

---

# Learning by Abstraction: The Neural State Machine

---

**Drew A. Hudson**

Stanford University

353 Serra Mall, Stanford, CA 94305  
dorarad@cs.stanford.edu

**Christopher D. Manning**

Stanford University

353 Serra Mall, Stanford, CA 94305  
manning@cs.stanford.edu

## Abstract

We introduce the Neural State Machine, seeking to bridge the gap between the neural and symbolic views of AI and integrate their complementary strengths for the task of visual reasoning. Given an image, we first predict a probabilistic graph that represents its underlying semantics and serves as a structured world model. Then, we perform sequential reasoning over the graph, iteratively traversing its nodes to answer a given question or draw a new inference. In contrast to most neural architectures that are designed to closely interact with the raw sensory data, our model operates instead in an abstract latent space, by transforming both the visual and linguistic modalities into semantic concept-based representations, thereby achieving enhanced transparency and modularity. We evaluate our model on VQA-CP and GQA, two recent VQA datasets that involve compositionality, multi-step inference and diverse reasoning skills, achieving state-of-the-art results in both cases. We provide further experiments that illustrate the model’s strong generalization capacity across multiple dimensions, including novel compositions of concepts, changes in the answer distribution, and unseen linguistic structures, demonstrating the qualities and efficacy of our approach.

## 1 Introduction

Language is one of the most marvelous feats of the human mind. The emergence of a compositional system of symbols that can distill and convey from rich sensory experiences to creative new ideas has been a major turning point in the evolution of intelligence, and made a profound impact on the nature of human cognition [19, 78, 13]. According to Jerry Fodor’s Language of Thought hypothesis [22, 72], thinking itself posses a language-like compositional structure, where elementary concepts combine in systematic ways to create compound new ideas or thoughts, allowing us to make “infinite use of finite means” [18] and fostering human’s remarkable capacities of abstraction and generalization [51].

Indeed, humans are particularly adept at making abstractions of various kinds: We make analogies and form **concepts** to generalize from given instances to unseen examples [70]; we see things in context, and build compositional **world models** to represent objects and understand their interactions and subtle relations, turning raw sensory signals into high-level semantic knowledge [64]; and we deductively draw inferences via conceptual rules and statements to proceed from known facts to novel conclusions [32, 40]. Not only are humans capable of learning, but we are also talented at **reasoning**.

Ideas about compositionality, abstraction and reasoning greatly inspired the classical views of artificial intelligence [74, 65], but have lately been overshadowed by the astounding success of deep learning over a wide spectrum of real-world tasks [33, 63, 82]. Yet, even though neural networks are undoubtedly powerful, flexible and robust, recent work has repeatedly demonstrated their flaws, showing how they struggle to generalize in a systematic manner [50], overly adhere to superficial and potentially misleading statistical associations instead of learning true causal relations [1, 42], strongly depend on large amounts of data and supervision [25, 51], and sometimes behave in surprising and

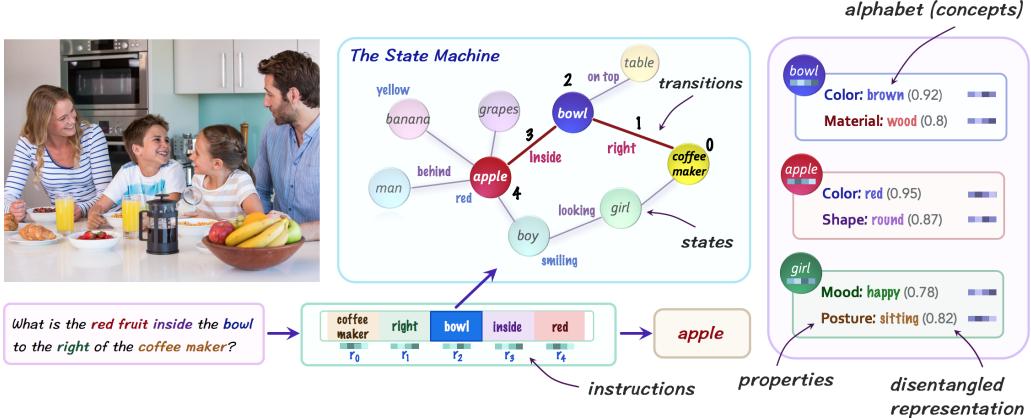


Figure 1: The Neural State Machine is a graph network that simulates the computation of an automaton. For the task of VQA, the model constructs a probabilistic scene graph to capture the semantics of a given image, which it then treats as a state machine, traversing its states as guided by the question to perform sequential reasoning.

worrisome ways [26, 20]. The sheer size and statistical nature of these models that support robustness and versatility are also what hinder their interpretability, modularity, and soundness.

Motivated to alleviate these deficiencies and bring the neural and symbolic approaches more closely together, we propose the Neural State Machine, a differentiable graph-based model that simulates the operation of an automaton, and explore it in the domain of visual reasoning and compositional question answering. Essentially, we proceed through two stages: *modeling* and *inference*. Starting from an image, we first generate a probabilistic scene graph [43, 49] that captures its underlying semantic knowledge in a compact form. Nodes correspond to objects and consist of structured representations of their properties, and edges depict both their spatial and semantic relations. Once we have the graph, we then treat it as a *state machine* and simulate an iterative computation over it, aiming to answer questions or draw inferences. We translate a given natural language question into a series of soft instructions, and feed them one-at-a-time into the machine to perform sequential reasoning, using attention to traverse its states and compute the answer.

Drawing inspiration from Bengio’s consciousness prior [12], we further define a set of semantic embedded *concepts* that describe different entities and aspects of the domain, such as various kinds of objects, attributes and relations. These concepts are used as the vocabulary that underlies both the scene graphs derived from the image as well as the reasoning instructions obtained from the question, effectively allowing both modalities to “speak the same language”. Whereas neural networks typically interact directly with raw observations and dense features, our approach encourages the model to reason instead in a semantic and factorized abstract space, which enables the disentanglement of structure from content and improves its modularity.

We demonstrate the value and performance of the Neural State Machine on two recent Visual Question Answering (VQA) datasets: GQA [41] which focuses on real-world visual reasoning and multi-step question answering, as well as VQA-CP [3], a recent split of the popular VQA dataset [2, 27] that has been designed particularly to evaluate generalization. We achieve state-of-the-art results on both tasks under single-model settings, substantiating the robustness and efficiency of our approach in answering challenging compositional questions. We then construct new splits leveraging the associated structured representations provided by GQA and conduct further experiments that provide significant evidence for the model’s strong generalization skills across multiple dimensions, such as novel compositions of concepts and unseen linguistic structures, validating its versatility under changing conditions.

Our model ties together two important qualities: abstraction and compositionality, with the respective key innovations of representing meaning as a structured attention distribution over an internal vocabulary of disentangled concepts, and capturing sequential reasoning as the iterative computation of a differentiable state machine over a semantic graph. We hope that creating such neural form of a classical model of computation will encourage and support the integration of the connectionist and symbolic methodologies in AI, opening the door to enhanced modularity, versatility, and generalization.



