The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

FiLM: Visual Reasoning with a General Conditioning Layer

Ethan Perez,[1,2] **Florian Strub,**[4] **Harm de Vries,**[1] **Vincent Dumoulin,**[1] **Aaron Courville**[1,3]
[1]MILA, Université de Montréal, [2]Rice University, [3]CIFAR Fellow,
[4]Univ. Lille, CNRS, Centrale Lille, Inria, UMR 9189 CRIStAL France
ethanperez@rice.edu, florian.strub@inria.fr, mail@harmdevries.com,{dumouliv,courvila}@iro.umontreal.ca

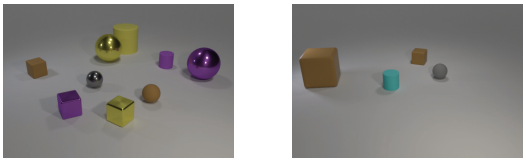
Abstract

We introduce a general-purpose conditioning method for neural networks called **FiLM**: **F**eature-wi**L**inear **M**odulation. FiLM layers influence neural network computation via a simple, feature-wise affine transformation based on conditioning information. We show that FiLM layers are highly effective for visual reasoning — answering image-related questions which require a multi-step, high-level process — a task which has proven difficult for standard deep learning methods that do not explicitly model reasoning. Specifically, we show on visual reasoning tasks that FiLM layers 1) halve state-of-the-art error for the CLEVR benchmark, 2) modulate features in a coherent manner, 3) are robust to ablations and architectural modifications, and 4) generalize well to challenging, new data from few examples or even zero-shot.

1 Introduction

The ability to reason about everyday visual input is a fundamental building block of human intelligence. 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, if a model made from general-purpose components could learn to visually reason, such an architecture would likely be more widely applicable across domains.

To understand if such a general-purpose architecture exists, we take advantage of the recently proposed CLEVR dataset (Johnson et al. 2017a) that tests visual reasoning via question answering. Examples from CLEVR are shown in Figure 1. Visual question answering, the general task of asking questions about images, has its own line of datasets (Malinowski and Fritz 2014; Geman et al. 2015; Antol et al. 2015) which generally focus on asking a diverse set of simpler questions on images, often answerable in a single glance. From these datasets, a number of 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). In particular, these methods tend to exploit biases in the data rather than capture complex underlying structure behind reasoning (Goyal et al. 2017).



(a) **Q:** *What number of cylinders are small purple things or yellow rubber things?* **A:** ***2***

(b) **Q:** *What color is the other object that is the same shape as the large brown matte thing?* **A:** ***Brown***

Figure 1: CLEVR examples and FiLM model answers.

In this work, we show that a general model architecture can achieve strong visual reasoning with a method we introduce as **FiLM**: **F**eature-wi**L**inear **M**odulation. 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, ranging from counting to comparing, for example. 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. We also introduce a novel FiLM-based zero-shot generalization method that further improves and validates FiLM's generalization capabilities.

2 Method

Our model processes the question-image input using FiLM, illustrated in Figure 2. We start by explaining FiLM and then describe our particular model for visual reasoning.

2.1 Feature-wise Linear Modulation

FiLM learns to adaptively influence the output of a neural network by applying an affine transformation, or FiLM, to the network's intermediate features, based on some input. More formally, FiLM learns functions f and h which output $\gamma_{i,c}$ and $\beta_{i,c}$ as a function of input $\boldsymbol{x_i}$:

$$\gamma_{i,c} = f_c(\boldsymbol{x}_i) \qquad \beta_{i,c} = h_c(\boldsymbol{x}_i), \tag{1}$$

where $\gamma_{i,c}$ and $\beta_{i,c}$ modulate a neural network's activations $\boldsymbol{F}_{i,c}$, whose subscripts refer to the i^{th} input's c^{th} feature or feature map, via a feature-wise affine transformation:

$$FiLM(\boldsymbol{F}_{i,c}|\gamma_{i,c}, \beta_{i,c}) = \gamma_{i,c}\boldsymbol{F}_{i,c} + \beta_{i,c}. \tag{2}$$

f and h can be arbitrary functions such as neural networks. Modulation of a target neural network's processing can be based on the same input to that neural network or some other input, as in the case of multi-modal or conditional tasks. For CNNs, f and h thus modulate the per-feature-map distribution of activations based on $\boldsymbol{x_i}$, agnostic to spatial location.

In practice, it is easier to refer to f and h as a single function that outputs one $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ vector, since, for example, it is often beneficial to share parameters across f and h for more efficient learning. We refer to this single function as the FiLM generator. We also refer to the network to which FiLM layers are applied as the Feature-wise Linearly Modulated network, the FiLM-ed network.

FiLM layers empower the FiLM generator to manipulate feature maps of a target, FiLM-ed network by scaling them up or down, negating them, shutting them off, selectively thresholding them (when followed by a ReLU), and more. Each feature map is conditioned independently, giving the FiLM generator moderately fine-grained control over activations at each FiLM layer.

As FiLM only requires two parameters per modulated feature map, it is a scalable and computationally efficient conditioning method. In particular, FiLM has a computational cost that does not scale with the image resolution.

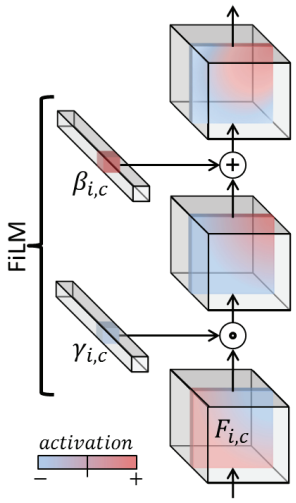


Figure 2: A single FiLM layer for a CNN. The dot signifies a Hadamard product. Various combinations of γ and β can modulate individual feature maps in a variety of ways.

2.2 Model

Our FiLM model consists of a FiLM-generating linguistic pipeline and a FiLM-ed visual pipeline as depicted in Figure 3. The FiLM generator processes a question $\boldsymbol{x_i}$ using a Gated Recurrent Unit (GRU) network (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 $(\boldsymbol{\gamma}^n_{i,\cdot}, \boldsymbol{\beta}^n_{i,\cdot})$ for each n^{th} residual block via affine projection.

