Image Captioning

Neural Networks and Deep Learning,
Artificial Intelligence, UAB, 2023

Eduard Florin Hogea
Júlia Garcia Torné
Josep Maria Rocafort Ferrer

CONTENT

PROJECT OVERVIEW

- Main objective: Develop a system that can generate natural language descriptions for input images.

→ Image Captioning Model →

" A boy hits a baseball with the bat. "

DATASET DESCRIPTION

Flickr 8k Dataset

8,000 Images

Multiple human-annotated captions for each image

" A man in an orange hat starring at something. "

" A person standing on a frozen lake. "

" A young boy runs aross the street. "

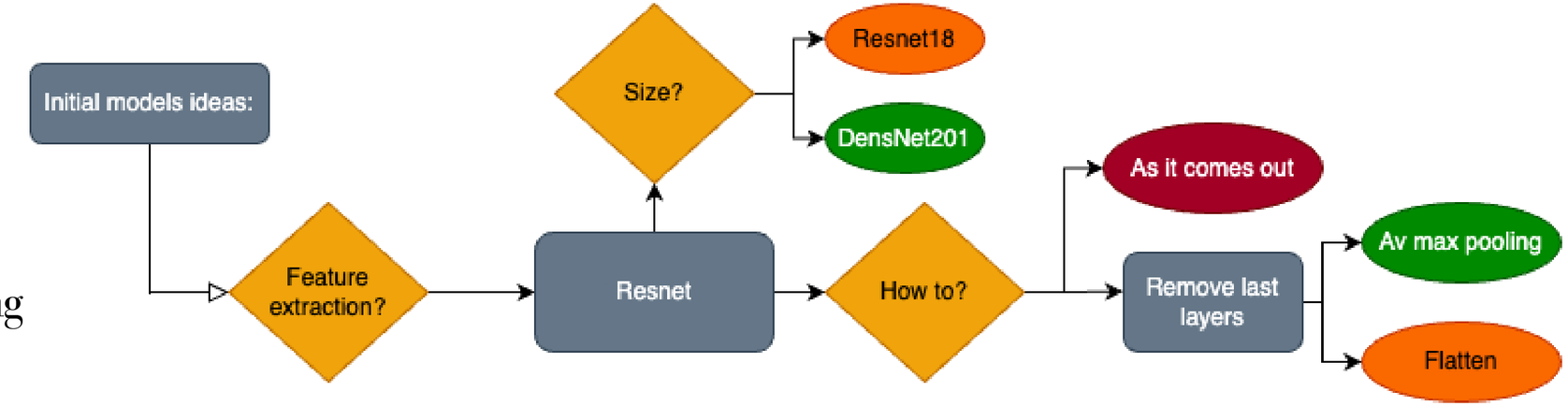
- 80 % for Training
- 20% for Validation

