Naive and object detection for detailed image captioning

Omar, Alexio, Jimmy, Andres

University of Lorraine

November 18, 2022

Overview

1. Reminder
2. Goal of the project
3. Data Pre-processing
4. Text quality enhancement
5. Future works

Reminder

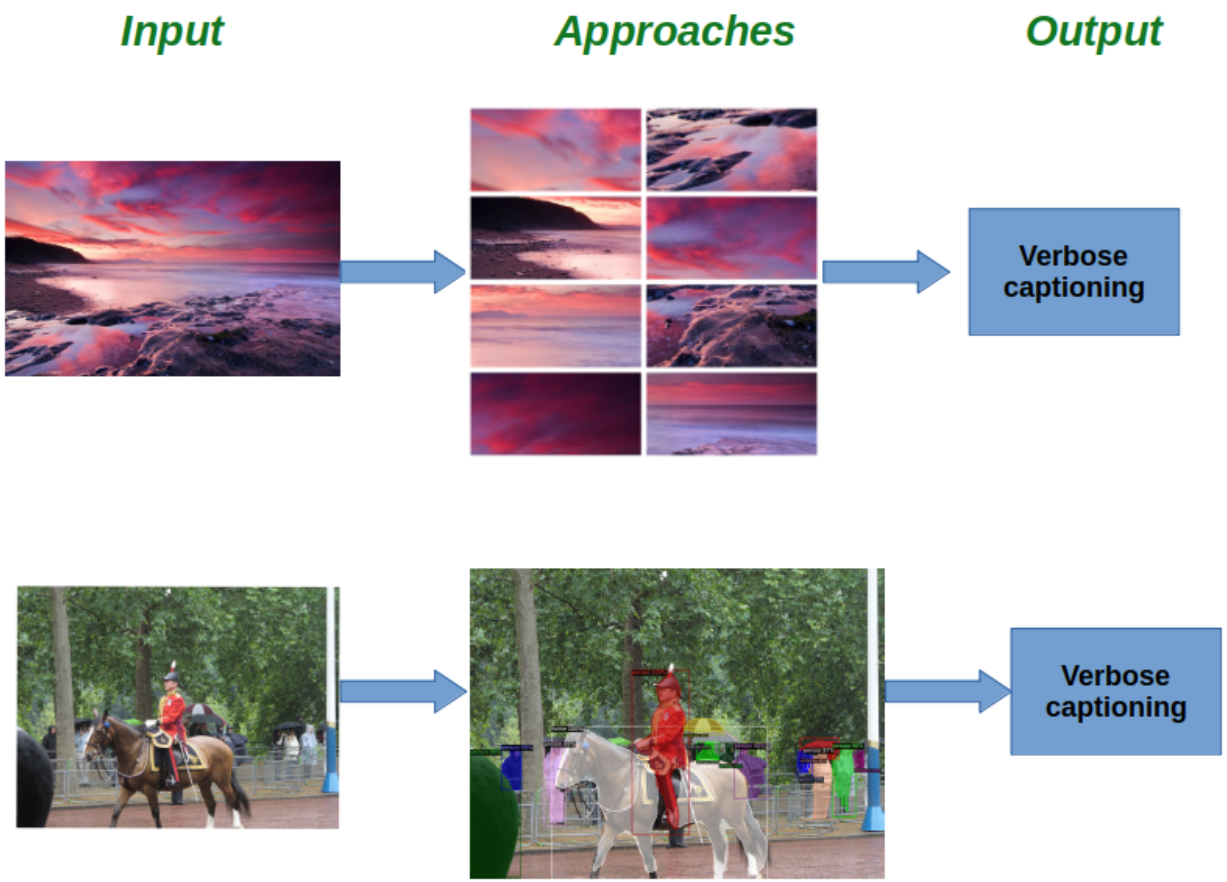

Figure 1: Image cutting and semantic detection

Goal of the project

- Image description for visually impaired people
- Make it easier to create dataset for image generation by allowing longer text as description.

Automating data loading

Goal

To standardize our images, regardless of type, for our downstream tasks.

