From Recognition to Cognition : Visual Common Sense Reasoning

Purvanshi Mehta

November 26, 2019

1 Summary

1.1 Problem

Humans have the ability to infer what is happening in a scene beyond what is visually obviously and without any other context. Machines for now lack this capability. The authors address this problem by introducing the task of Visual Common sense reasoning.

1.2 Innovation

The previous datasets suffer from common sense bias and there is a lack of large datasets in the community because of the expensive labelling of data.

Firstly this issue is addressed by introducing a **Adversarial matching Scheme** and introducing **Recognition to Cognition Networks**

1.3 Contributions

The contribution of the paper are three-fold

- Dataset Collection - VCR dataset has been collected from movie and video clips and has been filtered through an *interestingness filter*. Annotations were crowd sourced through Amazon Mechanical Turk
- A new method was formulated which allows for any 'language generation' dataset to be turned into Multiple Choice Test by using *Maximum-weight bipartite matching*
- The task of predicting the correct answer along with the correct explanation was divided into three modules - 1) Grounding - forming combined embeddings of the image and the language 2) Contextualization - Attention on the query and response and 3) Reasoning - BiLSTM was used to reason over the attended to produce the desired output.

1.4 Evaluation

Strong baselines were used to evaluate the VCR dataset. The recognition to cognition network achieves state of the art on the dataset. Their claims are also supported by strong albation studies.

1.5 Substantiation

Their claim of biasness in the previous datasets is solved to some extent as they provide 4 rationales or explanations for each of the option chosen by the network. they also consider a prediction as correct if the network predicts both the explanation and the answer correct.

2 Review

- **Using Multimodal Embeddings** - The Grounding module which finds combined image and query module by using BiLSTM can be improved by initializing with multimodal representations. Which explicitly means - *Person and (its image)* could be initially brought into shared representation instead of extracting features individually and then passing through a BiLSTM. Various techniques for shared embeddings could be used like DeVise, using CCA (Canonical Correlation Analysis) techniques to bring the embedding space as close to each other as possible.

- Two models have been trained - First for generating the answer ($Q- > A$) and then from the query and the answer - the rationale ($QA- > R$). An extension could be to directly learn one model which gives us the rationale and the answer i.e. Q− >AR - instead of doing this in two steps we can make this a one step model

- Can we leverage **pre training** on other image and language datasets to improve the accuarcy of the Visual common sense reasoning dataset? Another question to ask here is that does using this dataset as a pre-training module increase common sense reasoning on other datasets.

- Common sense in humans is also derived from **world priors**. In the current VCR dataset a image is provided and a model is expected to learn common sense from the image. Can priors be added to learn common sense? Such a task can be handled by learning important semantic concepts from ConceptNet.

- This point provides certain questions to ask before proceeding with research in this area. The worldly knowledge acquired by humans is unsupervised unlike the current Deep Learning techniques. Another criticism of such models is that they are highly data specific. There are several core questions to be asked in the domain of research -

– How can we learn common sense reasoning in an **Unsupervised** fashion? Can we use world knowledge sources like scene graphs and knowledge graphs to learn from unlabelled data.
– Is **question answering the best way** to learn common sense reasoning?
– Can we learn common sense from the large amount of unlabelled **video data** preset in today's world. As common sense can also be thought of as sequential. For example - If you throw a ball, it will eventually fall on the ground.