- |   |  |   |   |
|---|--|---|---|
| 1) What is the <b>giraffe</b> looking at?<br><b>person</b> ✓<br>2) Is the <b>fence</b> in front of the <b>giraffe</b> made of metal? <b>no</b> ✓<br>3) Is the <b>woman's shirt</b> blue or yellow? <b>blue</b> ✓<br>4) On which side of the image is the <b>person</b> ? <b>right</b> ✓<br>5) Is there a <b>child</b> behind the <b>giraffe</b> ? <b>no</b> ✗ | 1) What is the <b>fruit</b> to the right of the <b>salad</b> ? <b>strawberries</b> ✓<br>2) Is the <b>fork</b> to the right of the <b>salad</b> ? <b>no</b> ✓<br>3) Is the <b>plate</b> white and square? <b>no</b> ✓<br>4) Is the <b>cup</b> behind the round <b>plate</b> ? <b>yes</b> ✓<br>5) What is the <b>plate</b> made of? <b>paper</b> ✗ | 1) Are there either <b>scarves</b> or <b>hats</b> that are not pink? <b>no</b> ✓<br>2) Do the <b>bear's dress</b> and the <b>person's shirt</b> have the same color? <b>yes</b> ✓<br>3) Is the <b>bear</b> sitting or standing? <b>sitting</b> ✓<br>4) What is the green <b>object</b> that the <b>bear</b> is sitting on? <b>book</b> ✓<br>5) Is the <b>bear</b> wearing white <b>shoes</b> ? <b>yes</b> ✗ | 1) Are there either a <b>chair</b> or a <b>clock</b> in the image? <b>no</b> ✓<br>2) Are there any <b>flowers</b> behind the <b>bed</b> on the left of the <b>room</b> ? <b>yes</b> ✓<br>3) What color is the <b>appliance</b> on the right? <b>black</b> ✓<br>4) Is the <b>carpet</b> brown or blue? <b>brown</b> ✓<br>5) Is the <b>TV</b> turned on? <b>yes</b> ✗ |
|---|--|---|---|

Figure 2: Question examples along with answers predicted by the NSM. The questions involve diverse reasoning skills such as multi-step inference, relational and spatial reasoning, logic and comparisons.

## 2 Related work

Our model connects to multiple lines of research, including works about compositionality [14, 38], concept acquisition [36, 81], and neural computation [28, 62, 7]. Several works have explored the discovery and use of visual concepts in the contexts of reinforcement or unsupervised learning [35, 17] as well as in classical computer vision [21, 77]. Others have argued for the importance of incorporating strong inductive biases into neural architectures [14, 10, 6, 8], and indeed, there is a growing body of research that seeks to introduce different forms of structural priors inspired by computer architectures [29, 80, 40] or theory of computation [4, 23], aiming to bridge the gap between the symbolic and neural paradigms.

We explore our model in the context of VQA, a challenging multimodal task that has gained substantial attention in recent years [27, 79, 40]. Prior work commonly relied on dense visual features produced by either CNNs [83, 86] or object detectors [5], with a few recent models that use the relationships among objects to augment those features with contextual information from each object’s surroundings [52, 75, 66]. We move further in this direction, performing iterative reasoning over inferred scene graphs, and in contrast to prior models, incorporate higher-level semantic concepts to represent both the visual and linguistic modalities in a shared and sparser manner that facilitates their interaction.

Closest to the NSM is a model called MAC [40] we have developed in prior work, a recurrent network that applies attention-based operations to perform sequential reasoning. In fact, our new model follows the high-level structure proposed by MAC, where the new decoder (section 3.3) is analogous to the control unit, and the model simulation (section 3.4) is parallel to the read-write operation. At the same time, the Neural State Machine differs from MAC in two crucial respects: First, we reason over graph structures rather than directly over spatial maps of visual features, traversing the graph by successively shifting attention across its nodes and edges. Second, key to our model is the notion of semantic concepts that are used to express knowledge about both the visual and linguistic modalities, instead of working directly with the raw observational features. As our findings suggest, these ideas contribute significantly to the model’s performance compared to MAC and enhance its compositionality, transparency and generalization skills.

## 3 The Neural State Machine

The Neural State Machine is a graph-based network that simulates the computation of a finite automaton [37], and is explored here in the context of VQA, where we are given an image and a question and asked to provide an answer. We go through two stages – modeling and inference, the first to construct the state machine, and the second to simulate its operation.

In the modeling stage, we transform both the visual and linguistic modalities into abstract representations. The image is decomposed into a probabilistic *graph* that represents its semantics – the objects, attributes and relations in the depicted visual scene (section 3.2), while the question is converted into a sequence of reasoning *instructions* (section 3.3) that have to be performed in order to answer it.

In the inference stage (section 3.4), we treat the graph as a state machine, where the nodes, the objects within the image, correspond to states, and the edges, the relations between the objects, correspond to transitions. We then simulate a serial computation by iteratively feeding the machine with the instructions derived from the question and traversing its states, which allows us to perform sequential reasoning over the semantic visual scene, as guided by the question, to arrive at the answer.

We begin with a formal definition of the machine. In simple terms, a state machine is a computational model that consists of a collection of states, which it iteratively traverses while reading a sequence of inputs, as determined by a transition function. In contrast to the classical deterministic versions, the neural state machine defines an initial distribution over the states, and then performs a fixed number of computation steps  $N$ , recurrently updating the state distribution until completion. Formally, we define the neural state machine as a tuple  $(C, S, E, \{r_i\}_{i=0}^N, p_0, \delta)$ :

- $C$  the model’s alphabet, consisting of a set of concepts, embedded as learned vectors.
- $S$  a collection of states.
- $E$  a collection of directed edges that specify valid transitions between the states.
- $r_i$  a sequence of instructions, each of dimension  $d$ , that are passed in turn as an input to the transition function  $\delta$ .
- $p_0 : S \rightarrow [0, 1]$  a probability distribution of the initial state.
- $\delta_{S,E} : p_i \times r_i \rightarrow p_{i+1}$  a state transition function: a neural module that at each step  $i$  considers the distribution  $p_i$  over the states as well as an input instruction  $r_i$ , and uses it to redistribute the probability along the edges, yielding an updated state distribution  $p_{i+1}$ .

### 3.1 Concept vocabulary

In contrast to many common networks, the neural state machine operates over a discrete set of concepts. We create an embedded concept vocabulary  $C$  for the machine (initialized with GloVe [68]), that will be used to capture and represent the semantic content of input images. The vocabulary is grouped into  $L + 2$  properties such as object identity  $C_O = C_0$  (e.g. *cat*, *shirt*), different types of attributes  $C_A = \bigcup_{i=1}^L C_i$  (e.g. *colors*, *materials*) and relations  $C_R = C_{L+1}$  (e.g. *holding*, *behind*), all derived from the Visual Genome dataset [49] (see section 7.3 for details). We similarly define a set of embeddings  $D$  for each of the property types (such as “color” or “shape”).

In using the notion of concepts, we draw a lot of inspiration from humans, who are known for their ability to learn concepts and use them for tasks that involve abstract thinking and reasoning [11, 9, 30, 67]. In the following sections, rather than using raw and dense sensory input features directly, we represent both the visual and linguistic inputs in terms of our vocabulary, finding the most relevant concepts that they relate to. By associating such semantic concepts with raw sensory information from both the image and the question, we are able to derive higher-level representations that abstract away from irrelevant raw fine-grained statistics tied to each modality, and instead capture only the semantic knowledge necessary for the task. That way we can effectively cast both modalities onto the same space to facilitate their interaction, and, as discussed in section 4, improve the model’s compositionality, robustness and generalization skills.