PROJECT FLOWCHART

Not working

Potentially working

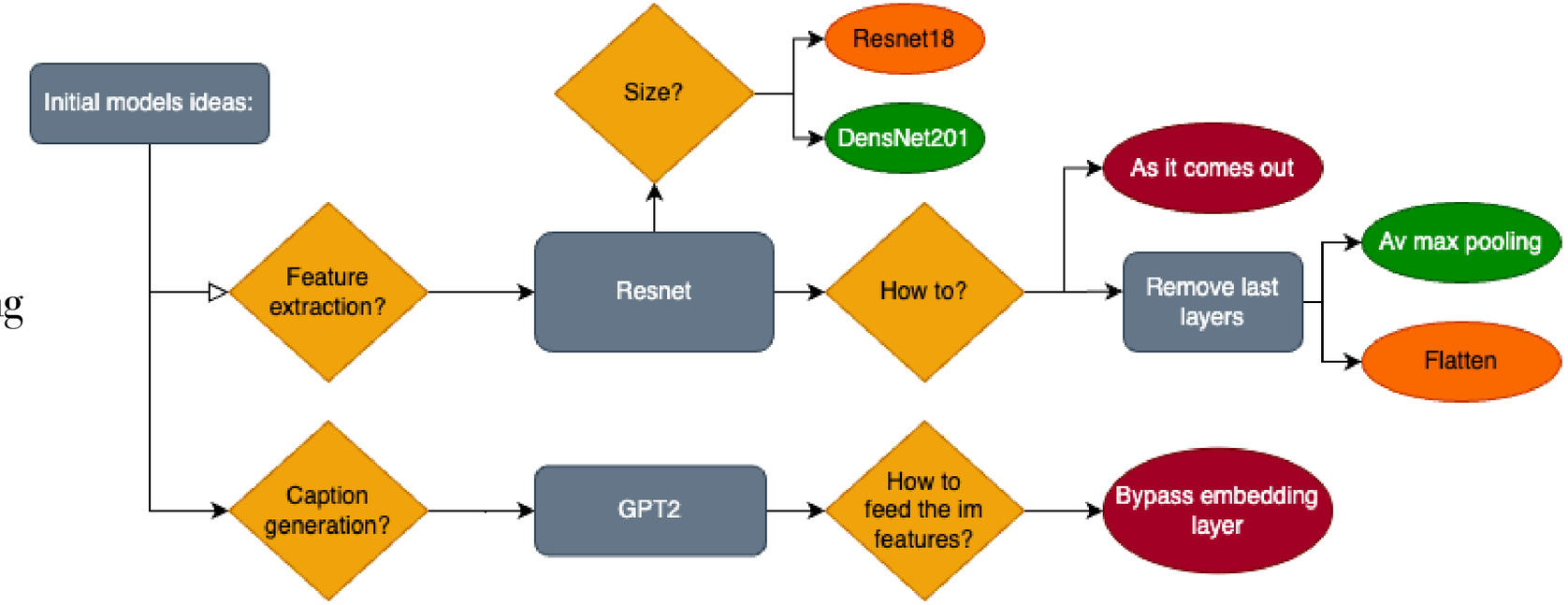
Working



Not working

Potentially working

