1 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 (Perez, de Vries, Strub, Dumoulin, & Courville, 2017).

A model which is made from general-purpose components and can learn to visually reason, will likely be more widely applicable across domains (Perez, Strub, De Vries, Dumoulin, & Courville, 2018).

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 (Johnson et al., 2017).

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., 2017).

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, Perez et al. (2018) 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 (De Vries et al., 2017) to a FiLM model can improve the performance and solve the above FiLM’s drawback.

2 Method and Implementation

Our model processes the multi-modal question-image input using a RNN and CNN combined via FiLM and Conditional Batch Normalization (CBN).

Firstly, we will start by explaining FiLM and CBN and next in order we will describe our model with its modifications and additions.

2.1 FiLM: Feature-wise Linear Modulation

FiLM learns to adaptively influence the output of a neural network by applying an affine transformation, to the network’s intermediate features, based on some input. More formally, FiLM learns functions f and h which output

and as a function of input:

where and modulate a neural network’s activations whose subscripts refer to the input’s feature or feature map, via a feature-wise affine transformation:

f and h can be arbitrary functions such as neural networks.  
(Perez et al., 2018)

2.2 CBN: Conditional batch normalization

BN has been shown to accelerate training and improve generalization by reducing

covariate shift throughout the network [18]. To explain BN, we define  
 as a mini batch of N samples, where F corresponds to input feature maps whose subscripts c, h, w refers to the feature map at the spatial location (h, w). We also define and as per-channel, trainable scalars and as a constant damping factor for numerical stability.  
BN is defined at training time as follows:

Conditional Batch Normalization (CBN) [14, 15, 16] instead learns to output new BN parameters and as a function of some input :

where f and h are arbitrary functions such as neural networks.  
(Perez et al., 2017)

2.3 Our Model

Our model consists of a linguistic pipeline and a visual pipeline as depicted in   
Figure 1. The linguistic pipeline processes a question q using a Gated Recurrent Unit (GRU) (Chung et al.2014) with 4096 hidden units that takes in learned,   
200-dimensional word embeddings. The final GRU hidden state is a question embedding, from which the model predicts for each residual block via affine projection. We also wrapped the GRU model with Pytorch’s Parallel model to increase processing speed. The visual pipeline extracts 128 14 x 14 image feature

maps from a resized, 224 x 224 image input using either a CNN trained from scratch or a fixed, pre-trained feature extractor with a learned layer of 3 x 3 convolutions. The CNN trained from scratch consists of 4 layers with 128 4 x 4 kernels each, ReLU activations, conditional batch normalization and dropout. The classifier is implemented as was presented by Perez and his collegues (2018).

Furthermore, we had to adjust the dataset, in the sense that we have reduced the data type to int32 down from int64 and used around half the dataset instead of all of it. These adjustments were necessary because of server limitations. The server wouldn’t let us run the model for more than 24hrs.