### 3.2 States and edge transitions

In order to create the state machine, we construct a probabilistic scene graph that specifies the objects and relations in a given image, and serves us as the machine’s state graph, where objects correspond to states and relations to valid transitions. Multiple models have been proposed for the task of *scene graph generation* [84, 85, 16, 88]. Here, we largely follow the approaches of Yang et al. [85] and Chen et al. [16] in conjunction with a variant of the Mask R-CNN object detector [34] proposed by Hu et al. [39]. Further details regarding the graph generation can be found in section 7.4.

By using such a graph generation model, we can infer a scene graph that consists of: (1) A set of **object nodes**  $S$  from the image, each accompanied by a bounding box, a mask, dense visual features, and a collection of discrete probability distributions  $\{P_i\}_{i=0}^L$  for each of the object’s  $L + 1$  **semantic properties** (such as its color, material, shape, etc.), defined over the concept vocabulary  $\{C_i\}_{i=0}^L$  presented above; (2) A set of **relation edges** between the objects, each associated with a probability



Figure 3: A visualization of object masks from the inferred scene graphs, which form the basis for our model.

distribution  $P_{L+1}$  of its semantic type (e.g. *on top of*, *eating*) among the concepts in  $C_{L+1}$ , and corresponding to a valid transition between the machine’s states.

Once we obtain the sets of state nodes and transition edges, we proceed to computing structured embedded representations for each of them. For each state  $s \in S$  that corresponds to an object in the scene, we define a set of  $L + 1$  property variables  $\{s^j\}_{j=0}^L$  and assign each of them with

$$s^j = \sum_{c_k \in C_j} P_j(k) c_k$$

Where  $c_k \in C_j$  denotes each embedded concept of the  $j^{th}$  property type and  $P_j$  refers to the corresponding property distribution over these concepts, resulting in a soft-binding of concepts to each variable. To give an example, if an object is recognized by the object detector as likely to be e.g. *red*, then its *color* variable will be assigned to an averaged vector close to the embedding of the “*red*” concept. Edge representations are computed in a similar manner, resulting in matching embeddings of their relation type:  $e' = \sum_{c_k \in C_{L+1}} P_{L+1}(k) c_k$  for each edge  $e \in E$ .

Consequently, we obtain a set of structured representations for both the nodes and the edges that underlie the state machine. Note that by associating each object and relation in the scene with not one, but a collection of vectors that capture each of their semantic properties, we are able to create disentangled representations that encapsulate the statistical particularities of the raw image and express it instead through a factorized discrete distribution over a vocabulary of embedded semantic concepts, aiming to encourage and promote higher compositionality.

### 3.3 Reasoning instructions

In the next step, we translate the question into a sequence of reasoning instructions (each expressed in terms of the concept vocabulary  $C$ ), which will later be read by the state machine to guide its computation. The translation process consists of two steps: tagging and decoding.

We begin by embedding all the question words using GloVe (dimension  $d = 300$ ). We process each word with a tagger function that either translates it into the most relevant concept in our vocabulary or alternatively keeps it intact, if it does not match any of them closely enough. Formally, for each embedded word  $w_i$  we compute a similarity-based distribution

$$P_i = \text{softmax}(w_i^T \mathbf{W} C)$$

Where  $\mathbf{W}$  is initialized to the identity matrix and  $C$  denotes the matrix of all embedded concepts along with an additional learned default embedding  $c'$  to account for structural or other non-content words.

Next, we translate each word into a concept-based representation:

$$v_i = P_i(c') w_i + \sum_{c \in C \setminus \{c'\}} P_i(c) c$$

Intuitively, a content word such as *apples* will be considered mostly similar to the concept *apple* (by comparing their GloVe embeddings), and thus will be replaced by the embedding of that term, whereas function words such as *who*, *are*, *how* will be deemed less similar to the semantic concepts and hence will stay close to their original embedding. Overall, this process allows us to “normalize” the question, by transforming content words to their matching concepts, while keeping function words mostly unaffected.

Finally, we process the normalized question words with an attention-based encoder-decoder, drawing inspiration from [40]: Given a question of  $P$  normalized words  $V^{P \times d} = \{v_i\}_{i=1}^P$ , we first pass it through an LSTM encoder, obtaining the final state  $q$  to represent the question. Then, we roll-out a recurrent decoder for a fixed number of steps  $N + 1$ , yielding  $N + 1$  hidden states  $\{h_{i=0}^N\}$ , and transform each of them into a corresponding reasoning instruction:

$$r_i = \text{softmax}(h_i V^T) V$$

Here, we compute attention over the normalized question words at each decoding step. By repeating this process for all  $N + 1$  steps, we decompose the question into a series of reasoning instructions that selectively focus on its various parts, accomplishing the goal of this stage.

### 3.4 Model simulation

Having all the building blocks of the state machine ready, the graph of states  $S$  and edges  $E$ , the instruction series  $\{r_i\}_{i=0}^N$ , and the concept vocabulary  $C = \bigcup_{i=0}^{L+1} C_i$ , we can now simulate the machine's sequential computation. Basically, we will begin with a uniform initial distribution  $p_0$  over the states (the objects in the image's scene), and at each reasoning step  $i$ , read an instruction  $r_i$  as derived from the question, and use it to redistribute our attention over the states (the objects) by shifting probability along the edges (their relations).

Formally, we perform this process by implementing a neural module for the state transition function  $\delta_{S,E} : p_i \times r_i \rightarrow p_{i+1}$ . At each step  $i$ , the module takes a distribution  $p_i$  over the states as an input and computes an updated distribution  $p_{i+1}$ , guided by the instruction  $r_i$ . Our goal is to determine what next states to traverse to  $(p_{i+1})$  based on the states we are currently attending to  $(p_i)$ . To achieve that, we perform a couple of steps.

Recall that in section 3.2 we define for each object a set of  $L + 1$  variables, representing its different properties (e.g. identity, color, shape). We further assigned each edge with a variable that similarly represents its relation type. Our first goal is thus to find the instruction *type*: the property type that is most relevant to the instruction  $r_i$  – basically, to figure out what the instruction is about. We compute the distribution  $R_i = \text{softmax}(r_i^T \circ D)$  over the  $L + 2$  embedded properties  $D$ , defined in section 3.1. We further denote  $R_i(L + 1) \in [0, 1]$  that corresponds to the relation property as  $r'_i$ , measuring the degree to which that reasoning instruction is concerned with semantic relations (in contrast to other possibilities such as e.g. objects or attributes).

Once we know what the instruction  $r_i$  is looking for, we can use it as a guiding signal while traversing the graph from the current states we are focusing on to their most relevant neighbors. We compare the instruction to all the states  $s \in S$  and edges  $e \in E$ , computing for each of them a relevance score:

$$\gamma_i(s) = \sigma \left( \sum_{j=0}^L R_i(j)(r_i \circ \mathbf{W}_j s^j) \right) \quad (1)$$

$$\gamma_i(e) = \sigma(r_i \circ \mathbf{W}_{L+1} e') \quad (2)$$

Where  $\sigma$  is a non-linearity,  $\{s^j\}_{j=0}^L$  are the state variables corresponding to each of its properties, and  $e'$  is the edge variable representing its type. We then get relevance scores between the instruction  $r_i$  and each of the variables, which are finally averaged for each state and edge using  $R_i$ .

