Face2Text

J. Neeraja, Eshwar Nukala, Shrey Jani
IITG.AI

Introduction

Recent development in deep learning has made a lot of problems comprehensible which were previously inconceivable. Writing a computer programme which takes as input an image and outputs its description is one such problem. In Face2Text, we tackle a derivative of this problem where input is an image of a face and output is the description capturing various facial features and emotional state.

Our motivation:

- To use transfer-learning and either establish its usefulness or inability to yield result in our task.
- To explore the use of both Computer Vision(CV) and Natural Language Processing(NLP) in novel real world applications.
- The machine generated description can act as an aid to the face recognition task in many real world applications such as describing facial features to the blind.

(a) Male example

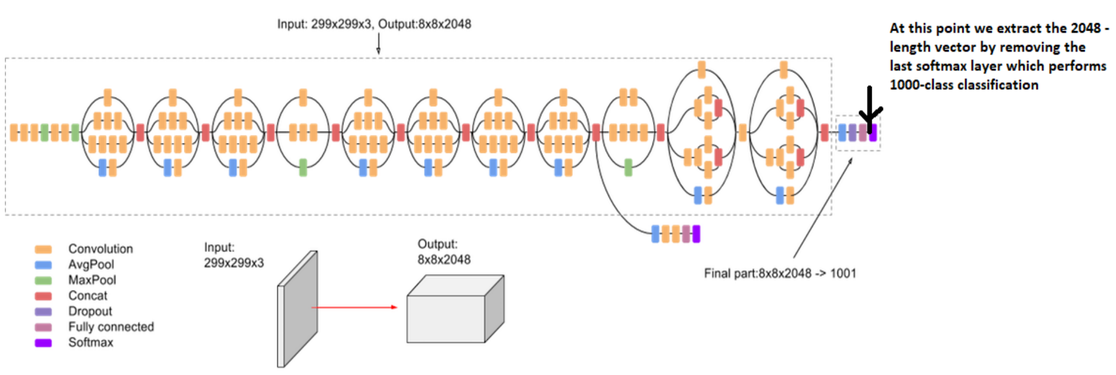
- I see a serious man. Such facial expressions indicate that the man is highly committed and dedicated to his work
- A middle eastern gentleman struggling with an administrative problem
- criminal
- Longish face, receding hairline although the rest is carefully combed with a low parting on the person's left. Groomed mustache. Could be middle-eastern or from the Arab world. Double chin and an unhappy face. Very serious looking.

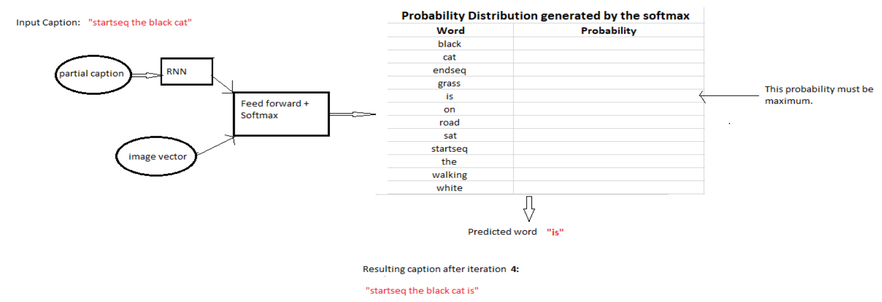
(b) Female example

- blonde hair, round face, thin long nose
- While female , American stylish blonde hair and blue or green eyes wearing a suit , public speaks person
- Middle aged women, blond (natural ?) well groomed (maybe over groomed). Seems to be defending/justifying herself to a crowd/audience. Face of remorse/regret of something she has done.
- An attractive woman with a lovely blonde hair style, she looks pretty seductive with her red lips. She looks like a fashion queen for her age.

Figure 1: Examples of descriptions collected for two faces.

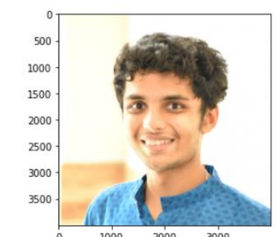
Techniques



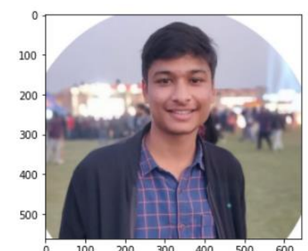


We used Inception-net model to extract image features into a vector which along with an 'in-sequence' is inputted into a LSTM layer. The LSTM layer predicts the next word ('out-sequence') in the description which in turn serves as 'in-sequence' for the next prediction. The whole description is generated in 'one word at a time' fashion.

Results

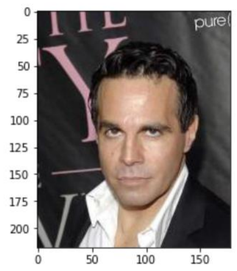


a young man with short brown hair and small dark e yes his nose is small and his lips are thin he is smiling and his upper teeth are visible

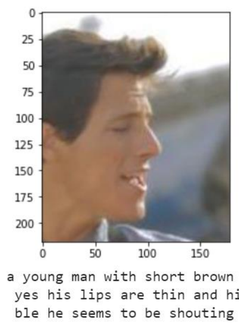


a young man with short dark hair and small dark ey es his nose is small and his lips are thin he is s miling and his upper teeth are visible

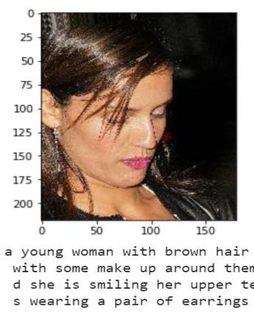
a young woman with brown hair and small dark eyes with some make up around them her lips are thin an d her upper teeth are visible she is smiling



a young man with short dark hair and small dark ey es his lips are thin and a stubble beard is growin g around them he looks serious



a young man with short brown hair and small dark e yes his lips are thin and his upper teeth are visi ble he seems to be shouting



a young woman with brown hair and small dark eyes with some make up around them her lips are thin an d she is smiling her upper teeth are visible she i s wearing a pair of earrings

Conclusion

- The model performed very well at generating grammatically correct and coherent sentences.
- Gender was predicted with a high accuracy
- There is not a lot of diversity the description structure. We believe this to be a dataset related issue. This may be overcome by using novel NLP models such as *attention to some extent*.
- We used cross-entropy loss function to train our model. After training for 20 epochs the loss function settled at 1.0631.

References

- Face2Text: Collecting an Annotated Image Description Corpus for the Generation of Rich Face Descriptions
- What is the Role of Recurrent Neural Networks (RNNs) in an Image Caption Generator?