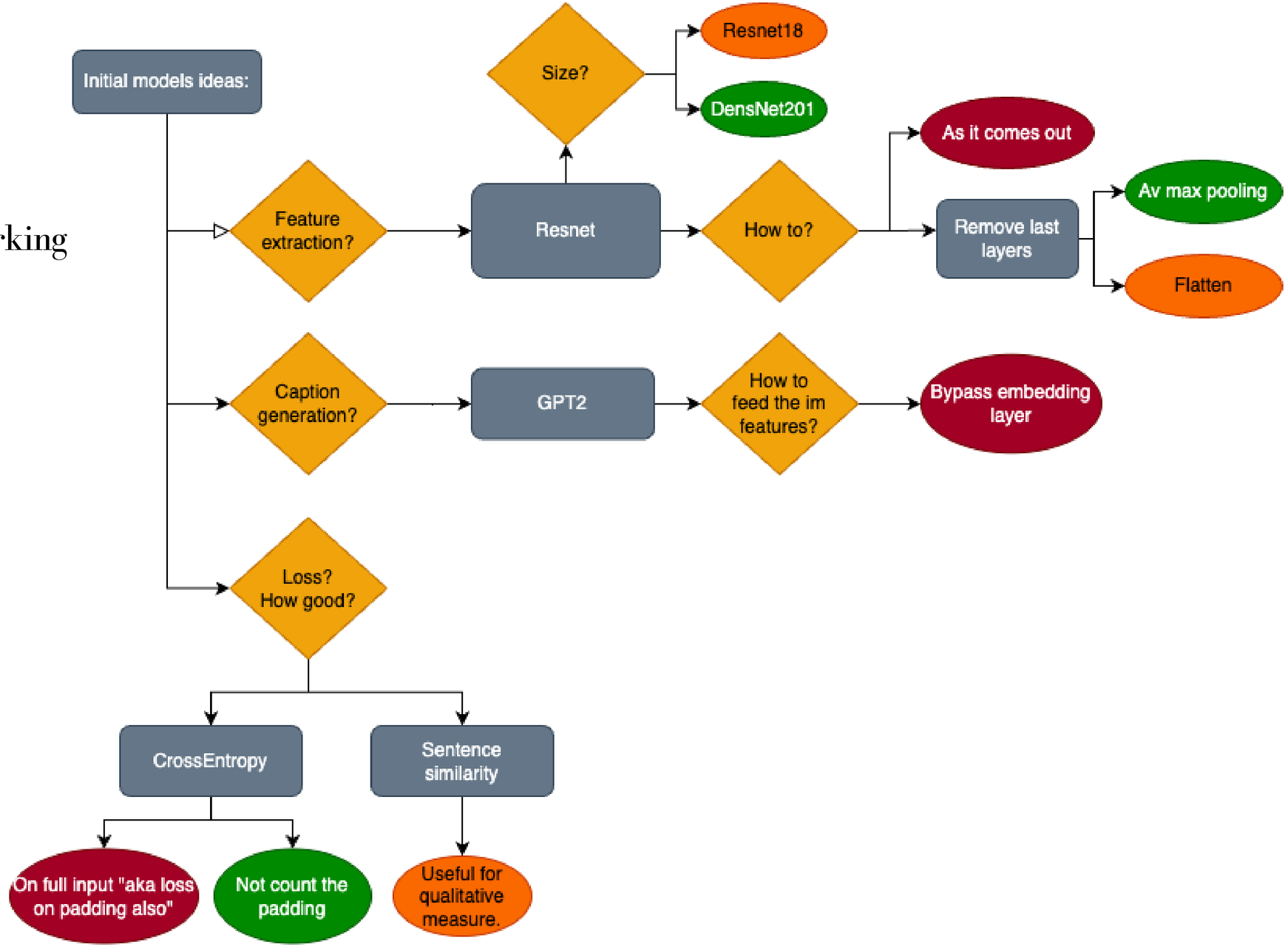
Working



Not working

Potentially working

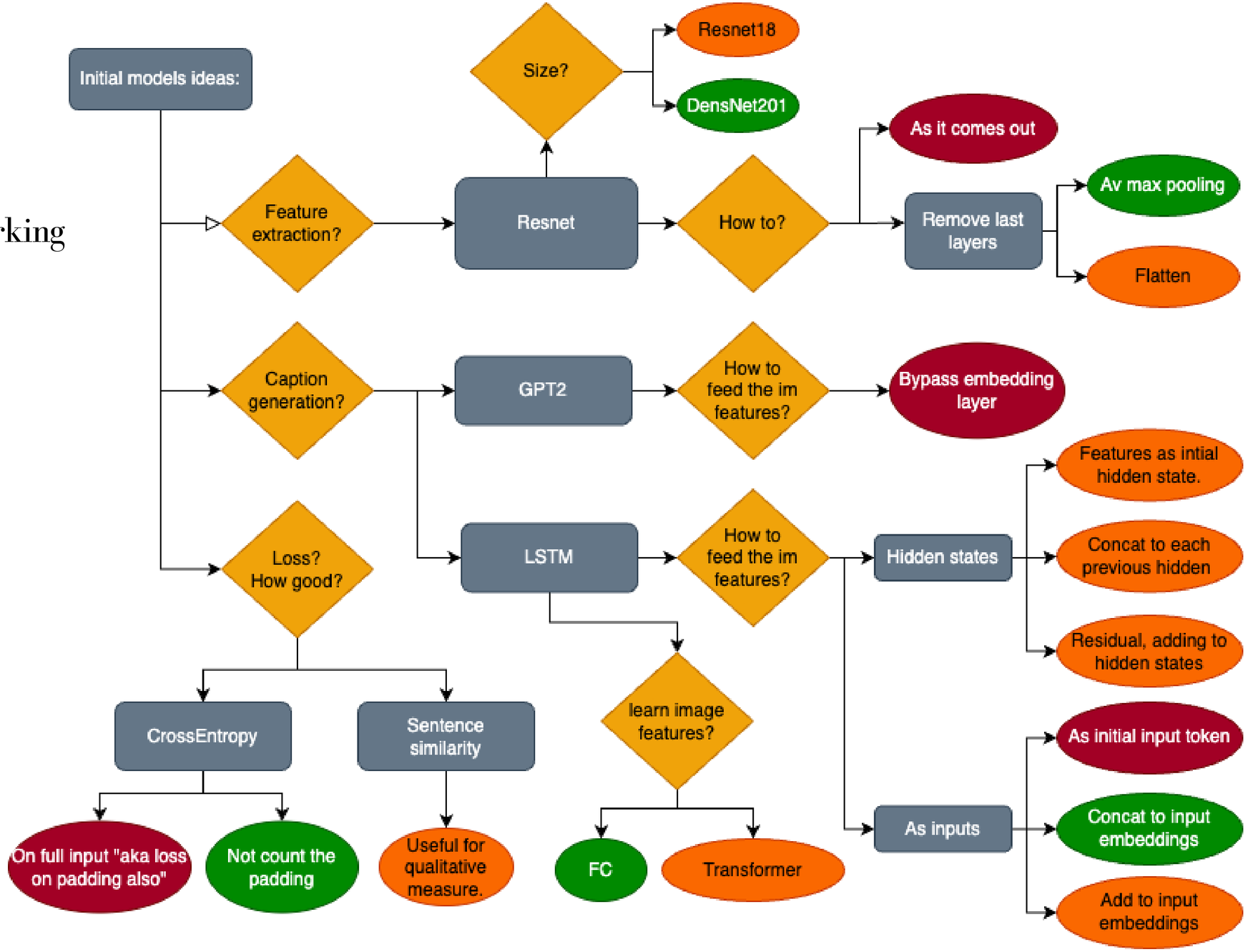
Working



Not working