Having a relevance score for both the nodes and the edges, we can use them to achieve the key goal of this section: shifting the model's attention  $p_i$  from the current nodes (states)  $s \in S$  to their most relevant neighbors – the next states:

$$p_{i+1}^s = \text{softmax}_{s \in S}(\mathbf{W}_s \cdot \gamma_i(s)) \quad (3)$$

$$p_{i+1}^r = \text{softmax}_{s \in S}(\mathbf{W}_r \cdot \sum_{(s',s) \in E} p_i(s') \cdot \gamma_i((s',s))) \quad (4)$$

$$p_{i+1} = r'_i \cdot p_{i+1}^r + (1 - r'_i) \cdot p_{i+1}^s \quad (5)$$

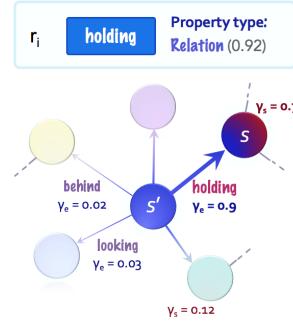


Figure 4: A visualization of a graph traversal step, where attention is being shifted from one node to its neighbor along the most relevant edge.

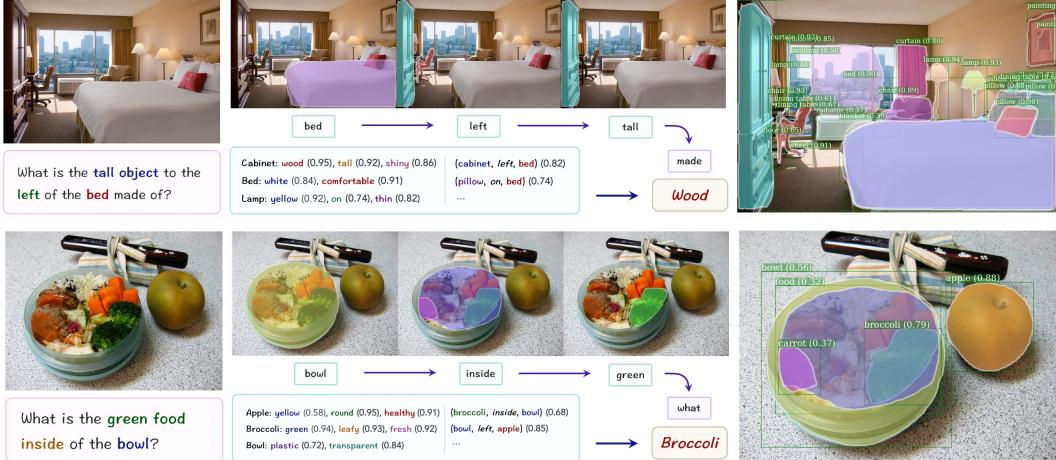


Figure 5: A visualization of the NSM’s reasoning process: given an image and a question (left side), the model first builds a probabilistic scene graph (the blue box and the image on the right), and translates the question into a series of instructions (the green and purple boxes, where for each instruction we present its closest concept (or word) in vector space (section 3.1)). The model then performs sequential reasoning over the graph, attending to relevant object nodes in the image’s scene as guided by the instructions, to iteratively compute the answer.

Here, we compute the distribution over the next states  $p_{i+1}$  by averaging two probabilities  $p_{i+1}^s$  and  $p_{i+1}^r$ : the former is based on each potential next state’s own *internal* properties, while the latter considers the next states *contextual* relevance, relative to the current states the model attends to. Overall, by repeating this process over  $N$  steps, we can simulate the iterative computation of the neural state machine.

After completing the final computation step, and in order to predict an answer, we use a standard 2-layer fully-connected softmax classifier that receives the concatenation of the question vector  $q$  as well as an additional vector  $m$  that aggregates information from the machine’s final states:

$$m = \sum_{s \in S} p_N(s) \left( \sum_{j=0}^L R_N(j) \cdot s^j \right) \quad (6)$$

Where  $m$  reflects the information extracted from the final states as guided by the final reasoning instruction  $r_N$ : averaged first by the reasoning instruction *type*, and then by the attention over the final states, as specified by  $p_N$ .

Overall, the above process allows us to perform a differentiable traversal over the scene graph, guided by the sequence of instructions that were derived from the question: Given an image and a question, we have first inferred a graph to represent the objects and relations in the image’s scene, and analogously decomposed the question into a sequence of reasoning instructions. Notably, we have expressed both the graph and the instructions in terms of the shared vocabulary of semantic concepts, translating them both into the same “internal language”. Then, we simulate the state machine’s iterative operation, and over its course of computation, are successively shifting our attention across the nodes and edges as we ground each instruction in the graph to guide our traversal. Essentially, this allows us to locate each part of the question in the image, and perform sequential reasoning over the objects and relations in the image’s scene graph until we finally arrive at the answer.

## 4 Experiments

We evaluate our model (NSM) on two recent VQA datasets: (1) The GQA dataset [41] which focuses on real-world visual reasoning and compositional question answering, and (2) VQA-CP (version 2) [3], a split of the VQA dataset [27] that has been particularly designed to test generalization skills across changes in the answer distribution between the training and the test sets. We achieve state-of-the-art performance both for VQA-CP, and, under single-model settings, for GQA. To further explore the generalization capacity of the NSM model, we construct two new splits for GQA that test

Table 1: GQA scores for the single-model settings, including official baselines and top submissions

Model	Binary	Open	Consistency	Validity	Plausibility	Distribution	Accuracy
Human [41]	91.20	87.40	98.40	98.90	97.20	-	89.30
Global Prior [41]	42.94	16.62	51.69	88.86	74.81	93.08	28.90
Local Prior [41]	47.90	16.66	54.04	84.33	84.31	13.98	31.24
Language [41]	61.90	22.69	68.68	96.39	<b>87.30</b>	17.93	41.07
Vision [41]	36.05	1.74	62.40	35.78	34.84	19.99	17.82
Lang+Vis [41]	63.26	31.80	74.57	96.02	84.25	7.46	46.55
BottomUp [5]	66.64	34.83	78.71	96.18	84.57	5.98	49.74
MAC [40]	71.23	38.91	81.59	96.16	84.48	5.34	54.06
SK T-Brain*	77.42	43.10	90.78	96.26	85.27	7.54	59.19
PVR*	77.69	43.01	90.35	<b>96.45</b>	84.53	5.80	59.27
GRN	77.53	43.35	88.63	96.18	84.71	6.06	59.37
Dream	77.84	43.72	91.71	96.38	85.48	8.40	59.72
LXRT	77.76	44.97	92.84	96.30	85.19	8.31	60.34
<b>NSM</b>	<b>78.94</b>	<b>49.25</b>	<b>93.25</b>	<b>96.41</b>	84.28	<b>3.71</b>	<b>63.17</b>

generalization over both the questions’ content and structure, and perform experiments based on them that provide substantial evidence for the strong generalization skills of our model across multiple dimensions. Finally, performance diagnosis, ablation studies and visualizations are presented in section 7.2 to shed more light on the inner workings of the model and its qualitative behavior.