The visual pipeline extracts 128 14×14 image feature maps from a resized, 224×224 image input using either a CNN trained from scratch or a fixed, pre-trained feature extractor with a learned layer of 3×3 convolutions. The CNN trained from scratch consists of 4 layers with 128 4×4 kernels each, ReLU activations, and batch normalization, similar to prior work on CLEVR (Santoro et al. 2017). The fixed feature extractor outputs the *conv4* layer of a ResNet-101 (He et al. 2016) pre-trained on ImageNet (Russakovsky et al. 2015) to match prior work on CLEVR (Johnson et al. 2017a; 2017b). Image features are processed by several — 4 for our model — FiLM-ed residual blocks (ResBlocks) with 128 feature maps and a final classifier. The classifier consists of a 1×1 convolution to 512 feature maps, global max-pooling, and a two-layer MLP with 1024 hidden units that outputs a softmax distribution over final answers.

Each FiLM-ed ResBlock starts with a 1×1 convolution followed by one 3×3 convolution with an architecture as depicted in Figure 3. We turn the parameters of batch normalization layers that immediately precede FiLM layers off. Drawing from prior work on CLEVR (Hu et al. 2017; Santoro et al. 2017) and visual reasoning (Watters et al. 2017), we concatenate two coordinate feature maps indicating relative x and y spatial position (scaled from -1 to 1) with the image features, each ResBlock's input, and the classifier's input to facilitate spatial reasoning.

We train our model end-to-end from scratch with

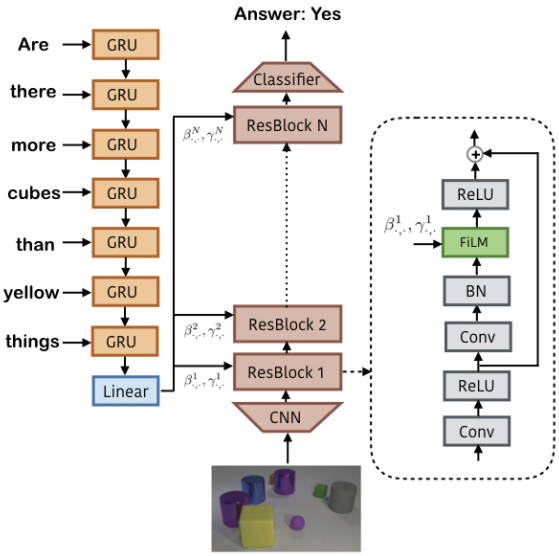


Figure 3: The FiLM generator (left), FiLM-ed network (middle), and residual block architecture (right) of our model.

Adam (Kingma and Ba 2015) (learning rate $3e^{-4}$), weight decay ($1e^{-5}$), batch size 64, and batch normalization and ReLU throughout FiLM-ed network. Our model uses only image-question-answer triplets from the training set without data augmentation. We employ early stopping based on validation accuracy, training for 80 epochs maximum. Empirically, we found FiLM had a large capacity, so many architectural and hyperparameter choices were for added regularization.

We stress that our model relies *solely* on feature-wise affine conditioning to use question information influence the visual pipeline behavior to answer questions. This approach differs from classical visual question answering pipelines which fuse image and language information into a single embedding via element-wise product, concatenation, attention, and/or more advanced methods (Yang et al. 2016; Lu et al. 2016; Anderson et al. 2017).

3 Related Work

FiLM can be viewed as a generalization of Conditional Normalization (CN) methods. CN replaces the parameters of the feature-wise affine transformation typical in normalization layers, as introduced originally (Ioffe and Szegedy 2015), with a learned function of some conditioning information. Various forms of CN have proven highly effective across a number of domains: Conditional Instance Norm (Dumoulin, Shlens, and Kudlur 2017; Ghiasi et al. 2017) and Adaptive Instance Norm (Huang and Belongie 2017) for image stylization, Dynamic Layer Norm for speech recognition (Kim, Song, and Bengio 2017), and Conditional Batch Norm for general visual question answering on complex scenes such as VQA and GuessWhat?! (de Vries et al. 2017). This work complements our own, as we seek to show that feature-wise affine conditioning is effective for multi-step reasoning and understand the underlying mechanism behind its success.

Notably, prior work in CN has not examined whether the affine transformation must be placed directly after normalization. Rather, prior work includes normalization in the method name for instructive purposes or due to implementation details. We investigate the connection between FiLM and normalization, finding it not strictly necessary for the affine transformation to occur directly after normalization. Thus, we provide a unified framework for all of these methods through FiLM, as well as a normalization-free relaxation of this approach which can be more broadly applied.

Beyond CN, there are many connections between FiLM and other conditioning methods. A common approach, used for example in Conditional DCGANs (Radford, Metz, and Chintala 2016), is to concatenate constant feature maps of conditioning information with convolutional layer input. Though not as parameter efficient, this method simply results in a feature-wise conditional bias. Likewise, concatenating conditioning information with fully-connected layer input amounts to a feature-wise conditional bias. Other approaches such as WaveNet (van den Oord et al. 2016a) and Conditional PixelCNN (van den Oord et al. 2016b) directly add a conditional feature-wise bias. These approaches are equivalent to FiLM with $\boldsymbol{\gamma} = \mathbf{1}$, which we compare FiLM to in the Experiments section. In reinforcement learning, an alternate formulation of FiLM has been used to train one game-conditioned deep Q-network to play ten Atari games (Kirkpatrick et al. 2017), though FiLM was neither the focus of this work nor analyzed as a major component.

Other methods gate an input's features as a function of that same input, rather than a separate conditioning input. These methods include LSTMs for sequence modeling (Hochreiter and Schmidhuber 1997), Convolutional Sequence to Sequence for machine translation (Gehring et al. 2017), and even the ImageNet 2017 winning model, Squeeze and Excitation Networks (Hu, Shen, and Sun 2017). This approach amounts to a feature-wise, conditional scaling, restricted to between 0 and 1, while FiLM consists of both scaling and shifting, each unrestricted. In the Experiments section, we show the effect of restricting FiLM's scaling to between 0 and 1 for visual reasoning. We find it noteworthy that this general approach of feature modulation is effective across a variety of settings and architectures.