A screenshot of a cell phone

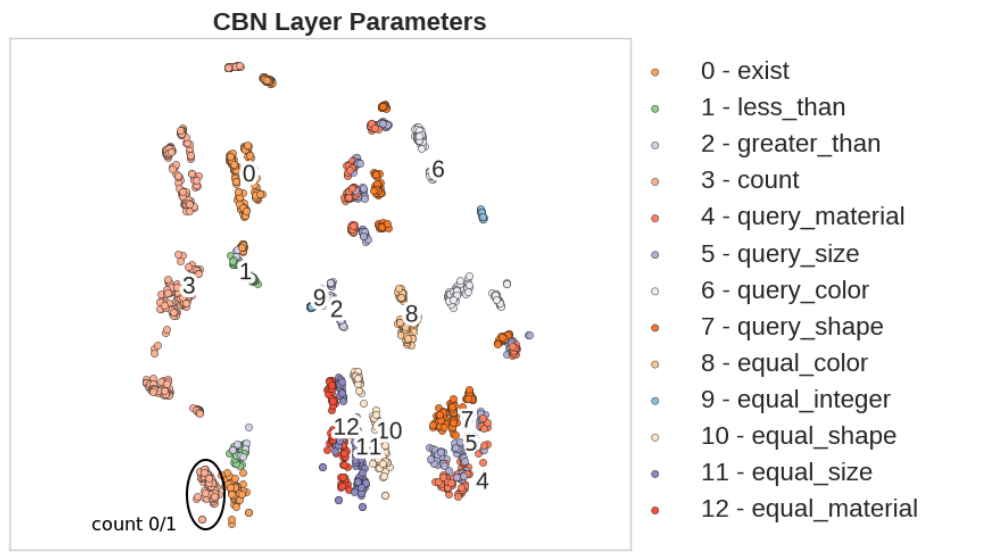
Description automatically generated

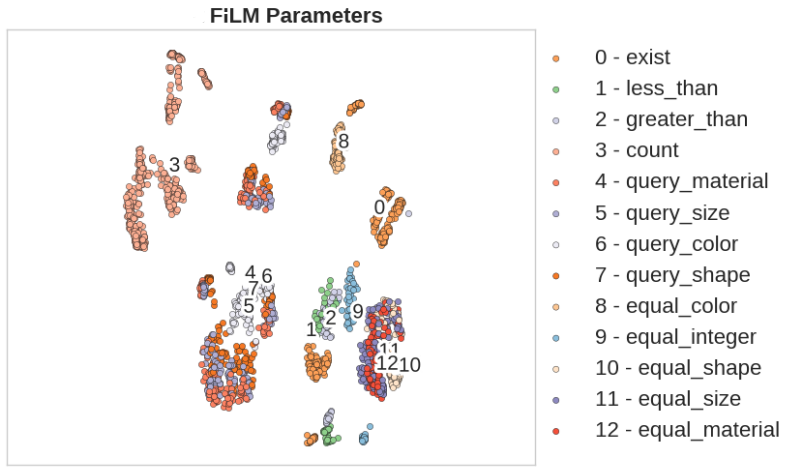
Figure 1: The linguistic pipeline (left), visual pipeline (middle),

and residual block architecture (right) of our model.

2.4 Theoretical Motivation

Both FiLM and CBN have comparable performances on the CLEVR dataset. However, each of them is slightly better than the other in different questions, for example FiLM has better accuracy with comparing questions but on the other hand, CBN is better for counting. We believe the accuracy difference is caused by the difference in spatial reasoning of CBN and FiLM.





From that, we assume that the combination of the two layers might result in a different spatial reasoning all together and as a result there will be an improvement in performances.

2.5 CLEVR dataset

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 (Johnson et al., 2017).

It is a generated dataset of 700K (image, question, answer, program) tuples. Images contain 3D-rendered objects of various shapes, materials, colors, and sizes. Questions are multi-step and compositional in nature, as shown in Figure 2. They range from counting questions (“How many green objects have the same size as the green metallic block?”) to comparison questions (“Are there fewer tiny yellow cylinders than yellow metal cubes?”) and can be 40+ words long. Answers are each

one word from a set of 28 possible answers (Perez et al., 2017).

Experiments

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

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.

References

De Vries, H., Strub, F., Mary, J., Larochelle, H., Pietquin, O., & Courville, A. (2017). Modulating early visual processing by language. *Advances in Neural Information Processing Systems*, *2017*-*Decem*, 6595–6605.

Johnson, J., Fei-Fei, L., Hariharan, B., Zitnick, C. L., Van Der Maaten, L., & Girshick, R. (2017). CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, *2017*-*Janua*, 1988–1997. https://doi.org/10.1109/CVPR.2017.215

Perez, E., de Vries, H., Strub, F., Dumoulin, V., & Courville, A. (2017). Learning Visual Reasoning Without Strong Priors. Retrieved from http://arxiv.org/abs/1707.03017

Perez, E., Strub, F., De Vries, H., Dumoulin, V., & Courville, A. (2018). FiLM: Visual reasoning with a general conditioning layer. *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, 3942–3951. Retrieved from https://arxiv.org/abs/1709.07871

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.

Examples from CLEVR are shown in Figure 1.

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,

Programs are an additional supervisory signal consisting of step-by-step instructions, such as filter shape[cube], relate[right],

and count, on how to answer the question. Program labels are difficult to generate or come by for real world datasets. Our model avoids using this extra supervision, learning to reason effectively directly from linguistic and visual input.