Both our model and implemented baselines are trained to minimize the cross-entropy loss of the predicted candidate answer (out of the top 2000 possibilities), using a hidden state size of  $d = 300$  and, unless otherwise stated, length of  $N = 8$  computation steps for the MAC and NSM models. Please refer to section 7.5 for further information about the training procedure, implementation details, hyperparameter configuration and data preprocessing, along with complexity analysis of the NSM model. The model has been implemented in Tensorflow, and will be released along with the features and instructions for reproducing the described experiments.

#### 4.1 Compositional question answering

We begin by testing the model on the GQA task [41], a recent dataset that features challenging compositional questions that involve diverse reasoning skills in real-world settings, including spatial reasoning, relational reasoning, logic and comparisons. We compare our performance both with baselines, as appear in [41], as well as with the top-5 single and top-10 ensemble submissions to the GQA challenge.<sup>1</sup> For single-model settings, to have a fair comparison, we consider all models that, similarly to ours, did not use the strong program supervision as an additional signal for training, but rather learn directly from the questions and answers.

As table 1 shows, we achieve state-of-the-art performance for a single-model across the dataset’s various metrics (defined in [41]) such as accuracy and consistency. For the ensemble setting, we compute a majority vote of 10 instances of our model, achieving the 3rd highest score compared to the 52 submissions that have participated in the challenge<sup>1</sup> (table 2), getting significantly stronger scores compared to the 4th or lower submissions.

Note that while several submissions (marked with \*) use the associated functional programs that GQA provides with each question as a strong supervision during train time, we intentionally did not use them in training our model, but rather aimed to learn the task directly using the question-answer pairs only. These results serve as an indicator for the ability of the model to successfully address questions that involve different forms of reasoning (see section 7 for examples), and especially multi-step inference, which is particularly common in GQA.

#### 4.2 Generalization experiments

Motivated to measure the generalization capacity of our model, we perform experiments over three different dimensions: (1) changes in the answer distribution between the training and the test sets, (2) contextual generalization for concepts learned in isolation, and (3) unseen grammatical structures.

<sup>1</sup>The official leaderboard mixes up single-model and ensemble results – we present here separated scores for each track.

Table 2: GQA ensemble

Model	Accuracy
Kakao*	73.33
270	70.23
NSM	<b>67.25</b>
LXRT	62.71
GRN	61.22
MSM	61.09
DREAM	60.93
SK T-Brain*	60.87
PKU	60.79
Musan	59.93

Table 3: VQA-CPv2

Model	Accuracy
SAN [86]	24.96
HAN [59]	28.65
GVQA [3]	31.30
RAMEN [73]	39.21
BAN [46]	39.31
MuRel [15]	39.54
ReGAT [52]	40.42
<b>NSM</b>	<b>45.80</b>

Table 4: GQA generalization

Model	Content	Structure
Global Prior	8.51	14.64
Local Prior	12.14	18.21
Vision	17.51	18.68
Language	21.14	32.88
Lang+Vis	24.95	36.51
BottomUp [5]	29.72	41.83
MAC [40]	31.12	47.27
<b>NSM</b>	<b>40.24</b>	<b>55.72</b>

<i>Structure Generalization</i>		<i>Content Generalization</i>	
training	testing	training	testing
What is the <obj> <b>covered by</b> ? Is there a <obj> in the <b>image</b> ? What is the <obj> <b>made of</b> ? What's the <b>name</b> of the <obj> that is <attr>?	What is <b>covering</b> the <obj>? Do you see any <obj>s in the <b>photo</b> ? What <b>material makes up</b> the <obj>? What is the <attr> <obj> <b>called</b> ?	Only questions that <b>do not</b> refer to any type of <b>food</b> or <b>animal</b> (do not include any word from these categories)	Only questions that refer to <b>foods</b> or <b>animals</b> (include a word from one of these categories)

Figure 6: Our new generalization splits for GQA, evaluating generalization over (1) content: where test questions ask about novel concepts, and (2) structure: where test questions follow unseen linguistic patterns.

First, we measure the performance on VQA-CP [3], which provides a new split of the VQA2 dataset [27], where the answer distribution is kept different between the training and the test sets (e.g. in the training set, the most common color answer is *white*, whereas in the test set, it is *black*). Such settings reduce the extent to which models can circumvent the need for genuine scene understanding skills by exploiting dataset biases and superficial statistics [1, 44, 27], and are known to be particularly difficult for neural networks [51]. Here, we follow the standard VQA1/2 [27] accuracy metric for this task (defined in [3]). Table 3 presents our performance compared to existing approaches. We can see that NSM surpasses alternative models by a large margin.

We perform further generalization studies on GQA, leveraging the fact that the dataset provides grounding annotations of the question words. For instance, a question such as “*What color is the book on the table?*” is accompanied by the annotation  $\{4 : (“book”, n_0), 7 : (“table”, n_1)\}$  expressing the fact that e.g. the 4<sup>th</sup> word refers to the *book* object node. These annotations allow us to split the training set in two interesting ways to test generalization over both *content* and *structure* (see figure 6 for an illustration of each split):

**Content:** Since the annotations specify which objects each question refers to, and by using the GQA ontology, we can identify all the questions that are concerned with particular object types, e.g. foods, or animals. We use this observation to split the training set by excluding all question-answer pairs that refer to these categories, and measure the model’s generalization over them. Note however, that the object detector module described in section 3.2 is still trained over all the scene graphs including those objects – rather, the goal of this split is to test whether the model can leverage the fact that it was trained to identify a particular object in isolation, in order to answer unseen questions about that type of object without any further question training.

**Structure:** We can use the annotations described above as masks over the objects (see figure 6 for examples), allowing us to divide the questions in the training set into linguistic pattern groups. Then, by splitting these groups into two separated sets, we can test whether a model is able to generalize from some linguistic structures to unseen ones.

Table 4 summarizes the results for both settings, comparing our model to the baselines released for GQA [41], all using the same training scheme and input features. We can see that here as well, NSM performs significantly better than the alternative approaches, testifying to its strong generalization capacity both over concepts it has not seen any questions about (but only learned in isolation), as well as over questions that involve novel linguistic structures. In our view, these results point to the strongest quality of our approach. Several prior works have argued for the great potential of abstractions and compositionality in enhancing models of deep learning [8, 10]. Our results suggest that incorporating these notions may indeed be highly beneficial to creating models that are more capable in coping with changing conditions and can better generalize to novel situations.

## 5 Conclusion

In this paper, we have introduced the Neural State Machine, a graph-based network that simulates the operation of an automaton, and demonstrated its versatility, robustness and high generalization skills on the tasks of real-world visual reasoning and compositional question answering. By incorporating the concept of a state machine into neural networks, we are able to introduce a strong structural prior that enhances compositionality both in terms of the *representation*, by having a structured graph to serve as our world model, as well as in terms of the *computation*, by performing sequential reasoning over such graphs. We hope that our model will help in the effort to integrate symbolic and connectionist approaches more closely together, in order to elevate neural models from sensory and perceptual tasks, where they currently shine, into the domains of higher-level abstraction, knowledge representation, compositionality and reasoning.