Potentially working

Working



FINAL MODEL ARCHITECTURE

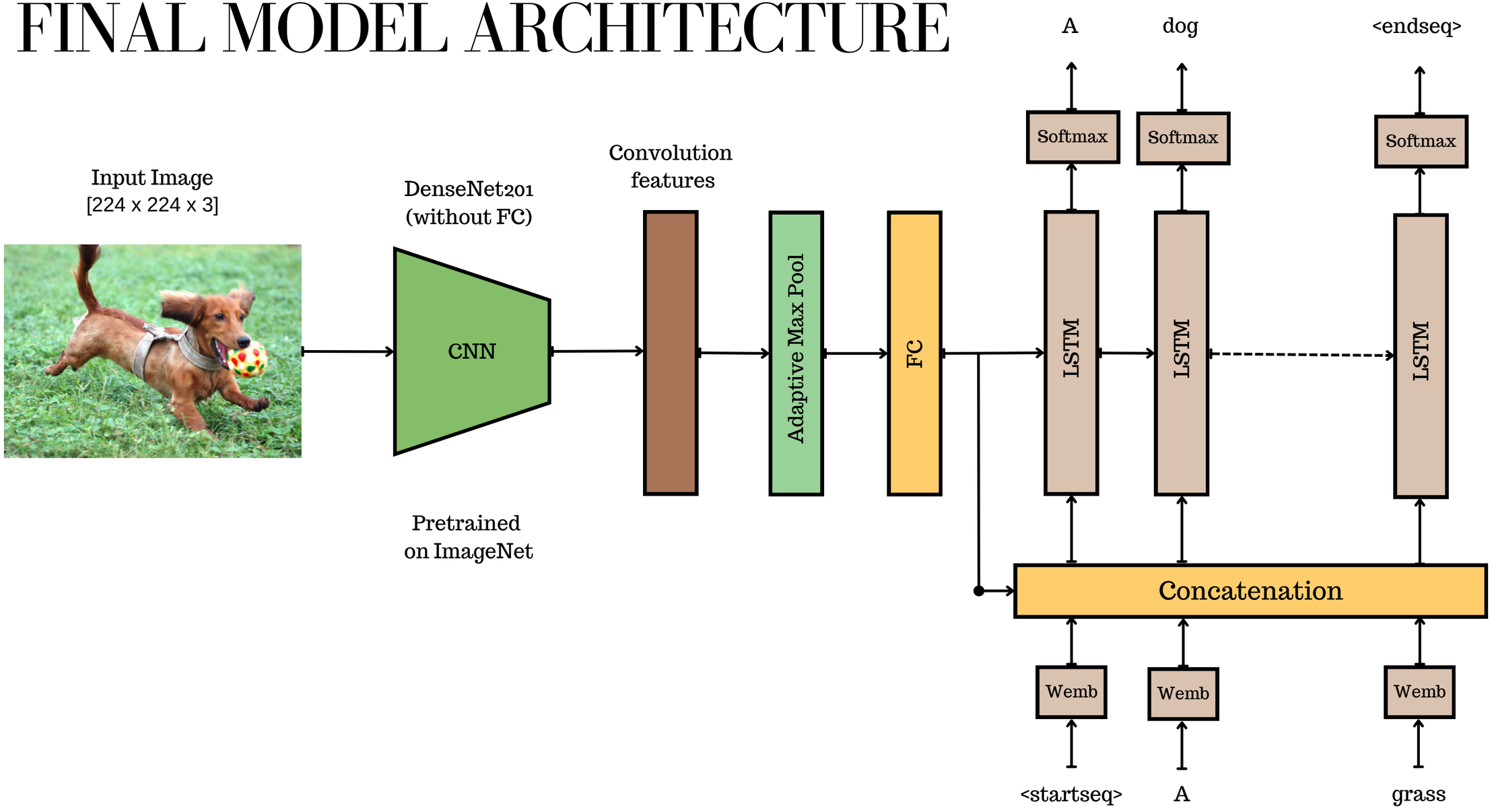
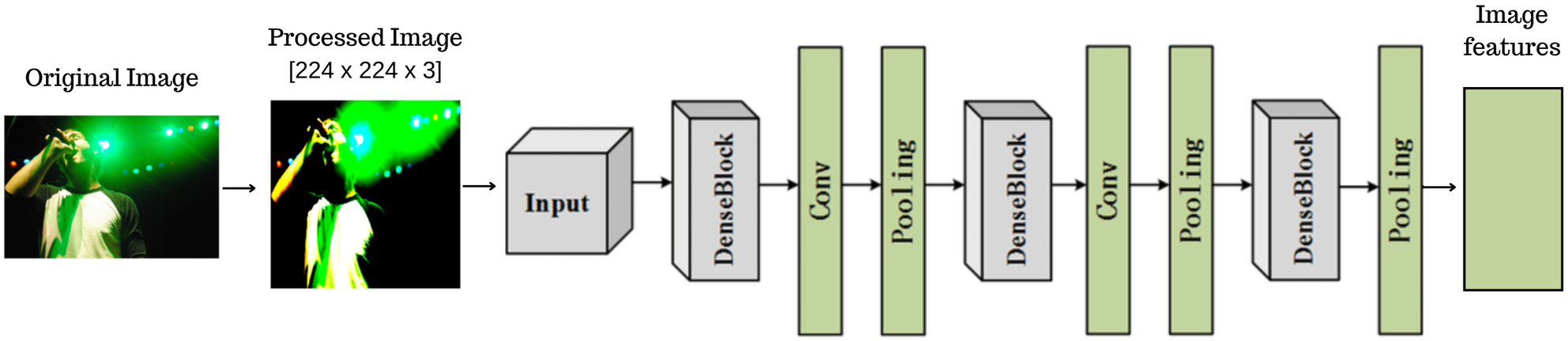


