Assignment 2
COL 775: Deep Learning. Semester II, 2022-23.
Due Date: Friday May 12, 2023. 11:59 pm (IST).

April 5, 2023

1 Temporal and Causal Reasoning from Videos

Humans are capable of understanding complex phenomena which are combined experiences obtained through various modalities(auditory, visual, tactile...). They are also quite adept at abstracting individual concepts from a large amount of data, and performing reasoning over them. Can machines be trained to perform such level of reasoning? For this purpose, researchers have proposed the CLEVRER dataset[Yi et al., 2020]. The CLEVRER dataset is a dataset on video-question answering. The dataset comes with questions pertaining to a video consisting of objects which interact in a complex manner. In this assignment, we wish to study the capabilities of Deep Learning based models to answer such questions.

1.1 CLEVRER

The name CLEVRER has been inspired from CLEVR [Johnson et al., 2016] dataset, which is a dataset consisting of questions on static images. However, CLEVRER extends this to videos which adds additional challenges, namely temporal and counterfactual reasoning. An example from the dataset can be seen here. The dataset consists of four categories of questions:

1. **Descriptive**:
   Question: What shape is the object that collides with the cyan cylinder?
2. **Explanatory**:
   Question: Which of the following is responsible for the gray cylinder's colliding with the cube? a) The presence of the sphere b) The collision between the gray cylinder and the cyan cylinder
3. **Predictive**:
   Question: Which event will happen next a) The cube collides with the red object b) The cyan cylinder collides with the red object

4. **Counterfactual**:
   Question: Without the gray object, which event will not happen? a) The cyan cylinder collides with the sphere b) The red object and the sphere collide

   For your convenience, the dataset has already been downloaded on HPC in `/scratch/cse/dual/cs5180404/col775/A2` along with a README to understand the format of the data.

1.2 Baselines[8 marks]

To get you started on the assignment, there are two baselines that we require you to train and test. The test set has also been given on HPC. You are required to generate the test labels and submit them on the evaluation server. The details have been described in the Competition section below.

1. **CNN+BERT**: An entire video can be broken down into frames, and each frame can be encoded separately using a ResNet[He et al., 2015] encoder. For encoding the question(and outputs), we can use a BERT[Devlin et al., 2019] based model. Using these powerful models, the task boils down to classification and the entire model can be trained end-to-end. There are various design decisions which can arise in this implementation including but not limited to:

   - Which version of BERT/ResNet works best?
   - How should we aggregate all the encoded frames of the video to form a unified representation of the entire video?
   - How to add temporal information to all the frames?

   We wish that you explore all these choices and make the best decisions to improve your models.

2. **VideoCLIP** As introduced in the lectures, CLIP is a model which can be used to understand visual-language concepts. Recently, VideoCLIP [Xu et al., 2021] has been introduced, which is aimed at understanding video-language concepts. Like CLIP, VideoCLIP has also been trained using contrastive learning. Pre-trained VideoCLIP model can be found here We expect you to fine-tune VideoCLIP on this dataset, and report its performance. The way CLIP can be used for classification problems has been covered in the lectures, and in a similar way, VideoCLIP can also be used.

1.3 Competition[12 marks]

The CLEVRER evaluation server has been hosted online and will be used for evaluating your scores. The submissions format has been explained with the help of a sample submission as well. For this part, the grading for the entire set

of submissions will be relative. Note that you only have 3 submissions allowed in a day, so be judicious with them and try to get started early. Optimize your models as well as you can!

1.4 Submission Instructions and Report

For all parts of the assignment, you must clearly describe all design choices and hyperparameters in your report. Include all train/val loss curves. Include any other interesting findings you obtain during your study. Ensure that you save your models. We might randomly evaluate certain submissions on the test set. For this, you need to upload your inference script and models. Detailed submissions instructions will be released on Moodle soon.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning, 2016.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding, 2021.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. Clevrer: Collision events for video representation and reasoning, 2020.