Scene Description

Abirami Dhayalan
The University of Texas at Dallas
abirami.dhayalan@utdallas.edu

Arjun Sridhar
The University of Texas at Dallas
arjun.sridhar@utdallas.edu

Batul Petiwala
The University of Texas at Dallas
batul.petiwala@utdallas.edu

1. Problem Statement

Scene description is an issue that is still being looked at in Computer Vision today. Analyzing the environment and describing the scene can be useful in many applications. For our project, we want to caption an image, where the caption will accurately describe the actions and such taking place in the scene.

2. Approach

This project falls under the application-oriented category. Deep learning methods can be applied here. Specifically, CNNs will be used to identify key features present in the scene. For our project, we will try to expand these methods to accurately recognize a scene in an image, which can include multiple important features that can be used to describe the scene.

To get the text that describes the scene, we can apply image captioning methods using networks such as LSTM or RNN.

There are 2 approaches we can use. The first is a top-down approach, where we identify the description of the whole environment, and then identify key features in the environment that are used to describe the scene [2]. The second approach is a bottom-up approach, where we first identify key features present in the image, and then build up from those key features to produce a description that describes the overall scene present [2].

3. Data

The data we will be using initially is the Places dataset, which consists of over 10 million images and more than 400 unique scenes [1]. Figure 1 shows an example of an image that is present in the dataset.

For the image above, it falls under the category of 'food_court' in the dataset, and has some attributes such as 'socializing', 'eating', and 'working'. Our model will attempt to describe this scene in a coherent way that accurately describes the scene, such as outputting 'a group of people eating and socializing' for example.

Figure 1. Example image from the dataset

4. Evaluation

Evaluation of Scene Description models can be performed using metrics such as BLEU, METEOR, ROUGE or CIDEr, all of which mainly measure the word overlap between generated and reference captions [3].

References

[1] Zhou, Bolei and Lapedriza, Agata and Khosla, Aditya and Oliva, Aude and Torralba, Antonio, "Places: A 10 million Image Database for Scene Recognition", 2017, IEEE Transactions on Pattern Analysis and Machine Intelligence

[2] P. G. Pawar and V. Devendran, ”Scene Understanding: A Survey to See the World at a Single Glance,” 2019 2nd International Conference on Intelligent Communication and Computational Techniques (ICCT), 2019, pp. 182-186, doi: 10.1109/ICCT46177.2019.8969051.

[3] Xu, Ke, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel and Yoshua Bengio. “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention.” ICML (2015).