IMAGE FEATURE EXTRACTION

1. Pre-process images →

```
preprocess = transforms.Compose([
    transforms.Resize((224, 224)), # Resize the image: ResNet model  - > (224,224,3)
    transforms.ToTensor(), # Img to Python Tensor
    transforms.Normalize(mean=mean, std=std), # image = (image - mean) / std
])
```

2. Feature Extraction with a pretrained CNN: DenseNet201



DenseNet201 architecture

CAPTIONS PRE-PROCESSING

- Convert all text to lowercase.
- Remove non-letter characters: Delete punctuation marks, special characters, or any other non-alphabetical characters.
- Replace whitespace sequences.
- Remove single-letter words.
- Add start and end tokens: "startseq" and "endseq" tokens to the beginning and end of each caption.

" A child in a pink dress is climbing up a set of stairs in an entry way. "

↓

" startseq child in pink dress is climbing up set of stairs in an entry way endseq "

- Tokenization process:

The tokenizer will tokenize the caption by splitting it into words and assign a unique index to each word based on its position in the tokenizer's word index.

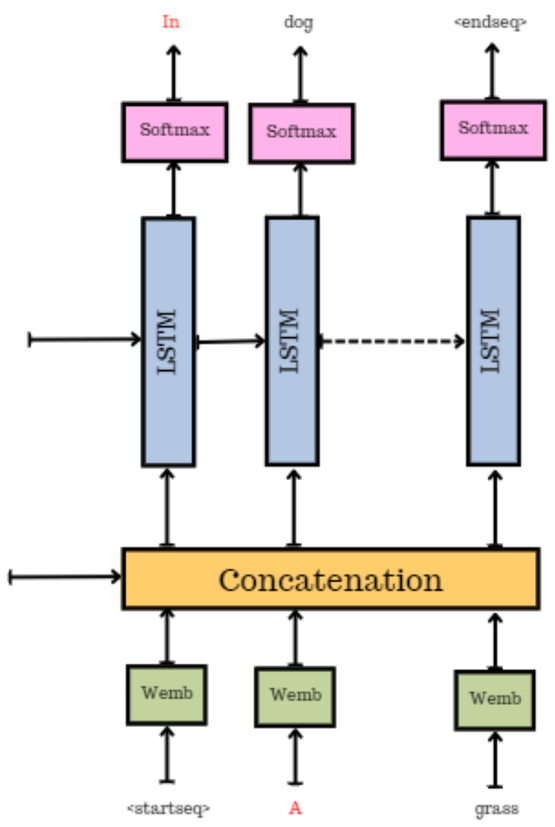
- The captions are transformed into tensors of length max_length.

" startseq child in pink dress is climbing up set of stairs in an entry way endseq "

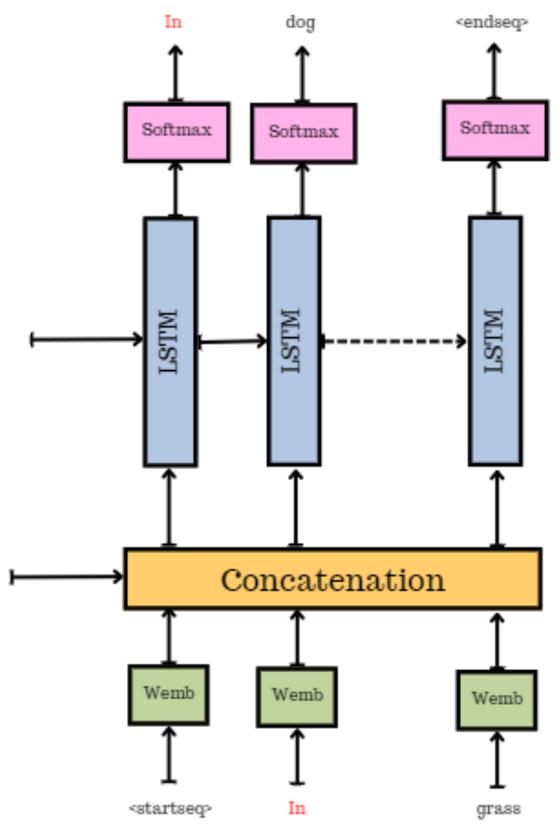
tensor([2, 43, 4, 90, 171, 7, 119, 51, 391, 12, 392, 4, 28, 5123, 668, 3, 0,0, ...])