## 6 Acknowledgments

We thank Samsung Electronics Co., Ltd. for partially funding this project. We also would like to thank the reviewers for their detailed and constructive comments, suggestions and feedback. Finally, we would like to thank the readers who have contacted us with useful comments and questions.

## References

- [1] Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. Analyzing the behavior of visual question answering models. In *EMNLP*, pp. 1955–1960, 2016.
- [2] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C Lawrence Zitnick, Devi Parikh, and Dhruv Batra. VQA: Visual question answering. *International Journal of Computer Vision*, 123(1):4–31, 2017.
- [3] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don’t just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4971–4980, 2018.
- [4] Igor Aleksander. The consciousness of a neural state machine. In *International Conference on Artificial Neural Networks*, pp. 212–217. Springer, 1994.
- [5] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and VQA. *arXiv preprint arXiv:1707.07998*, 2017.
- [6] Jacob Andreas. Measuring compositionality in representation learning. *arXiv preprint arXiv:1902.07181*, 2019.
- [7] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 39–48, 2016.
- [8] Jacob Andreas, Dan Klein, and Sergey Levine. Learning with latent language. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2166–2179, 2018.
- [9] Lawrence W Barsalou, W Kyle Simmons, Aron K Barbey, and Christine D Wilson. Grounding conceptual knowledge in modality-specific systems. *Trends in cognitive sciences*, 7(2):84–91, 2003.
- [10] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.
- [11] George Bealer. A theory of concepts and concept possession. *Philosophical Issues*, 9:261–301, 1998.
- [12] Yoshua Bengio. The consciousness prior. *arXiv preprint arXiv:1709.08568*, 2017.
- [13] Lera Boroditsky. How language shapes thought. *Scientific American*, 304(2):62–65, 2011.
- [14] Léon Bottou. From machine learning to machine reasoning. *Machine learning*, 94(2):133–149, 2014.

- [15] Remi Cadene, Hedi Ben-Younes, Matthieu Cord, and Nicolas Thome. Murel: Multimodal relational reasoning for visual question answering. *arXiv preprint arXiv:1902.09487*, 2019.
- [16] Tianshu Chen, Weihao Yu, Riquan Chen, and Liang Lin. Knowledge-embedded routing network for scene graph generation. *arXiv preprint arXiv:1903.03326*, 2019.
- [17] Edward Choi, Angeliki Lazaridou, and Nando de Freitas. Compositional obverter communication learning from raw visual input. *arXiv preprint arXiv:1804.02341*, 2018.
- [18] Noam Chomsky. *Aspects of the Theory of Syntax*, volume 11. MIT press, 2014.
- [19] Noam Chomsky. The language capacity: architecture and evolution. *Psychonomic bulletin & review*, 24(1):200–203, 2017.
- [20] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100, 2017.
- [21] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1778–1785. IEEE, 2009.
- [22] Jerry A Fodor. *The language of thought*, volume 5. Harvard university press, 1975.
- [23] Mikel L Forcada and Rafael C Carrasco. Finite-state computation in analog neural networks: steps towards biologically plausible models? In *Emergent neural computational architectures based on neuroscience*, pp. 480–493. Springer, 2001.
- [24] Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. Semantic compositional networks for visual captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5630–5639, 2017.
- [25] Marta Garnelo, Kai Arulkumaran, and Murray Shanahan. Towards deep symbolic reinforcement learning. *arXiv preprint arXiv:1609.05518*, 2016.
- [26] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
- [27] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pp. 6325–6334, 2017.
- [28] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural Turing machines. *arXiv preprint arXiv:1410.5401*, 2014.
- [29] Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, et al. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471–476, 2016.
- [30] James G Greeno. A perspective on thinking. *American Psychologist*, 44(2):134–141, 1989.
- [31] Akshay Kumar Gupta. Survey of visual question answering: Datasets and techniques. *arXiv preprint arXiv:1705.03865*, 2017.
- [32] Gilbert Harman. *Change in view: Principles of reasoning*. The MIT Press, 1986.
- [33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- [34] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.
- [35] Irina Higgins, Loic Matthey, Xavier Glorot, Arka Pal, Benigno Uria, Charles Blundell, Shakir Mohamed, and Alexander Lerchner. Early visual concept learning with unsupervised deep learning. *arXiv preprint arXiv:1606.05579*, 2016.
- [36] Irina Higgins, Nicolas Sonnerat, Loic Matthey, Arka Pal, Christopher P Burgess, Matko Bosnjak, Murray Shanahan, Matthew Botvinick, Demis Hassabis, and Alexander Lerchner. Scan: Learning hierarchical compositional visual concepts. *arXiv preprint arXiv:1707.03389*, 2017.

- [37] John E Hopcroft. *Introduction to automata theory, languages, and computation*. Pearson Education India, 2008.
- [38] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 804–813, 2017.
- [39] Ronghang Hu, Piotr Dollár, Kaiming He, Trevor Darrell, and Ross Girshick. Learning to segment everything. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4233–4241, 2018.
- [40] Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. In *International Conference on Learning Representations (ICLR)*, 2018.
- [41] Drew A Hudson and Christopher D Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [42] Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*, 2017.
- [43] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3668–3678, 2015.
- [44] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pp. 1988–1997. IEEE, 2017.
- [45] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1219–1228, 2018.
- [46] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *Advances in Neural Information Processing Systems*, pp. 1564–1574, 2018.
- [47] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [48] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.
- [49] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.
- [50] Brenden M Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. *arXiv preprint arXiv:1711.00350*, 2017.
- [51] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, 2017.
- [52] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. Relation-aware graph attention network for visual question answering. *arXiv preprint arXiv:1903.12314*, 2019.
- [53] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. Relation-aware graph attention network for visual question answering. *arXiv preprint arXiv:1903.12314*, 2019.
- [54] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. Scene graph generation from objects, phrases and region captions. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1261–1270, 2017.
- [55] Yikang Li, Wanli Ouyang, Bolei Zhou, Jianping Shi, Chao Zhang, and Xiaogang Wang. Factorizable net: An efficient subgraph-based framework for scene graph generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 335–351, 2018.
- [56] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2117–2125, 2017.