1. Resizing
2. Cropping (previously shown)
3. By object detection
   - This enables crops according to detected objects (in-progress)
   - It cannot be standardized across data sets

Pipeline visualization

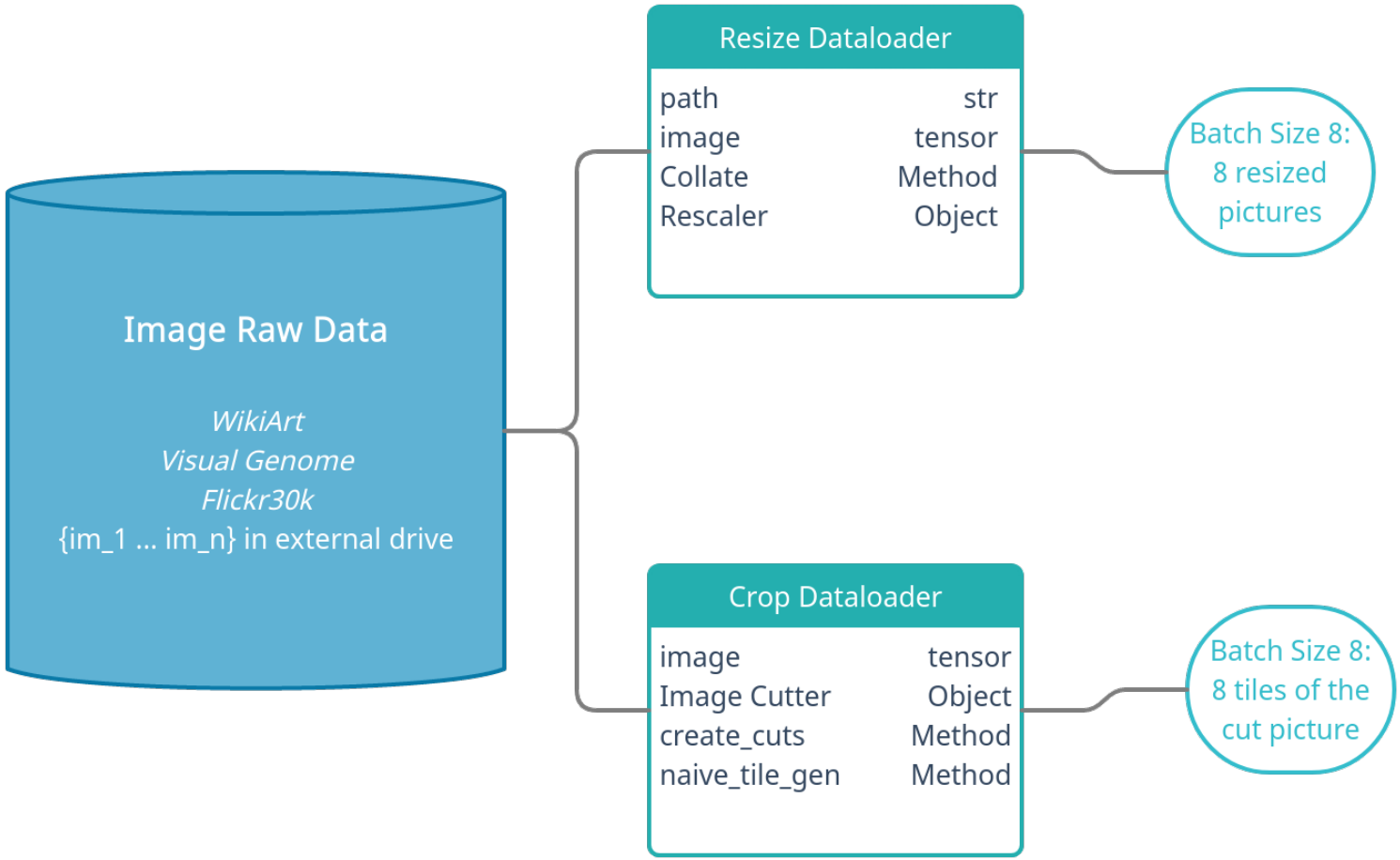

Figure 2: Different dataloaders depending on goal

Diving into Resize

Key functionality stems from the collate function, which has been redefined to take two scaler methods:

- Min
- Max

Diving into Resize

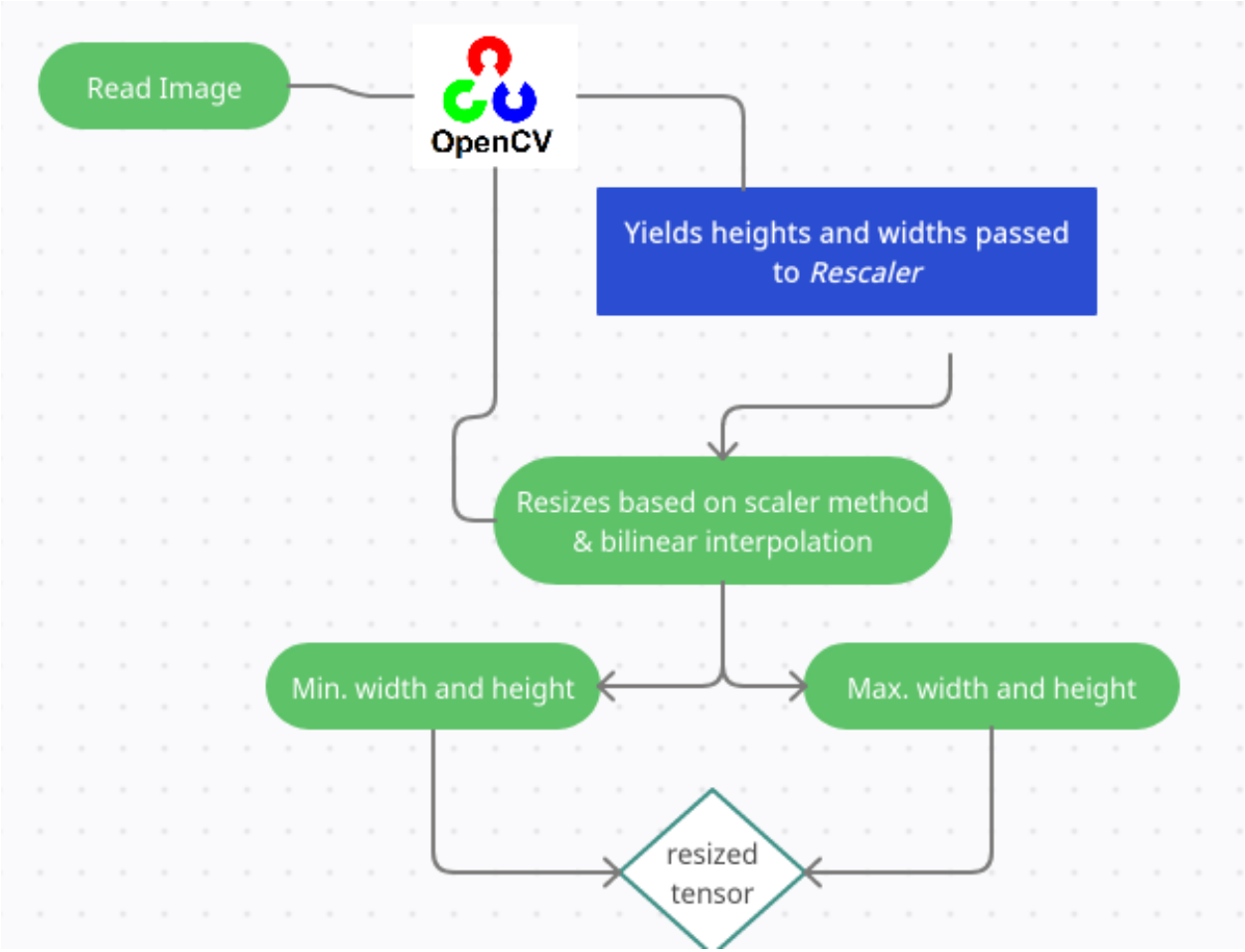

Figure 3: The collate/resize algorithm

Diving into Resize

- **Interpolation** Computes extra pixels in target image, but if it's smaller then the larger image has pixels that are not the same as nearby pixels. These are ignored. [1]
  - Bilinear interpolation is better for enlargement. It performs the interpolation in 2 dimensions and predicts the function to compute the color of the pixel.
  - What about the other methods? Nearest neighbor interpolation blurs out images, but retains sharpness of edges
  - Inter Area can be better for decimating images. It should be tested if only used the min scaler factor.
- **Collate** If we don't use a custom one, we need tensor data with the same dimensions [2]. We are padding dynamically during batch creation