TRAINING PROCESS: Two approaches

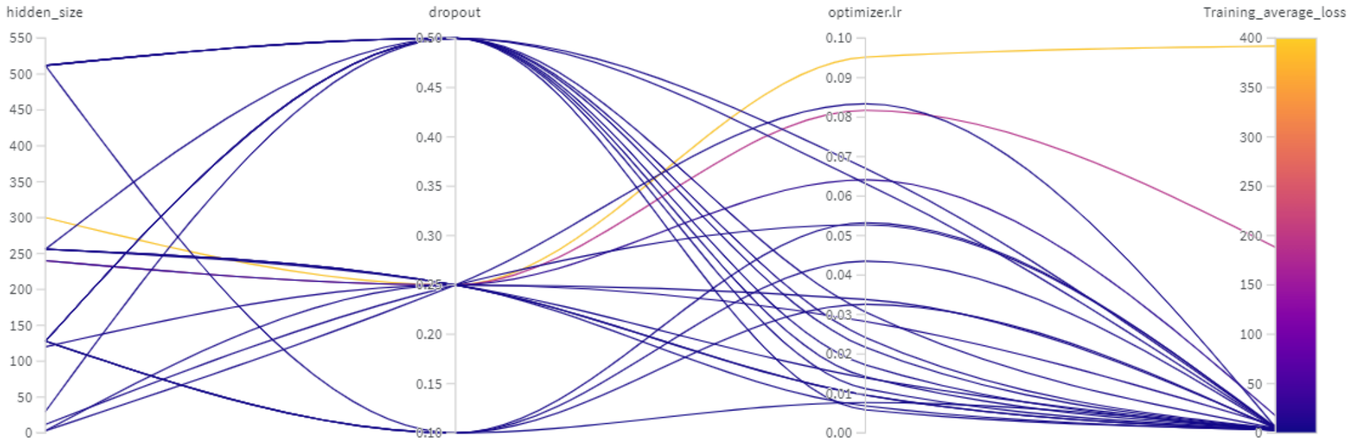
Ground Truth tokens as Input

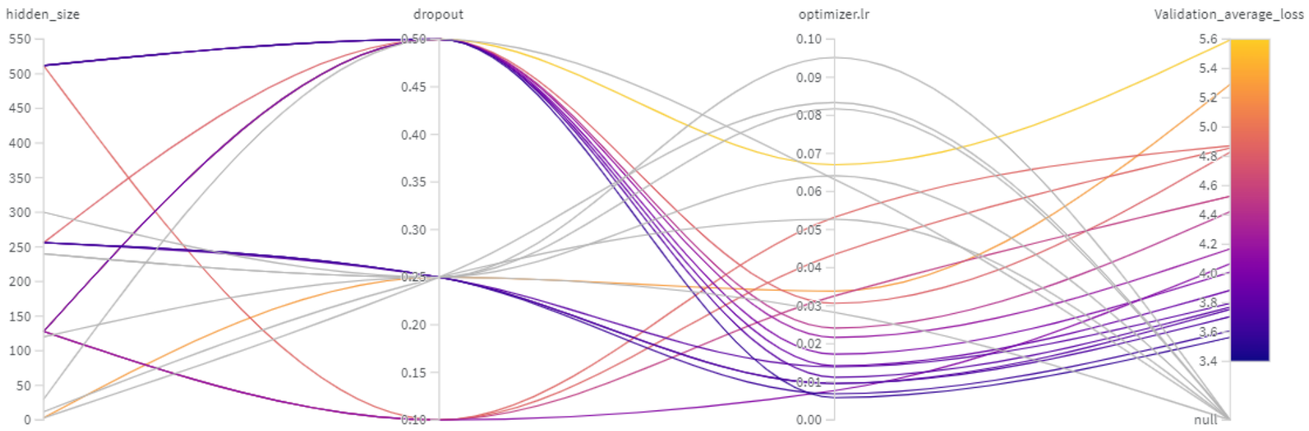


Learning from Model Predictions

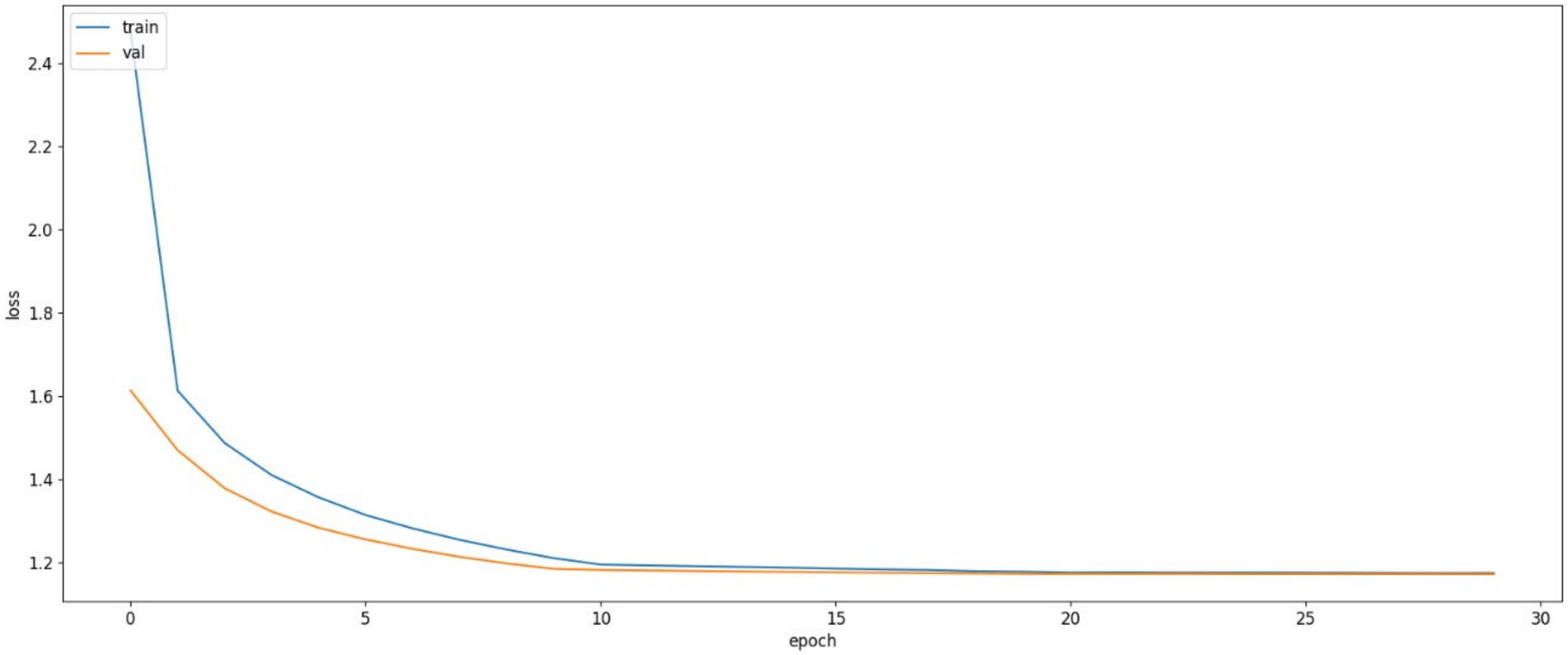


WEIGHTS AND BIASES: Optimizing Hyperparameters





Training and Validation Loss



RESULTS AND PERFORMANCE

- BLEU score

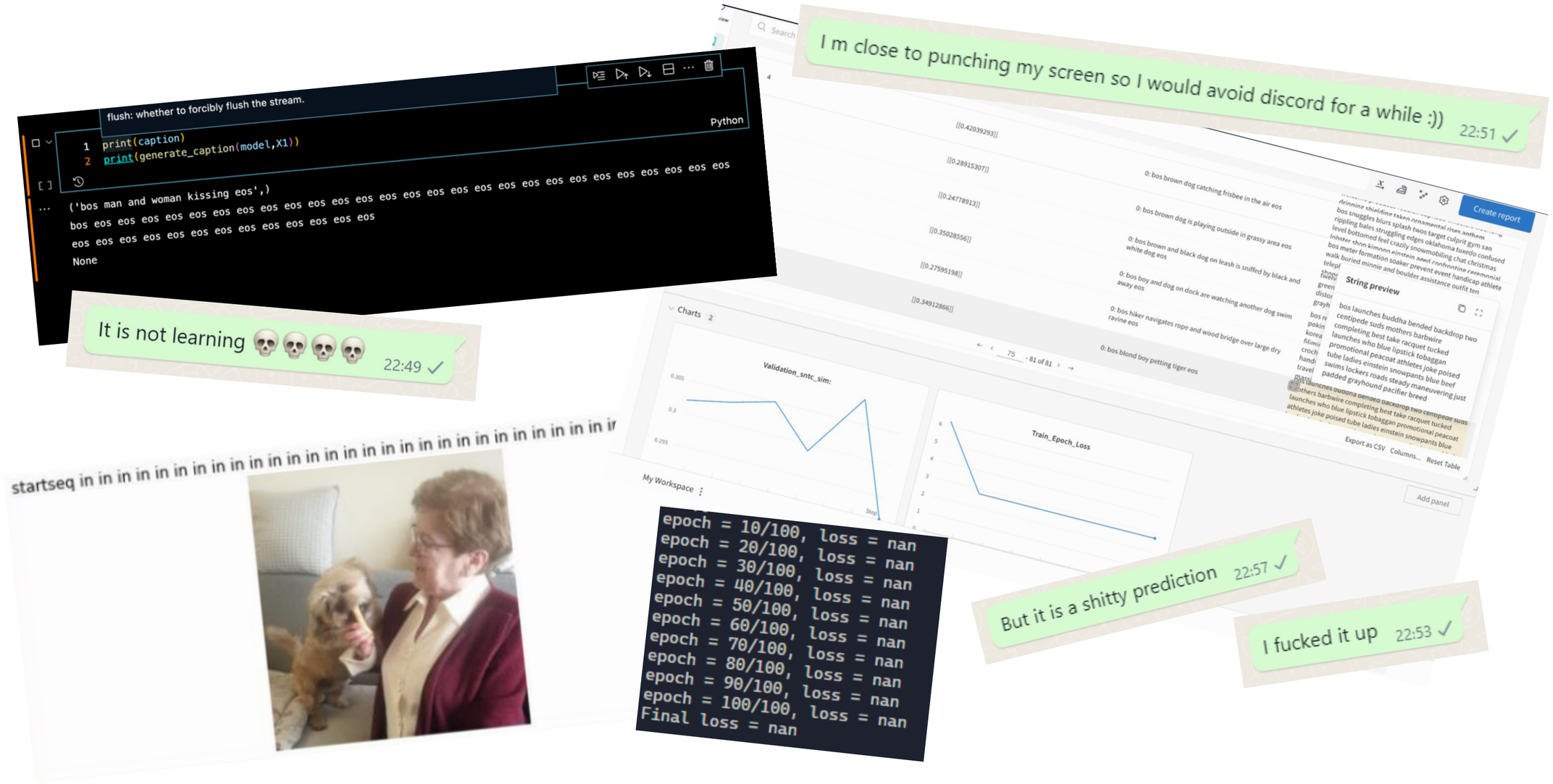
RESULTS: Testing on a different dataset

- Out-of-context dataset
- Poorly results

startseq on of and over three on beach in

startseq of the young down and ball is on runs in

CHALLENGES FACED



```
flush: whether to forcibly flush the stream.
print(caption)
print(generate_caption(model,X1))
('bos man and woman kissing eos',)
bos eos eos eos eos eos eos eos eos eos eos eos eos eos eos eos eos eos eos eos eos
bos eos eos eos eos eos eos eos eos eos eos eos eos
eos eos eos eos eos eos eos eos eos eos
None