There are even broader links between FiLM and other methods. For example, FiLM can be viewed as using one network to generate parameters of another network, making it a form of hypernetwork (Ha, Dai, and Le 2016). Also, FiLM has potential ties with conditional computation and mixture of experts methods, where specialized network subparts are active on a per-example basis (Jordan and Jacobs 1994; Eigen, Ranzato, and Sutskever 2014; Shazeer et al. 2017); we later provide evidence that FiLM learns to selectively highlight or suppress feature maps based on conditioning information. Those methods select at a sub-network level while FiLM selects at a feature map level.

In the domain of visual reasoning, one leading method is the Program Generator + Execution Engine model (Johnson et al. 2017b). This approach consists of a sequence-to-sequence Program Generator, which takes in a question

Model	**Overall**	Count	Exist	Compare Numbers	Query Attribute	Compare Attribute
Human (Johnson et al. 2017b)	92.6	86.7	96.6	86.5	95.0	96.0
Q-type baseline (Johnson et al. 2017b)	41.8	34.6	50.2	51.0	36.0	51.3
LSTM (Johnson et al. 2017b)	46.8	41.7	61.1	69.8	36.8	51.8
CNN+LSTM (Johnson et al. 2017b)	52.3	43.7	65.2	67.1	49.3	53.0
CNN+LSTM+SA (Santoro et al. 2017)	76.6	64.4	82.7	77.4	82.6	75.4
N2NMN* (Hu et al. 2017)	83.7	68.5	85.7	84.9	90.0	88.7
PG+EE (9K prog.)* (Johnson et al. 2017b)	88.6	79.7	89.7	79.1	92.6	96.0
PG+EE (700K prog.)* (Johnson et al. 2017b)	96.9	92.7	97.1	**98.7**	98.1	98.9
CNN+LSTM+RN†‡ (Santoro et al. 2017)	95.5	90.1	97.8	93.6	97.9	97.1
CNN+GRU+FiLM	**97.7**	**94.3**	99.1	96.8	99.1	99.1
CNN+GRU+FiLM‡	97.6	**94.3**	**99.3**	93.4	**99.3**	**99.3**

Table 1: CLEVR accuracy (overall and per-question-type) by baselines, competing methods, and FiLM. (*) denotes use of extra supervision via program labels. (†) denotes use of data augmentation. (‡) denotes training from raw pixels.

and outputs a sequence corresponding to a tree of composable neural modules, each of which is a two or three layer residual block. This tree of neural modules is assembled to form the Execution Engine that then predicts an answer from the image. This modular approach is part of a line of neural module network methods (Andreas et al. 2016a; 2016b; Hu et al. 2017), of which End-to-End Module Networks (Hu et al. 2017) have also been tested on visual reasoning. These models use strong priors by explicitly modeling the compositional nature of reasoning and by training with additional program labels, *i.e.* ground-truth step-by-step instructions on how to correctly answer a question. End-to-End Module Networks further build in model biases via per-module, hand-crafted neural architectures for specific functions. Our approach learns directly from visual and textual input without additional cues or a specialized architecture.

Relation Networks (RNs) are another leading approach for visual reasoning (Santoro et al. 2017). RNs succeed by explicitly building in a comparison-based prior. RNs use an MLP to carry out pairwise comparisons over each location of extracted convolutional features over an image, including LSTM-extracted question features as input to this MLP. RNs then element-wise sum over the resulting comparison vectors to form another vector from which a final classifier predicts the answer. We note that RNs have a computational cost that scales quadratically in spatial resolution, while FiLM's cost is independent of spatial resolution. Notably, since RNs concatenate question features with MLP input, a form of feature-wise conditional biasing as explained earlier, their conditioning approach is related to FiLM.

4 Experiments

First, we test our model on visual reasoning with the CLEVR task and use trained FiLM models to analyze what FiLM learns. Second, we explore how well our model generalizes to more challenging questions with the CLEVR-Humans task. Finally, we examine how FiLM performs in few-shot and zero-shot generalization settings using the CLEVR Compositional Generalization Test. Our code is available at https://github.com/ethanjperez/film.

4.1 CLEVR Task

CLEVR is a synthetic dataset of 700K (image, question, answer, program) tuples (Johnson et al. 2017a). Images contain 3D-rendered objects of various shapes, materials, colors, and sizes. Questions are multi-step and compositional in nature, as shown in Figure 1. 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. 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.

Baselines We compare against the following methods, discussed in detail in the Related Work section:

- **Q-type baseline**: Predicts based on a question's category.
- **LSTM**: Predicts using only the question.
- **CNN+LSTM**: MLP prediction over CNN-extracted image features and LSTM-extracted question features.
- **Stacked Attention Networks (CNN+LSTM+SA)**: Linear prediction over CNN-extracted image feature and LSTM-extracted question features combined via two rounds of soft spatial attention (Yang et al. 2016).
- **End-to-End Module Networks (N2NMN) and Program Generator + Execution Engine (PG+EE)**: Methods in which separate neural networks learn separate subfunctions and are assembled into a question-dependent structure (Hu et al. 2017; Johnson et al. 2017b).
- **Relation Networks (CNN+LSTM+RN)**: An approach which builds in pairwise comparisons over spatial locations to explicitly model reasoning's relational nature (Santoro et al. 2017).

Results FiLM achieves a new overall state-of-the-art on CLEVR, as shown in Table 1, outperforming humans and previous methods, including those using explicit models of reasoning, program supervision, and/or data augmentation.

