RecVis Project: Video Question Answering with Hierarchical Conditional Relation Networks

Nicolas DUFOUR
ENS Paris-Saclay
nicolas.dufourn@gmail.com

Abstract

The objective of this project is to reproduce and try to improve the results of the paper "Conditional Relation Networks for Video Question Answering." [1]. The goal is to be able to answer questions about some videos. The proposed architecture leverages a novel hierarchical architecture.

1. Motivation and Problem Definition

The task we are trying to solve consist in training a model that is going to be able to answer different questions about a video. The questions can span to a frame-wise question but also to the whole videos. For example, we can try to answer questions about an action. This needs a good spatio-temporal understanding from the model. Also, this task needs some semantic modeling to understand the question and formulate the correct answer.

2. Methodology

The first goal is going to reproduce the results from the original paper [1]. This paper introduces a novel RNN unit the Conditional Relation Network (CRN). The CRN allows to process the clips and condition the output according to other data such as the motion data of the clip, or the question. We will reproduce the results using Pytorch [2]. We will reproduce the results on 3 Datasets: MSRVTT-QA [3], MSVD-QA [4] and TGIF-QA [5].

Then we will try to use a BERT [6] model to replace the LSTM language model and see if we can obtain better results. Indeed, BERT is a more advanced model based on the Transformer [7] architecture and has allowed multiple breakthroughs in the NLP domain. We will use the pretrain model by Hugging Face [8]. We will use it as a feature extractor, and we can try to finetune it on the question dataset trying to predict randomly occluded words in sentences or next words.

We will also try to see what happen when we use other conditioning such as subtitles. We will try to evaluate the impact of subtitle conditioning on performances. We can either concatenate subtitle representation with clip motion or replace the motion conditioning by the subtitle conditioning. The dataset we will use is TVQA [9].

Finally, the initial CRN architecture is based on average pooling and a specific clip sampling. The idea is to try and see what happen when we replace these blocks by an attention mechanism [7]. Indeed, it could be interesting to select the clips and the features depending on their attention. The attention mechanism as allowed significant progress in many fields such as NLP, computer vision, and graph neural networks.

3. Datasets and evaluation

The 4 datasets we are going to use are MSRVTT-QA [3], MSVD-QA [4], TGIF-QA [5] and TVQA [9].

MSVD-QA dataset has a total number of 1,970 video clips and 50,505 question answer pairs.

MSRVTT-QA contains 10K video clips and 243k question answer pairs.

TGIF-QA contains 4 different types of questions: Repetition count, repeating action, state transition and framewise questions. The dataset consists of 165,165 QA pairs collected from 71,741 animated GIFs.

TVQA dataset is based on 6 sitcoms which gives 21,793 clips and 152545 questions.

Finally, we will use the accuracy metric to assess the results.

4. References

[1] Thao Minh Le, Vuong Le, Svetha Venkatesh, & Truyen Tran. (2020). Hierarchical Conditional Relation Networks for Video Question Answering.

[2] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., & Chintala, S.. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library.

[3] Xu, D., Zhao, Z., Xiao, J., Wu, F., Zhang, H., He, X., & Zhuang, Y.. Video Question Answering via Gradually Refined Attention over Appearance and Motion.

[4] Xu, D., Zhao, Z., Xiao, J., Wu, F., Zhang, H., He, X., & Zhuang, Y.. Video Question Answering via Gradually Refined Attention over Appearance and Motion.

[5] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, & Gunhee Kim. (2017). TGIF-QA: Toward Spatio-Temporal Reasoning in Visual Question Answering.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, & Kristina Toutanova. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, & Illia Polosukhin. (2017). Attention Is All You Need.

[8] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, & Alexander M. Rush. (2019). HuggingFace's Transformers: State-of-the-art Natural Language Processing.

[9] Lei, J., Yu, L., Bansal, M., & Berg, T.. (2018). TVQA: Localized, Compositional Video Question Answering.