[1] Interpolation in OpenCV Explained

[2] Collate in Torch Explained

Diving into Collate

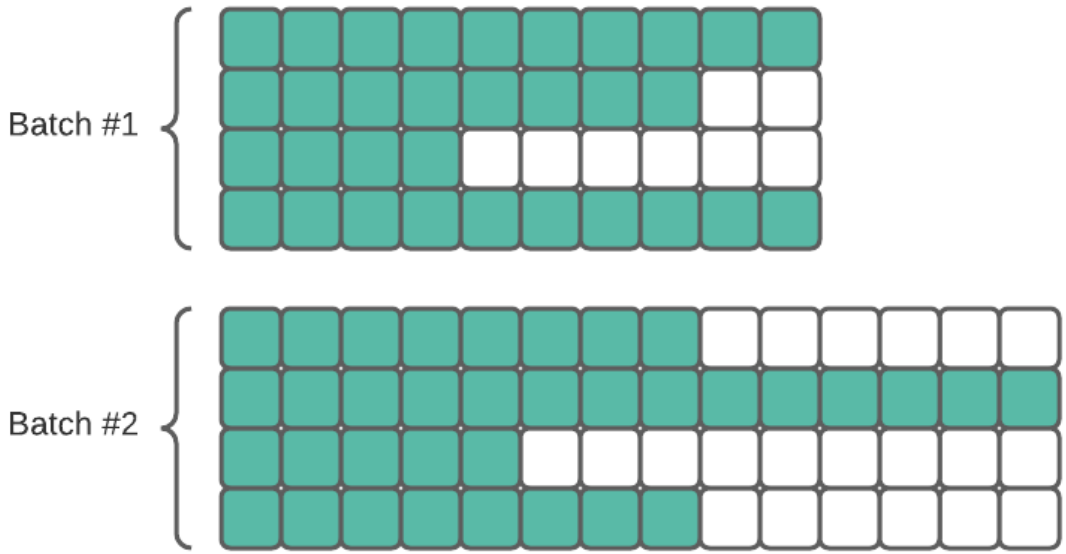

Figure 4: Image obtained from Understanding collate fn in Pytorch

No preprocessing

Figure 5: Air Terjun by Abdullah Suriosubroto

Resized

Figure 6: Downsized Painting

Cropping Image

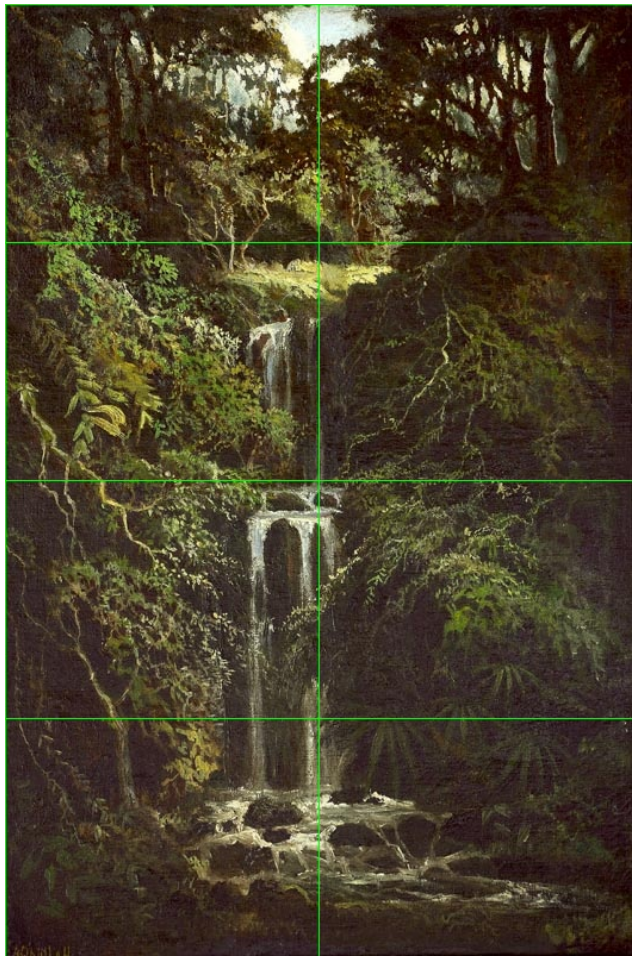

Figure 7: Naive cropped image

Idea

Upgrade the quality of the information from the sub-images captioning.
For example :

Figure 8: Image of a basket of apples

A basket of apples $\longrightarrow$ A dozen of bright red apples disposed in a basket. Some have their leaves.

Strategy

1. **Verbose captioning** *A basket of apples*
2. **Text completion** *A basket of apples , there are a dozen of bright red apples, some have leaves attached to them. The man takes the apples to eat them.*
3. **Paraphrasing** *A dozen of bright red apples disposed in a basket. Some have their leaves. They are taken by the man to be eaten.*
4. **Test adequacy with the image** *A dozen of bright red apples disposed in a basket. Some have their leaves.*

Then repeat this process several times until the size of the description is long enough.

Future Works

- Have a better captioning model.
- Use depth maps.
- Make the text enhancement work.
- Maybe do our own object detection model.
- Add analogies to select sub-images irrelevant for the paragraph generation