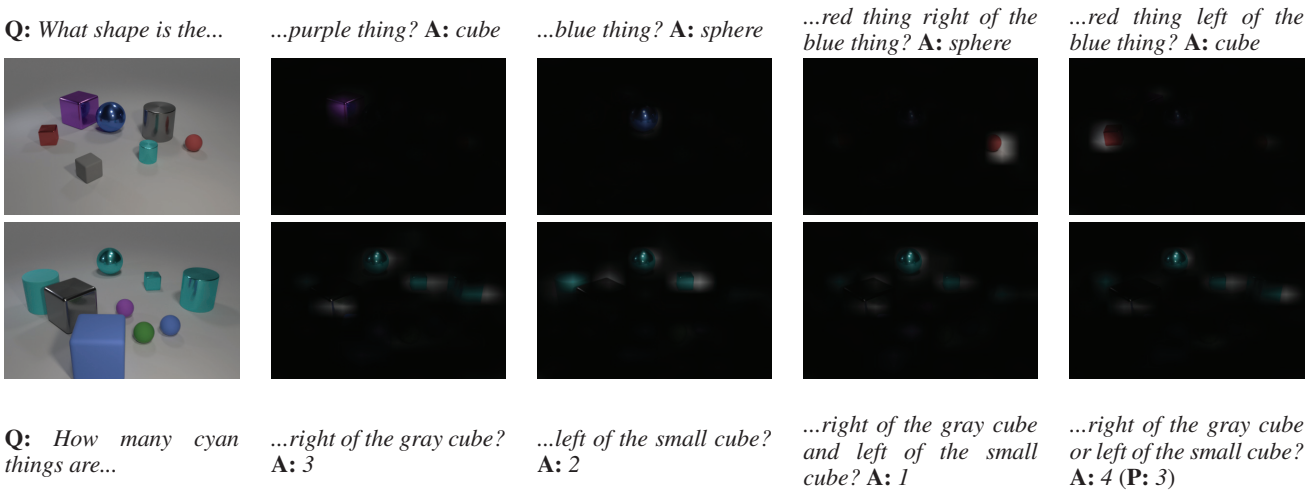


Figure 4: Visualizations of the distribution of locations which the model uses for its globally max-pooled features which its final MLP predicts from. FiLM correctly localizes the answer-referenced object (top) or all question-referenced objects (bottom), but not as accurately when it answers incorrectly (rightmost bottom). Questions and images used match (Johnson et al. 2017b).

For methods not using extra supervision, FiLM roughly halves state-of-the-art error (from 4.5% to 2.3%). Note that using pre-trained image features as input can be viewed as a form of data augmentation in itself but that FiLM performs equally well using raw pixel inputs. Interestingly, the raw pixel model seems to perform better on lower-level questions (*i.e.* querying and comparing attributes) while the image features model seems to perform better on higher-level questions (*i.e.* compare numbers of objects).

4.2 What Do FiLM Layers Learn?

To understand how FiLM visually reasons, we visualize activations to observe the net result of FiLM layers. We also use histograms and t-SNE (van der Maaten and Hinton 2008) to find patterns in the learned FiLM γ and β themselves.

Activation Visualizations Figure 4 visualizes the distribution of locations responsible for the globally-pooled features which the MLP in the model's final classifier uses to predict answers. These images reveal that the FiLM model predicts using features of areas near answer-related or question-related objects, as the high CLEVR accuracy also suggests. This finding highlights that appropriate feature modulation indirectly results in spatial modulation, as regions with question-relevant features will have large activations while other regions will not. This observation might explain why FiLM outperforms Stacked Attention, the next best method not explicitly built for reasoning, so significantly (21%); FiLM appears to carry many of spatial attention's benefits, while also influencing feature representation.

Figure 4 also suggests that the FiLM-ed network carries out reasoning throughout its pipeline. In the top example, the FiLM-ed network has localized the answer-referenced object alone before the MLP classifier. In the bottom example, the FiLM-ed network retains, for the MLP classifier, features on objects that are not referred to by the answer but are referred to by the question. The latter example provides evidence that the final MLP itself carries out some reasoning, using FiLM to extract relevant features for its reasoning.

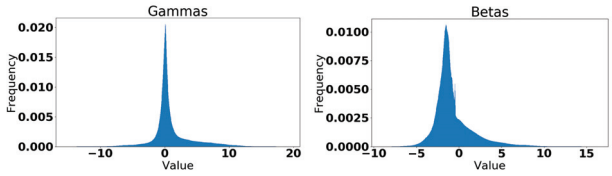


Figure 5: Histograms of $\gamma_{i,c}$ (left) and $\beta_{i,c}$ (right) values over all FiLM layers, calculated over the validation set.

FiLM Parameter Histograms To analyze at a lower level how FiLM uses the question to condition the visual pipeline, we plot γ and β values predicted over the validation set, as shown in Figure 5. γ and β values take advantage of a sizable range, varying from -15 to 19 and from -9 to 16, respectively. γ values show a sharp peak at 0, showing that FiLM learns to use the question to shut off or significantly suppress whole feature maps. Simultaneously, FiLM learns to upregulate a much more selective set of other feature maps with high magnitude γ values. Furthermore, a large fraction (36%) of γ values are negative; since our model uses a ReLU after FiLM, $\gamma < 0$ can cause a significantly different set of activations to pass the ReLU to downstream layers than $\gamma > 0$. Also, 76% of β values are negative, suggesting that FiLM also uses β to be selective about which activations pass the ReLU. We show later that FiLM's success is largely architecture-agnostic, but examining a particular model gives insight into the influence FiLM learns to exert in a specific case. Together, these findings suggest that FiLM learns to selectively upregulate, downregulate, and shut off feature maps based on conditioning information.

FiLM Parameters t-SNE Plot In Figure 6, we visualize FiLM parameter vectors $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ for 3,000 random valida-

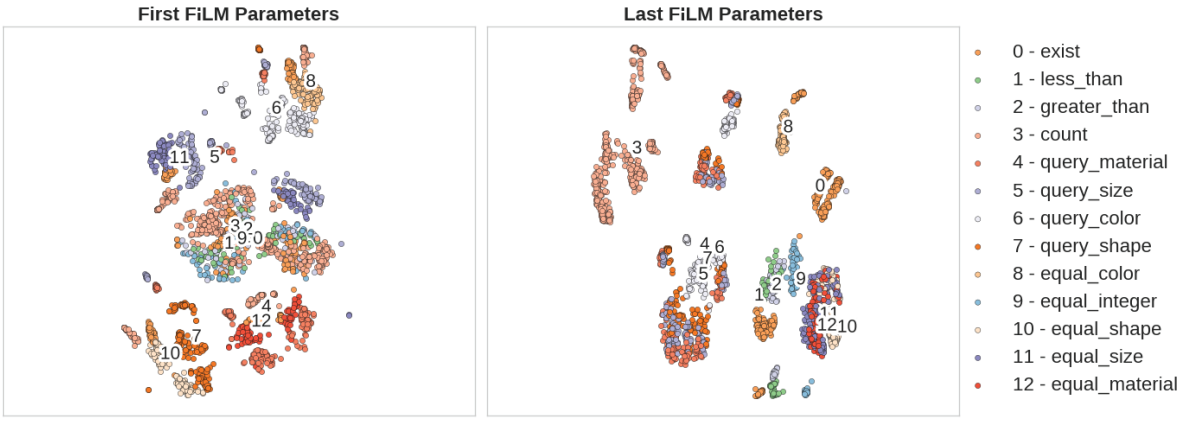


Figure 6: t-SNE plots of (γ, β) of the first (left) and last (right) FiLM layers of a 6-FiLM layer Network. FiLM parameters cluster by low-level reasoning functions in the first layer and by high-level reasoning functions in the last layer.