- [57] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 375–383, 2017.
- [58] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7219–7228, 2018.
- [59] Mateusz Malinowski, Carl Doersch, Adam Santoro, and Peter Battaglia. Learning visual question answering by bootstrapping hard attention. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3–20, 2018.
- [60] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. *arXiv preprint arXiv:1904.12584*, 2019.
- [61] David Mascharka, Philip Tran, Ryan Soklaski, and Arjun Majumdar. Transparency by design: Closing the gap between performance and interpretability in visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4942–4950, 2018.
- [62] Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. Key-value memory networks for directly reading documents. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1400–1409, 2016.
- [63] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.
- [64] David Navon. Forest before trees: The precedence of global features in visual perception. *Cognitive psychology*, 9(3):353–383, 1977.
- [65] Allen Newell. Physical symbol systems. *Cognitive science*, 4(2):135–183, 1980.
- [66] Will Norcliffe-Brown, Stathis Vafeias, and Sarah Parisot. Learning conditioned graph structures for interpretable visual question answering. In *Advances in Neural Information Processing Systems*, pp. 8334–8343, 2018.
- [67] Diane Pecher and Rolf A Zwaan. *Grounding cognition: The role of perception and action in memory, language, and thinking*. Cambridge University Press, 2005.
- [68] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.
- [69] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7263–7271, 2017.
- [70] Timothy T Rogers and James L McClelland. *Semantic cognition: A parallel distributed processing approach*. MIT press, 2004.
- [71] Adam Santoro, Felix Hill, David Barrett, Ari Morcos, and Timothy Lillicrap. Measuring abstract reasoning in neural networks. In *International Conference on Machine Learning*, pp. 4477–4486, 2018.
- [72] Susan Schneider. *The language of thought: A new philosophical direction*. Mit Press, 2011.
- [73] Robik Shrestha, Kushal Kafle, and Christopher Kanan. Answer them all! Toward universal visual question answering models. *arXiv preprint arXiv:1903.00366*, 2019.
- [74] Paul Smolensky. Connectionist AI, symbolic AI, and the brain. *Artificial Intelligence Review*, 1(2):95–109, 1987.
- [75] Damien Teney, Lingqiao Liu, and Anton van den Hengel. Graph-structured representations for visual question answering. *arXiv preprint*, 2017.
- [76] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.
- [77] Julia Vogel and Bernt Schiele. Semantic modeling of natural scenes for content-based image retrieval. *International Journal of Computer Vision*, 72(2):133–157, 2007.

- [78] Lev Semenovich Vygotsky. Thought and language. *Annals of Dyslexia*, 14(1):97–98, 1964.
- [79] Peng Wang, Qi Wu, Chunhua Shen, and Anton van den Hengel. The vqa-machine: Learning how to use existing vision algorithms to answer new questions. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, volume 4, 2017.
- [80] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. *arXiv preprint arXiv:1410.3916*, 2014.
- [81] Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton Van Den Hengel. What value do explicit high level concepts have in vision to language problems? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 203–212, 2016.
- [82] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.
- [83] Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. In *International conference on machine learning*, pp. 2397–2406, 2016.
- [84] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, 2017.
- [85] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. *arXiv preprint arXiv:1808.00191*, 2018.
- [86] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 21–29, 2016.
- [87] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. In *Advances in Neural Information Processing Systems*, pp. 1031–1042, 2018.
- [88] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5831–5840, 2018.

## 7 Supplementary material

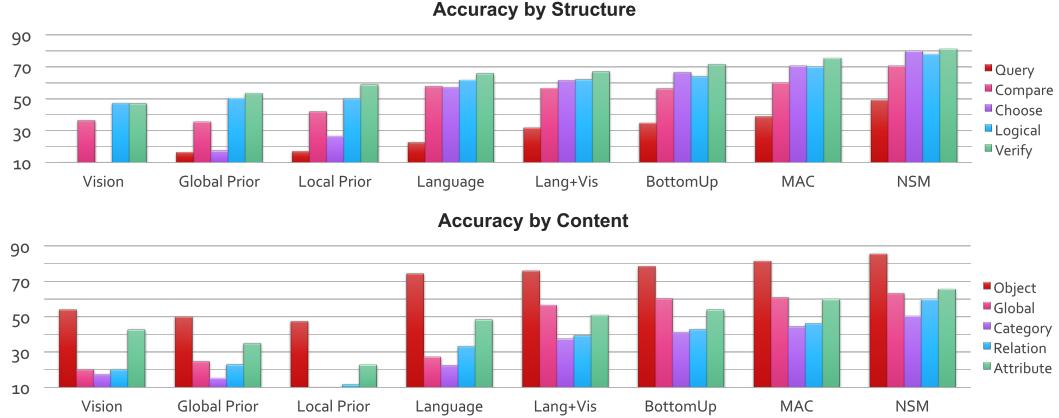


Figure 7: Accuracy of NSM and baselines for different structural and semantic question types of the GQA dataset. NSM surpasses the baselines for all question types, with the largest gains for **relational** and **comparative** questions, and high improvements for **attribute**, **logic** and **query** (open) questions.

### 7.1 Related work (full version)

Our model connects to multiple lines of research, including works about concept acquisition [36], visual abstractions [24, 81, 58, 57, 87], compositionality [14, 38], and neural computation [28, 62, 7]. Several works have explored the discovery and use of visual concepts in the contexts of reinforcement or unsupervised learning [35, 17] as well as in classical computer vision [21, 77]. Others have argued for the importance of incorporating strong inductive biases into neural architectures [14, 10, 6, 8], and indeed, there is a growing body of research that seeks to introduce different forms of structural priors inspired by computer architectures [29, 80, 40] or theory of computation [4, 23], aiming to bridge the gap between the symbolic and neural paradigms.

One such structural prior that underlies our model is that of the probabilistic scene graph [43, 49] which we construct and reason over to answer questions about presented images. Scene graphs provide a succinct representation of the image’s semantics, and have been effectively used for variety of applications such as image retrieval, captioning or generation [43, 54, 45]. Recent years have witnessed an increasing interest both in scene graphs in particular [84, 85, 16, 88, 55] as well as in graph networks in general [10, 48, 76] – a family of graph-structured models in which information is iteratively propagated across the nodes to turn their representations increasingly more contextual with information from their neighborhoods. In our work, we also use the general framework of graph networks, but in contrast to common approaches, we avoid the computationally-heavy state updates per each node, keeping the graph states static once predicted, and instead perform a series of refinements of one global attention distribution over the nodes, resulting in a soft graph traversal which better suits our need for supporting sequential reasoning.

We explore our model in the context of visual question answering [31], a challenging multi-modal task that has gained substantial attention over the last years. Plenty of models have been proposed [5, 79, 40], focusing on visual reasoning or question answering in both abstract and real-world settings [27, 44, 71, 41]. Typically, most approaches use either CNNs [83, 86] or object detectors [5] to derive visual features which are then compared to a fixed-size question embedding obtained by an LSTM. A few newer models use the relationships among objects to augment features with contextual information from each object’s surroundings [53, 75, 66]. We move further in this direction, performing iterative reasoning over the inferred scene graphs, and in contrast to prior models [52, 15], incorporate higher-level semantic concepts to represent both the visual and linguistic modalities in a shared and sparser manner to facilitate their interaction.

Other methods for visual reasoning such as [87, 60, 61] have similar motivation to ours of integrating neural and symbolic perspectives by learning and reasoning over structured representations. However, in contrast to our approach, those models heavily rely on symbolic execution, either through strong supervision of program annotations [61], non-differentiable python-based functions [87], or a

collection of hand-crafted modules specifically designed for each given task [60], and consequently, they have mostly been explored in artificial environments such as CLEVR [44]. Instead, our model offers a fully-neural and more general graph-based design that, as demonstrated by our experiments, successfully scales to real-world settings.

