Download the dataset and visualize some training image/caption pairs of your choice from both classes (radiology and non-radiology).

We would like to create a simple caption generation model. For that, we try to construct an encoder-decoder model, where CNN is used to extract features from images and LSTM is used to generate sentences for a given test image. The model architecture is presented in Figure.1.

[1] Pelka O, Koitka S, Rückert J et al (2018) Radiology objects in context (roco): a multimodal image dataset. In: 7th joint international workshop on computing and visualization for intravascular imaging and computer assisted stenting, CVII-STENT 2018, and the 3rd international workshop on large-scale annotation of biomedical data and expert label synthesis, LABELS 2018, held in conjunction with the 21th international conference on medical imaging and computer-assisted intervention, MICCAI 2018 11043:180–189.

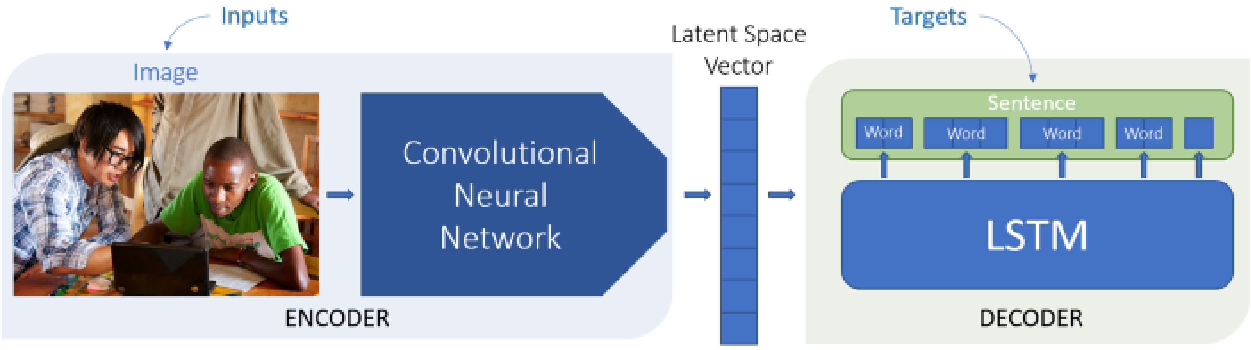


Figure 1: encoder-decoder model

Start with a simple code from https://medium.com/@stepanulyanin/captioning-images-with-pytorch-bc592e5fd1a3 to construct the captioning model. The main goal is to create a sentence that describes the visual content of the image. For that, follow these steps:

2. Keep the encoder and decoder architectures as the example (you can change the densnet121 into another pretrained model such as vgg16 or resnet50).
3. Create the vocabulary from the ROCO dataset (you can limit to only 2500 images and their captions for the training set and 500 for the validation set):
   - ✓ First, clean the captions by punctuation removal, stop words removal, lowercasing, tokenization and stemming.
   - ✓ Plot the word occurrence frequency curve after ranking the tokens. Check whether a power-law distribution can be fitted or not by plotting the log-log curve. Explain the results.
   - ✓ Calculate embeddings for the captions using word2vec and glove.
   - ✓ Visualize part of the word embedding space. Explain the results.
4. Create train_data_loader and val_data_loader from the training and validation sets, respectively. The loaders allow us to load the images and their associated captions to the model by batches.
5. Change the training loop to train the implemented model using the ROCO dataset.
6. Create a function to test your model on some samples from the test set (10 samples).
7. Calculate the similarity between the newly generated captions and the original captions using three different metrics.
8. Identify appropriate literature in the field of medical image captioning to provide reasonable findings of the results in the previous steps.