tion points with t-SNE. We analyze the deeper, 6-ResBlock version of our model, which has a similar validation accuracy as our 4-ResBlock model, to better examine how FiLM layers in different layers of a hierarchy behave. First and last layer FiLM $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ are grouped by the low-level and high-level reasoning functions necessary to answer CLEVR questions, respectively. For example, FiLM parameters for `equal_color` and `query_color` are close for the first layer but apart for the last layer. The same is true for shape, size and material questions. Conversely, `equal_shape`, `equal_size`, and `equal_material` FiLM parameters are grouped in the last layer but split in the first layer — likewise for other high level groupings such as integer comparison and querying. These findings suggest that FiLM layers learn a sort of function-based modularity without an architectural prior. Simply with end-to-end training, FiLM learns to handle not only different types of questions differently, but also different types of question sub-parts differently; the FiLM model works from low-level to high-level processes as is the proper approach. For models with fewer FiLM layers, such patterns also appear, but less clearly; these models must begin higher level reasoning sooner.

4.3 Ablations

Using the validation set, we conduct an ablation study on our best model to understand how FiLM learns visual reasoning. We show results for test time ablations in Figure 7, for architectural ablations in Table 2, and for varied model depths in Table 3. Without hyperparameter tuning, most architectural ablations and model depths outperform prior state-of-the-art on training from only image-question-answer triplets, supporting FiLM's overall robustness. Table 3 also shows using the validation set that our results are statistically significant.

Effect of γ and β To test the effect of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ separately, we trained one model with a constant $\boldsymbol{\gamma} = \mathbf{1}$ and another with $\boldsymbol{\beta} = \mathbf{0}$. With these models, we find a 1.5% and .5% accuracy drop, respectively; FiLM can learn to condition the CNN for visual reasoning through either biasing or scaling alone, albeit not as well as conditioning both together. This

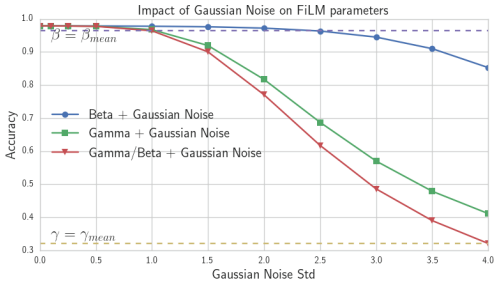


Figure 7: An analysis of how robust FiLM parameters are to noise at test time. The horizontal lines correspond to setting γ or β to their respective training set mean values.

result also suggests that $\boldsymbol{\gamma}$ is more important than $\boldsymbol{\beta}$.

To further compare the importance of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, we run a series of test time ablations (Figure 7) on our best, fully-trained model. First, we replace $\boldsymbol{\beta}$ with the mean $\boldsymbol{\beta}$ across the training set. This ablation in effect removes all conditioning information from $\boldsymbol{\beta}$ parameters during test time, from a model trained to use both $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$. Here, we find that accuracy only drops by 1.0%, while the same procedure on $\boldsymbol{\gamma}$ results in a 65.4% drop. This large difference suggests that, in practice, FiLM largely conditions through $\boldsymbol{\gamma}$ rather than $\boldsymbol{\beta}$. Next, we analyze performance as we add increasingly more Gaussian noise to the best model's FiLM parameters at test time. Noise in gamma hurts performance significantly more, showing FiLM's higher sensitivity to changes in $\boldsymbol{\gamma}$ than in $\boldsymbol{\beta}$ and corroborating the relatively greater importance of $\boldsymbol{\gamma}$.

Restricting γ To understand what aspect of $\boldsymbol{\gamma}$ is most effective, we train a model that limits $\boldsymbol{\gamma}$ to $(0, 1)$ using sigmoid, as many models which use feature-wise, multiplicative gating do. Likewise, we also limit $\boldsymbol{\gamma}$ to $(-1, 1)$ using $tanh$. Both restrictions hurt performance, roughly as much as removing conditioning from $\boldsymbol{\gamma}$ entirely by training with $\boldsymbol{\gamma} = \mathbf{1}$. Thus, FiLM's ability to scale features by large magnitudes appears to contribute to its success. Limiting $\boldsymbol{\gamma}$ to

Model	Overall
Restricted γ or β	
FiLM with $\beta := \mathbf{0}$	96.9
FiLM with $\gamma := \mathbf{1}$	95.9
FiLM with $\gamma := \sigma(\gamma)$	95.9
FiLM with $\gamma := tanh(\gamma)$	96.3
FiLM with $\gamma := exp(\gamma)$	96.3
Moving FiLM within ResBlock	
FiLM after residual connection	96.6
FiLM after ResBlock ReLU-2	97.7
FiLM after ResBlock Conv-2	97.1
FiLM before ResBlock Conv-1	95.0
Removing FiLM from ResBlocks	
No FiLM in ResBlock 4	96.8
No FiLM in ResBlock 3-4	96.5
No FiLM in ResBlock 2-4	97.3
No FiLM in ResBlock 1-4	21.4
Miscellaneous	
1×1 conv only, with no coord. maps	95.3
No residual connection	94.0
No batch normalization	93.7
Replace image features with raw pixels	97.6
Best Architecture	97.4±.4

Table 2: CLEVR val accuracy for ablations, trained with the best architecture with only specified changes. We report the standard deviation of the best model accuracy over 5 runs.

$(0, \infty)$ with exp also hurts performance, validating the value of FiLM's capacity to negate and zero out feature maps.

