COMPOSITIONAL VISUAL GENERATION WITH ENERGY BASED MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Humans are able to both learn quickly and rapidly adapt their knowledge. One major component is the ability to incrementally combine many simple concepts to accelerates the learning process. We show that energy based models are a promising class of models towards exhibiting these properties by directly combining probability distributions. This allows us to combine an arbitrary number of different distributions in a globally coherent manner. We show this compositionality property allows us to define three basic operators, logical conjunction, disjunction, and negation, on different concepts to generate plausible naturalistic images. Furthermore, by applying these abilities, we show that we are able to extrapolate concept combinations, continually combine previously learned concepts, and infer concept properties in a compositional manner.

1 INTRODUCTION

Humans are able to rapidly learn new concepts and continually integrate them among their prior knowledge. The key ingredient in enabling this is the ability to compose increasingly complex concepts out of simpler ones, and recombining and reusing concepts in novel ways (Fodor & Lepore, 2002). By combining a finite number of primitive components, humans can create an exponential number of new concepts, and use them to rapidly explain current and past experiences (Lake et al., 2017). We are interested in enabling such compositionality capabilities in machine learning systems, particularly in the generative modeling context.

Past efforts in machine learning to incorporate compositionality have attempted it in several distinct ways. One has been to decompose data into disentangled factors of variation and situate each datapoint in the resulting - typically continuous - factor vector space (Vedantam et al., 2018; Higgins et al., 2018). The factors can either be explicitly provided or learned in an unsupervised manner. In both cases, however, the dimensionality of the factor vector space is fixed and defined prior to training. This makes it difficult to introduce new factors of variation, which may be necessary to explain new data, or to differently taxonomize past data. Another approach to incorporate the compositionality is to spatially decompose an image into a collection of objects, each object slot occupying some pixels of the image defined by a segmentation mask (van Steenkiste et al., 2018; Greff et al., 2019). Such approaches can generate visual scenes with multiple objects, but may have difficulty in generating interacting effects between objects. These two incorporations of compositionality are typically seen as distinct, with very different underlying implementations.

In this work, we propose to implement compositionality ideas via energy based models (EBMs). Instead of an explicit vector of factors that is input to a generator function, or object slots that are blended to form an image, our unified treatment defines factors of variation and object slots via energy functions. Each factor is represented by an individual scalar energy function that takes as input an image and outputs a low energy value if the factor is exhibited in the image. Images that exhibit the factor can then be generated implicitly as a result of an MCMC process that minimizes the energy. Importantly, it is also possible to run MCMC process on some *combination* of energy functions to generate images that exhibit multiple factors or multiple objects, in a globally coherent manner.

There are several ways to combine energy functions. One can add or multiply distributions defined by the energy functions (as in mixtures (Shazeer et al., 2017; Greff et al., 2019) or products (Hinton, 2002) of experts). We view these as probabilistic instances of logical operators over concepts. Instead of using one, we consider three operators: logical conjunction, disjunction, and negation (illustrated in Figure 1). We can then flexibly and recursively combine multiple energy functions via these operators. More complex operators (such as implication) can be formed out of our base operators.

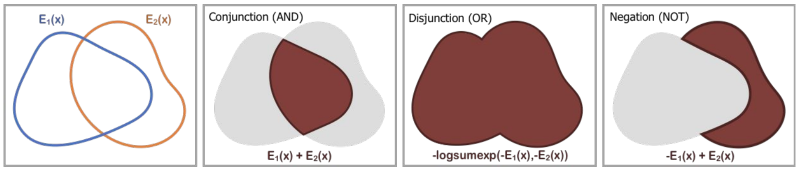


Figure 1: Illustration of logical composition operators over energy functions E_1 and E_2 (drawn as level sets).

EBMs with such logical composition operators enable several capabilities. They allow defining new concepts (factors) implicitly via examples. This is similar to learning to generate images in a few-shot setting (Reed et al., 2017), with the distinction that instead of learning to generate holistic images from few examples, we learn *properties* from examples in a way that can then be flexibly combined with other previously learned concepts. This allows new concepts to be added on demand in a continual manner by simply learning a new energy function from examples, and which again can be combined with all past concepts. Additionally, finely controllable image generation can be enabled by specifying the desired image via a collection of logical clauses, with applications to neural scene rendering (Eslami et al., 2018).

