Introduction

The ability to use language to reason about every-day visual input is a fundamental building block of human intelligence. Achieving this capacity to visually reason is thus a meaningful step towards artificial agents that truly understand the world (cbn paper).

A model which is made from general-purpose components and can learn to visually reason, will likely be more widely applicable across domains (film paper).

One of the ways to evaluate such a model is by using A Diagnostic Dataset for  
Compositional Language and Elementary Visual Reasoning (CLEVR), which is used to test visual reasoning via question answering (CLEVR website).

<https://cs.stanford.edu/people/jcjohns/clevr/>

Visual question answering is a general task of asking questions about images, has its own line of datasets which generally focus on asking a diverse set of simpler questions on images, often answerable in a single glance. From these datasets, several effective general-purpose deep learning models have emerged for visual question answering (Malinowski, Rohrbach, and Fritz 2015; Yang et al. 2016; Lu et al. 2016; Anderson et al. 2017). However, tests on CLEVR show that these general deep learning approaches struggle to learn structured, multi-step reasoning (Johnson et al. 2017a).

These models tend to exploit biases in the data rather than capture complex underlying structure behind reasoning (Goyal et al. 2017). In order to overcome this problem, film paper developed a general model architecture that can achieve strong visual reasoning which they termed as FiLM: Feature-wise Linear Modulation.

Film is a general-purpose conditioning method that is highly effective for visual reasoning, however one of its drawbacks is that it makes some logical mistakes that humans won’t do, for example: (example from film paper).

In this project we want to observe whether adding a CBN layer, which has proven highly effective for traditional visual question answering tasks[H. de Vries, F. Strub, J. Mary, H. Larochelle, O. Pietquin, and A. C. Courville, “Modulating early visual processing by language,” arXiv preprint arXiv:1707.00683, 2017. [Online]. Available: <http://arxiv.org/abs/1707.00683>] to a FiLM model can improve the performance and solve this specific FiLM’s drawback.

Method

The CLEVR's creators argues that artificial intelligence systems that can reason and answer questions about visual data, need diagnostic tests in order to analyze progress and be able to discover shortcomings. CLEVR presents a diagnostic dataset that tests a range of visual reasoning abilities. It contains minimal biases and has detailed annotations describing the kind of reasoning each question requires, and can be used to analyze a variety of modern visual reasoning systems, providing novel insights into their abilities and limitations.

Examples from CLEVR are shown in Figure 1.

A FiLM layer carries out a simple, feature-wise affine transformation on a neural network’s intermediate features, conditioned on an arbitrary input. In the case of visual reasoning, FiLM layers enable a Recurrent Neural Network (RNN) over an input question to influence Convolutional Neural Network (CNN) computation over an image. This process adaptively and radically alters the CNN’s behavior as a function of the input question, allowing the overall model to carry out a variety of reasoning tasks, for instance ranging from counting to comparing.

FiLM can be thought of as a generalization of Conditional Normalization, which has proven highly successful for image stylization (Dumoulin, Shlens, and Kudlur 2017; Ghiasi et al. 2017; Huang and Belongie 2017), speech recognition (Kim, Song, and Bengio 2017), and visual question answering (de Vries et al. 2017), demonstrating FiLM’s broad applicability. In this paper, which expands upon a shorter report (Perez et al. 2017), our key contribution is that we show FiLM is a strong conditioning method by showing the following on visual reasoning tasks:

1. FiLM models achieve state-of-the-art across a variety of visual reasoning tasks, often by significant margins.

2. FiLM operates in a coherent manner. It learns a complex, underlying structure and manipulates the conditioned network’s features in a selective manner. It also enables the CNN to properly localize question-referenced objects.

3. FiLM is robust; many FiLM model ablations still outperform prior state-of-the-art. Notably, we find there is no close link between normalization and the success of a conditioned affine transformation, a previously untouched assumption. Thus, we relax the conditions under which this method can be applied.

4. FiLM models learn from little data to generalize to more complex and/or substantially different data than seen during training.

Implementation

We introduce a FiLM model but instead of batch normalization layer we use a CBN layer as was presented by cbn paper.

Our goal in this project is to check if the new model that we built can produce better results than the FiLM’s results

Some have argued that for artificial agents to learn this complex, structured process, it is necessary to build in aspects of reasoning, such as compositionality (Hu et al. 2017; Johnson et al. 2017b) or relational computation (Santoro et al. 2017). However,