Conditional Normalization We perform an ablation study on the placement of FiLM to evaluate the relationship between normalization and FiLM that Conditional Normalization approaches assume. Unfortunately, it is difficult to accurately decouple the effect of FiLM from normalization by simply training our corresponding model without normalization, as normalization significantly accelerates, regularizes, and improves neural network learning (Ioffe and Szegedy 2015), but we include these results for completeness. However, we find no substantial performance drop when moving FiLM layers to different parts of our model's ResBlocks; we even reach the upper end of the best model's performance range when placing FiLM after the post-normalization ReLU in the ResBlocks. Thus, we decouple the name from normalization for clarity regarding where the fundamental effectiveness of the method comes from. By demonstrating this conditioning mechanism is not closely connected to normalization, we open the doors to applications other settings in which normalization is less common, such as RNNs and reinforcement learning, which are promising directions for future work with FiLM.

Repetitive Conditioning To understand the contribution of repetitive conditioning towards FiLM model success, we train FiLM models with successively fewer FiLM layers. Models with fewer FiLM layers, even a single FiLM layer, do not deviate far from the best model's performance, revealing that the model can reason and answer diverse questions successfully by modulating features even just once. This observation highlights the capacity of even one FiLM layer. Perhaps one FiLM layer can pass enough question information to the CNN to enable it to carry out reasoning later in the network, in place of the more hierarchical conditioning deeper FiLM models appear to use. We leave more in-depth investigation of this matter for future work.

Model	Overall	Model	Overall
1 ResBlock	93.5	6 ResBlocks	97.7
2 ResBlocks	97.1	7 ResBlocks	97.4
3 ResBlocks	96.7	8 ResBlocks	97.6
4 ResBlocks	97.4±.4	12 ResBlocks	96.9
5 ResBlocks	97.4		

Table 3: CLEVR val accuracy by FiLM model depth.

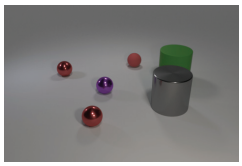
Spatial Reasoning To examine how FiLM models approach spatial reasoning, we train a version of our best model architecture, from image features, with only 1×1 convolutions and without feeding coordinate feature maps indicating relative spatial position to the model. Due to the global max-pooling near the end of the model, this model cannot transfer information across spatial positions. Notably, this model still achieves a high 95.3% accuracy, indicating that FiLM models are able to reason about space simply from the spatial information contained in a single location of fixed image features.

Residual Connection Removing the residual connection causes one of the larger accuracy drops. Since there is a global max-pooling operation near the end of the network, this finding suggests that the best model learns to primarily use features of locations that are repeatedly important throughout lower and higher levels of reasoning to make its final decision. The higher accuracies for models with FiLM modulating features inside residual connections rather than outside residual connections supports this hypothesis.

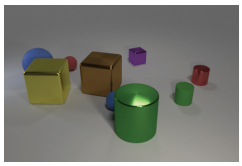
Model Depth Table 3 shows model performance by the number of ResBlocks. FiLM is robust to varying depth but less so with only 1 ResBlock, backing the earlier theory that the FiLM-ed network reasons throughout its pipeline.

4.4 CLEVR-Humans: Human-Posed Questions

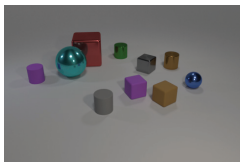
To assess how well visual reasoning models generalize to more realistic, complex, and free-form questions, the CLEVR-Humans dataset was introduced (Johnson et al. 2017b). This dataset contains human-posed questions on CLEVR images along with their corresponding answers. The number of samples is limited — 18K for training, 7K for validation, and 7K for testing. The questions were collected from Amazon Mechanical Turk workers prompted to ask questions that were likely *hard for a smart robot to answer*. As a result, CLEVR-Humans questions use more diverse vocabulary and complex concepts.



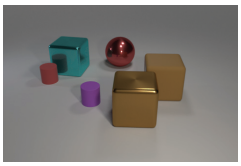
Q: *What object is the color of grass?* **A:** *Cylinder*



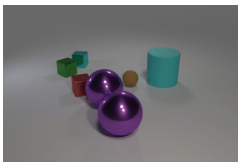
Q: *Which shape objects are partially obscured from view?* **A:** *Sphere*



Q: *What color is the matte object farthest to the right?* **A:** *Brown*



Q: *What shape is reflecting in the large cube?* **A:** *Cylinder*



Q: *If all cubical objects were removed what shaped objects would there be the most of?* **A:** *Sphere* (**P:** *Rubber*)

Figure 8: Examples from CLEVR-Humans, which introduces new words (underlined) and concepts. After fine-tuning on CLEVR-Humans, a CLEVR-trained model can now reason about obstruction, superlatives, and reflections but still struggles with hypothetical scenarios (rightmost). It also has learned human preference to primarily identify objects by shape (leftmost).

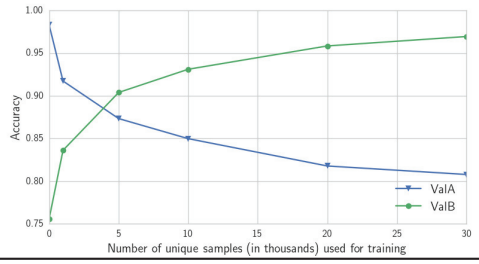
Model	Train CLEVR	Train CLEVR, fine-tune human
LSTM	27.5	36.5
CNN+LSTM	37.7	43.2
CNN+LSTM+SA+MLP	50.4	57.6
PG+EE (18K prog.)	54.0	66.6
CNN+GRU+FiLM	**56.6**	**75.9**

Table 4: CLEVR-Humans test accuracy, before (left) and after (right) fine-tuning on CLEVR-Humans data

Method To test FiLM on CLEVR-Humans, we take our best CLEVR-trained FiLM model and fine-tune its FiLM-generating linguistic pipeline alone on CLEVR-Humans. Similar to prior work (Johnson et al. 2017b), we do not update the visual pipeline on CLEVR-Humans to mitigate overfitting to the small training set.

Results Our model achieves state-of-the-art generalization to CLEVR-Humans, both before and after fine-tuning, as shown in Table 4, indicating that FiLM is well-suited to handle more complex and diverse questions. Figure 8 shows examples from CLEVR-Humans with FiLM model answers. Before fine-tuning, FiLM outperforms prior methods by a smaller margin. After fine-tuning, FiLM reaches a considerably improved final accuracy. In particular, the *gain* in accuracy made by FiLM upon fine-tuning is more than 50% greater than those made by other models; FiLM adapts data-efficiently using the small CLEVR-Humans dataset.