Closest to our work is a model called MAC [40], a recurrent network that applies attention-based operations to perform sequential reasoning. However, the Neural State Machine differs from MAC in two crucial respects: First, we reason over graph structures rather than directly over spatial maps of visual features, traversing the graph by successively shifting attention across its nodes and edges. Second, key to our model is the notion of semantic concepts that are used to express knowledge about both the visual and linguistic modalities, instead of working directly with the raw observational features. As our findings suggest, these ideas contribute significantly to the model’s performance compared to MAC and enhance its compositionality, transparency and generalization skills.

## 7.2 Ablation studies

To gain further insight into the relative contributions of different aspects of our model to its overall performance, we have conducted multiple ablation experiments, as summarized in table 5 and figure 4. First, we measure the impact of using our new visual features (section 3.2) compared to the default features used by the official baselines [41]. Such settings result in an improvement of 1.27%, confirming that most of the gain achieved by the NSM model (with 62.95% overall) stems from its inherent architecture rather than from the input features.

We then explore several ablated models: one where we do not perform any iterative simulation of the state machine, but rather directly predict the answer from its underlying graph structure (55.41%); and a second enhanced version where we perform a traversal across the states but without considering the typed relations among them, adopting instead a uniform fully-connected graph (58.83%). These experiments suggest that using the graph structure to represent the visual scene as well performing sequential reasoning over it by traversing its nodes, are both crucial to the model’s overall accuracy and have positive impact on its performance.

Next, we evaluate a variant of the model where instead of using the concept-based representations defined in section 3.1, we fallback to the more standard dense features to represent the various graph elements, which results in an accuracy of 58.48%. Comparing this score to that of the default model’s settings (62.95%) proves the significance of using the higher-level semantic representations to the overall accuracy of the NSM.

Finally, we measure the impact of varying the number of computation steps on the obtained results (figure 4), revealing a steady increase in accuracy as we perform a higher number of reasoning steps (until saturation for  $N = 8$  steps). These experiments further validate the effectiveness of sequential computation in addressing challenging questions, and especially compositional questions which are at the core of GQA.

## 7.3 Concept vocabulary

We define a vocabulary  $C$  of 1335 embedded concepts about object types  $C_0$  (e.g. *cat*, *shirt*), attributes, grouped into  $L$  types  $\{C_i\}_{i=1}^L$  (e.g. *colors*, *materials*), and relations  $C_{L+1}$  (e.g. *holding*, *on top of*), all initialized with GloVe [68]. Our concepts consist of 785 objects, 170 relations, and 303 attributes that are divided into 77 types, with most types being binary (e.g. *short/tall*, *light/dark*). All concepts are derived from the Visual Genome dataset [49], which provides human-annotated scene graphs for real-world images. We use the refined dataset supported by GQA [41], where the graphs are further cleaned-up and consolidated, resulting in a closed-vocabulary version of the dataset.

## 7.4 Scene graph generation

In order to generate scene graphs from given images, we re-implement a model that largely follows Yang et al. [85], Chen et al. [16] in conjunction with an object detector proposed by Hu et al. [39].

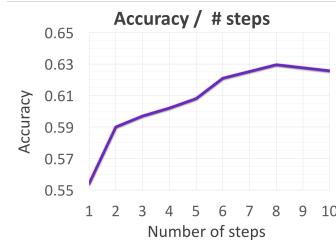


Figure 8: Accuracy as a function of the number of computation steps. Performance is reported on the GQA Test-Dev split.

Table 5: Ablations (reported on GQA Test-Dev)

Model	Accuracy
NSM (default)	$62.95 \pm 0.22$
No concepts	$58.48 \pm 0.14$
No relations (set)	$58.83 \pm 0.17$
No traversal	$55.41 \pm 0.35$
Visual features	$47.82 \pm 0.09$
Baseline (Lang+Vis)	$46.55 \pm 0.13$

Specifically, we use a variant of Mask R-CNN [34, 39] to obtain object detections from each image, using Hu et al. [39]’s official implementation along with ResNet-101 [33] and FPN [56] for features and region proposals respectively, and keep up to 50 detections, with a confidence threshold of 0.2. We train the object detector (and the following graph generation model) over a cleaner version of Visual Genome scene graphs [49], offered as part of the GQA dataset [41]. In particular, the detector heads are trained to classify both the object class and category (akin to YOLO’s hierarchical softmax [69]), as well as the object’s attributes per each *property* type, resulting in a set of probability distributions  $\{P_i\}_{i=0}^L$  over the object’s various properties (e.g. its *identity*, *color* or *shape*).

Once we obtain the graph nodes – the set of detected objects, we proceed to detecting their relations, mainly following the approach of Yang et al. [85]: First, we create a directed edge for each pair of objects that are in close enough proximity – as long as the relative distance between the objects in both axes is smaller than 15% of the image dimensions (which covers over 94% of the ground truth edges, and allows us to sparsify the graph, reducing the computation load). Once we set the graph structure, we process it through a graph attention network [76] to predict the identities of each relation, resulting in a probability distribution  $P_{L+1}$  over the relation type of each edge.

## 7.5 Implementation and training details

We train the model using the Adam optimization method [47], with a learning rate of  $10^{-4}$  and a batch size of 64. We use gradient clipping, and employ early stopping based on the validation accuracy, leading to a training process of approximately 15 epochs, equivalent to roughly 30 hours on a single Maxwell Titan X GPU. Both hidden states and word vectors have a dimension size of 300, the latter being initialized using GloVe [68]. During the training, we maintain exponential moving averages of the model weights, with a decay rate of 0.999, and use them at test time instead of the raw weights. Finally, we use ELU as non-linearity and dropout of 0.15 across the network: in the initial processing of images and questions, for the state and edge representations, and in the answer classifier. All hyperparameters were tuned manually (from the following ranges: learning rate [ $5 \cdot 10^{-5}, 10^{-4}, 5 \cdot 10^{-4}, 10^{-3}$ ], batch size [32, 64, 128], dropout [0.08, 0.1, 0.12, 0.15, 0.18, 0.2] and hidden and word dimensions [50, 100, 200, 300] to comply with GloVe provided sizes).

We have preprocessed all the questions by removing punctuation and keeping the top 5000 most common words, and use the standard training/test splits provided by the original datasets we have explored [41, 3]. For GQA, we use the more common “balanced” version that has been designed to reduce biases within the answer distribution (similar in motivation to the VQA2 dataset [27]), and includes 1.7M questions split into 70%/10%/10% for training, validation and test sets respectively. For VQA-CP, we use version v2 which consists of 438k/220k questions for training/test respectively. Finally, for the new generalization split, we have downsampled the data to have a ratio of 80%/20% for training/test over approximately 800k questions overall. All results reported are for a single-model settings, except the ensemble scores for GQA that compute majority vote over 10 models, and the ablation studies that are performed over 5 runs for each ablated version. To measure the confidence of the results, we have performed additional 5 runs of our best-performing model over both the GQA Test-Dev and VQA-CP, getting standard deviations of 0.22 and 0.31 respectively.

In terms of time complexity, the NSM is linear both in the number of question words  $P$ , and the size of the graph ( $S \leq 50$  states and  $E \approx O(S)$  due to the proximity-based pruning we perform while constructing the graph), multiplied by a constant  $N = 8$  of reasoning steps  $O(N(V + E + P))$ . Similarly, the space complexity of our model is  $O(V + E + P)$  since we do not have to store separated values for each computation step, but rather keep only the most recent ones.