```

It is not learning 💀💀💀💀 22:49 ✓

I m close to punching my screen so I would avoid discord for a while :)) 22:51 ✓

[[0.42039293]]	
[[0.28915307]]	0: bos brown dog catching frisbee in the air eos
[[0.24778913]]	0: bos brown dog is playing outside in grassy area eos
[[0.35028556]]	0: bos brown and black dog on leash is sniffed by black and white dog eos
[[0.27595198]]	0: bos boy and dog on dock are watching another dog swim away eos
[[0.34912866]]	0: bos hiker navigates rope and wood bridge over large dry ravine eos
	0: bos blond boy petting tiger eos

75 - 81 of 81

bos snuggles blurs splash twos target culprit gym san rippling bales struggling edges oklahoma tuxedo confused level bottomed feel crazily snowmobiling chat christmas bos meter formation soaker prevent event handicap athlete walk buried minnie and boulder assistance outfit ten

String preview

bos launches buddha bended backdrop two centipede suds mothers barbwire completing best take racquet tucked launches who blue lipstick tobaggan promotional peacoat athletes joke poised tube ladies einstein snowpants blue beef swims lockers roads steady maneuvering just padded grayhound pacifier breed

startseq in in in in in in in in in in in in in in in in in in in in in in in in in in in in

```
epoch = 10/100, loss = nan
epoch = 20/100, loss = nan
epoch = 30/100, loss = nan
epoch = 40/100, loss = nan
epoch = 50/100, loss = nan
epoch = 60/100, loss = nan
epoch = 70/100, loss = nan
epoch = 80/100, loss = nan
epoch = 90/100, loss = nan
epoch = 100/100, loss = nan
Final loss = nan
```

But it is a shitty prediction 22:57 ✓

I fucked it up 22:53 ✓

CONCLUSIONS

- Successfully developed a caption generation system.
- The system's performance did not meet our highest expectations.
- Acquired knowledge.
- Future improvements.