Notably, FiLM surpasses the prior state-of-the-art method, Program Generator + Execution Engine (PG+EE), after fine-tuning by 9.3%. Prior work on PG+EEs explains that this neural module network method struggles on questions which cannot be well approximated with the model's module inventory (Johnson et al. 2017b). In contrast, FiLM has the freedom to modulate existing feature maps, a fairly flexible and fine-grained operation, in novel ways to reason about new concepts. These results thus provide some evidence for the benefits of FiLM's general nature.



Method	Train A		Fine-tune B	
	A	B	A	B
CNN+LSTM+SA	80.3	68.7	75.7	75.8
PG+EE (18K prog.)	96.6	73.7	76.1	92.7
CNN+GRU+FiLM	**98.3**	75.6	80.8	**96.9**
CNN+GRU+FiLM 0-Shot	**98.3**	**78.8**	**81.1**	**96.9**

Figure 9: CoGenT results. FiLM ValB accuracy reported on ValB without the 30K fine-tuning samples (Figure). Accuracy before and after fine-tuning on 30K of ValB (Table).

4.5 CLEVR Compositional Generalization Test

To test how well models learn compositional concepts that generalize, CLEVR-CoGenT was introduced (Johnson et al. 2017a). This dataset is synthesized in the same way as CLEVR but contains two conditions: in Condition A, all cubes are gray, blue, brown, or yellow and all cylinders are red, green, purple, or cyan; in Condition B, cubes and cylinders swap color palettes. Both conditions contain spheres of all colors. CLEVR-CoGenT thus indicates how a model answers CLEVR questions: by memorizing combinations of traits or by learning disentangled or general representations.

Results We train our best model architecture on Condition A and report accuracies on Conditions A and B, before and after fine-tuning on B, in Figure 9. Our results indicate FiLM surpasses other visual reasoning models at learning general concepts. FiLM learns better compositional generalization even than PG+EE, which explicitly models compositionality and is trained with program-level supervision that specifically includes filtering colors and filtering shapes.

Sample Efficiency and Catastrophic Forgetting We show sample efficiency and forgetting curves in Figure 9. FiLM achieves prior state-of-the-art accuracy with $1/3$ as much fine-tuning data. However, our FiLM model still suffers from catastrophic forgetting after fine-tuning.

Zero-Shot Generalization FiLM's accuracy on Condition A is much higher than on B, suggesting FiLM has memorized attribute combinations to an extent. For example, the model learns a bias that cubes are not cyan, as learning this training set bias helps minimize training loss.

To overcome this bias, we develop a novel FiLM-based zero-shot generalization method. Inspired by word embedding manipulations, *e.g. "King" - "Man" + "Woman" = "Queen"* (Mikolov et al. 2013), we test if linear manipulation extends to reasoning with FiLM. We compute $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ for *"How many cyan cubes are there?"* via the linear combination of questions in the FiLM parameter space: *"How many cyan spheres are there?"* + *"How many brown cubes are there?"* − *"How many brown spheres are there?"*. With this $(\boldsymbol{\gamma}, \boldsymbol{\beta})$, our model can correctly count cyan cubes. We show another example of this method in Figure 10.

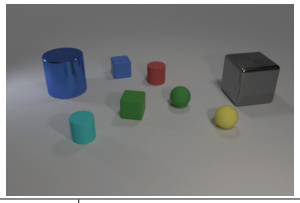
We evaluate this method on validation B, using a parser to automatically generate the right combination of questions. We test previously reported CLEVR-CoGenT FiLM models with this method and show results in Figure 9. With this method, there is a 3.2% overall accuracy gain when training on A and testing for zero-shot generalization on B. Yet this method could only be applied to $1/3$ of questions in B. For these questions, model accuracy starts at 71.5% and jumps to 80.7%. Before fine-tuning on B, the accuracy between zero-shot and original approaches on A is identical, likewise for B after fine-tuning. We note that difference in the predicted FiLM parameters between these two methods is negligible, likely causing the similar performance.

We achieve these improvements without specifically training our model for zero-shot generalization. Our method simply allows FiLM to take advantage of any concept disentanglement in the CNN after training. We also observe that convex combinations of the FiLM parameters – *i.e.* between *"How many cyan things are there?"* and *"How many brown things are there?"* – often monotonically interpolates the predicted answer between the answers to endpoint questions. These results highlight, to a limited extent, the flexibility of FiLM parameters for meaningful manipulations.

As implemented, this method has many limitations. However, approaches from word embeddings, representation learning, and zero-shot learning can be applied to directly optimize $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ for analogy-making (Bordes et al. 2013; Guu, Miller, and Liang 2015; Oh et al. 2017). The FiLM-ed network could directly train with this procedure via backpropagation. A learned model could also replace the parser. We find such avenues promising for future work.

5 Conclusion

We show that a model can achieve strong visual reasoning using general-purpose Feature-wise Linear Modulation layers. By efficiently manipulating a neural network's intermediate features in a selective and meaningful manner using FiLM layers, a RNN can effectively use language to modulate a CNN to carry out diverse and multi-step reasoning tasks over an image. Our ablation study suggests that FiLM is resilient to architectural modifications, test time ablations, and even restrictions on FiLM layers themselves. Notably, we provide evidence that FiLM's success is not closely connected with normalization as previously assumed. Thus, we open the door for applications of this approach to settings where normalization is less common, such as RNNs and reinforcement learning. Our findings also suggest that FiLM models can generalize better, more sample efficiently, and even zero-shot to foreign or more challenging data. Overall, the results of our investigation of FiLM in the case of visual reasoning complement broader literature that demonstrates the success of FiLM-like techniques across many domains, supporting the case for FiLM's strength not simply within a single domain but as a general, versatile approach.



Question	What is the blue big cylinder made of?
(1) Swap shape	What is the blue big **sphere** made of?
(2) Swap color	What is the **green** big cylinder made of?
(3) Swap shape/color	What is the **green** big **sphere** made of?

Figure 10: A CLEVR-CoGenT example. The combination of concepts *"blue"* and *"cylinder"* is not in the training set. Our zero-shot method computes the original question's FiLM parameters via linear combination of three other questions' FiLM parameters: (1) + (2) - (3). This method corrects our model's answer from *"rubber"* to *"metal"*.

6 Acknowledgements

We thank the developers of PyTorch (pytorch.org) and (Johnson et al. 2017b) for open-source code which our implementation was based off. We thank Mohammad Pezeshki, Dzmitry Bahdanau, Yoshua Bengio, Nando de Freitas, Hugo Larochelle, Laurens van der Maaten, Joseph Cohen, Joelle Pineau, Olivier Pietquin, Jérémie Mary, César Laurent, Chin-Wei Huang, Layla Asri, Max Smith, and James Ough for helpful discussions and Justin Johnson for CLEVR test evaluations. We thank NVIDIA for donating a DGX-1 computer used in this work. We also acknowledge FRQNT through the CHIST-ERA IGLU project and CPER Nord-Pas de Calais/FEDER DATA Advanced data science and technologies 2015-2020 for funding our work. Lastly, we thank acronymcreator.net for the acronym FiLM.

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2017. Bottom-up and top-down attention for image captioning and vqa. In *VQA Workshop at CVPR*.

Andreas, J.; Marcus, R.; Darrell, T.; and Klein, D. 2016a. Learn-

ing to compose neural networks for question answering. In *NAACL*.

Andreas, J.; Rohrbach, M.; Darrell, T.; and Klein, D. 2016b. Neural module networks. In *CVPR*.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. VQA: Visual Question Answering. In *ICCV*.

Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; and Yakhnenko, O. 2013. Translating embeddings for modeling multi-relational data. In Burges, C. J. C.; Bottou, L.; Welling, M.; Ghahramani, Z.; and Weinberger, K. Q., eds., *NIPS*. Curran Associates, Inc. 2787–2795.

Chung, J.; Gülçehre, Ç.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *Deep Learning Workshop at NIPS*.

de Vries, H.; Strub, F.; Mary, J.; Larochelle, H.; Pietquin, O.; and Courville, A. C. 2017. Modulating early visual processing by language. In *NIPS*.

Dumoulin, V.; Shlens, J.; and Kudlur, M. 2017. A learned representation for artistic style. In *ICLR*.

Eigen, D.; Ranzato, M.; and Sutskever, I. 2014. Learning factored representations in a deep mixture of experts. In *ICLR Workshops*.

Gehring, J.; Auli, M.; Grangier, D.; Yarats, D.; and Dauphin, Y. N. 2017. Convolutional sequence to sequence learning. In *ICML*.

Geman, D.; Geman, S.; Hallonquist, N.; and Younes, L. 2015. Visual turing test for computer vision systems. volume 112, 3618–3623. National Acad Sciences.

Ghiasi, G.; Lee, H.; Kudlur, M.; Dumoulin, V.; and Shlens, J. 2017. Exploring the structure of a real-time, arbitrary neural artistic stylization network. *CoRR abs/1705.06830*.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *CVPR*.

Guu, K.; Miller, J.; and Liang, P. 2015. Traversing knowledge graphs in vector space. In *EMNLP*.

Ha, D.; Dai, A.; and Le, Q. 2016. Hypernetworks. In *ICLR*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural Comput.* 9(8):1735–1780.

Hu, R.; Andreas, J.; Rohrbach, M.; Darrell, T.; and Saenko, K. 2017. Learning to reason: End-to-end module networks for visual question answering. In *ICCV*.

Hu, J.; Shen, L.; and Sun, G. 2017. Squeeze-and-Excitation Networks. In *ILSVRC 2017 Workshop at CVPR*.

Huang, X., and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*.

Ioffe, S., and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*.

Johnson, J.; Hariharan, B.; van der Maaten, L.; Fei-Fei, L.; Zitnick, C. L.; and Girshick, R. B. 2017a. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*.

Johnson, J.; Hariharan, B.; van der Maaten, L.; Hoffman, J.; Li, F.; Zitnick, C. L.; and Girshick, R. B. 2017b. Inferring and executing programs for visual reasoning. In *ICCV*.

Jordan, M. I., and Jacobs, R. A. 1994. Hierarchical mixtures of experts and the em algorithm. *Neural Comput.* 6(2):181–214.

Kim, T.; Song, I.; and Bengio, Y. 2017. Dynamic layer normalization for adaptive neural acoustic modeling in speech recognition. In *InterSpeech*.

Kingma, D. P., and Ba, J. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; Hassabis, D.; Clopath, C.; Kumaran, D.; and Hadsell, R. 2017. Overcoming catastrophic forgetting in neural networks. *National Academy of Sciences* 114(13):3521–3526.

Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2016. Hierarchical question-image co-attention for visual question answering. In *NIPS*.

Malinowski, M., and Fritz, M. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. In *NIPS*.

Malinowski, M.; Rohrbach, M.; and Fritz, M. 2015. Ask your neurons: A neural-based approach to answering questions about images. In *ICCV*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*.

Oh, J.; Singh, S.; Lee, H.; and Kholi, P. 2017. Zero-shot task generalization with multi-task deep reinforcement learning. In *ICML*.

Perez, E.; de Vries, H.; Strub, F.; Dumoulin, V.; and Courville, A. C. 2017. Learning visual reasoning without strong priors. In *MLSLP Workshop at ICML*.

Radford, A.; Metz, L.; and Chintala, S. 2016. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. S.; Berg, A. C.; and Li, F. 2015. Imagenet large scale visual recognition challenge. *IJCV* 115(3):211–252.

Santoro, A.; Raposo, D.; Barrett, D. G.; Malinowski, M.; Pascanu, R.; Battaglia, P.; and Lillicrap, T. 2017. A simple neural network module for relational reasoning. *CoRR abs/1706.01427*.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *ICLR*.

van den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; and Kavukcuoglu, K. 2016a. Wavenet: A generative model for raw audio. *CoRR abs/1609.03499*.

van den Oord, A.; Kalchbrenner, N.; Espeholt, L.; Vinyals, O.; Graves, A.; and Kavukcuoglu, K. 2016b. Conditional image generation with pixelcnn decoders. In *NIPS*.

van der Maaten, L., and Hinton, G. 2008. Visualizing data using t-sne. *JMLR* 9(Nov):2579–2605.

Watters, N.; Tacchetti, A.; Weber, T.; Pascanu, R.; Battaglia, P.; and Zoran, D. 2017. Visual interaction networks. *CoRR abs/1706.01433*.

Yang, Z.; He, X.; Gao, J.; Deng, L.; and Smola, A. J. 2016. Stacked attention networks for image question answering. In *CVPR*.