Our contributions are as follows: first, while composition of energy-based models has been proposed in abstract settings before (Hinton, 2002), we show that it can be used to generate plausible natural images. Second, we propose to combine energy models based on logical operators which can be chained recursively, allowing controllable generation based on a collection of logical clauses. Third, we demonstrate unique advantages of such an approach, such as extrapolation to concept combinations, continual addition of new energy functions, and ability to infer concept properties.

2 Method

In this section, we first give a background overview of EBMs and then define three different basic logic operators on them. The components of these operators can be learnt independently and incrementally combined to support continual learning. Furthermore, the operators themselves can be combined to support nested compositions.

2.1 Energy Based Models

EBMs represent data by learning an unnormalized probability distribution across the data. For each data point $\mathbf{x}$, an energy function $E_\theta(\mathbf{x})$, parameterized by a neural network, outputs a scalar real energy such that

$$p_\theta(x) \propto e^{-E_\theta(x)}. \tag{1}$$

To train an EBM on a data distribution p_D, we follow the methodology defined in (Du & Mordatch, 2019), where a Monte Carlo estimate (Equation 2) of maximum likelihood is minimized.

$$\nabla_\theta \mathcal{L} = \mathbb{E}_{x^+ \sim p_D} E_\theta(x^+) - \mathbb{E}_{x^- \sim p_\theta} E_\theta(x^-). \tag{2}$$

To sample x^- from p_θ for both training and generation, we use MCMC based off Langevin dynamics (Du & Mordatch, 2019). Samples are initialized from uniform random noise and are iteratively refined following Equation 3

$$\tilde{\mathbf{x}}^k = \tilde{\mathbf{x}}^{k-1} - \frac{\lambda}{2} \nabla_{\mathbf{x}} E_\theta(\tilde{\mathbf{x}}^{k-1}) + \omega^k, \; \omega^k \sim \mathcal{N}(0, \lambda), \tag{3}$$

where k is the k^{th} iteration step and λ is the step size. We refer to each iteration of Langevin dynamics as a negative sampling step. We note that this form of sampling allows us to generate samples from distributions composed of p_θ and other distributions by using the gradient of the modified distribution. We use this ability to generate from multiple distributions that allow various different forms of compositionality that we detail below.

2.2 Composition of Energy-Based Models

We next present different ways that EBMs can compose. We consider a set of independently trained EBMs, $E(\mathbf{x}|c_1), E(\mathbf{x}|c_2), \ldots, E(\mathbf{x}|c_n)$, which are learned conditional distributions on underlying latents c_i. Latents we consider including position, size, color, gender, hair style, and age, which we refer to as concepts. Figure 2 shows three concepts and their combinations.

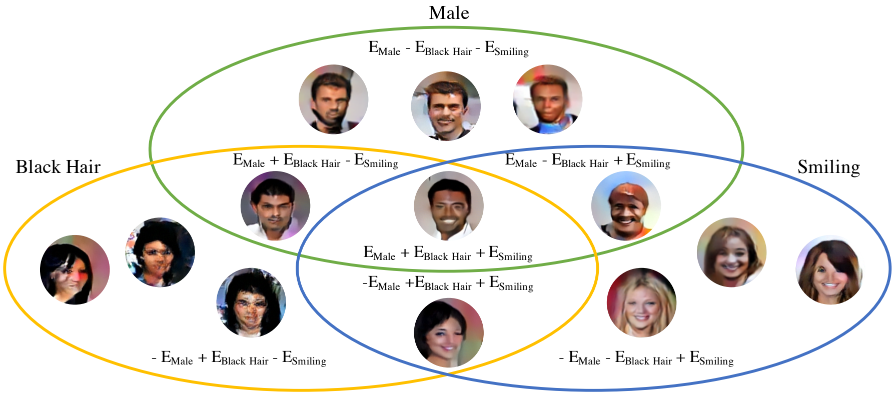

Male
E_{Male} - $E_{\text{Black Hair}}$ - E_{Smiling}
Black Hair
E_{Male} + $E_{\text{Black Hair}}$ - E_{Smiling}
E_{Male} - $E_{\text{Black Hair}}$ + E_{Smiling}
Smiling
E_{Male} + $E_{\text{Black Hair}}$ + E_{Smiling}
-E_{Male} +$E_{\text{Black Hair}}$ + E_{Smiling}
- E_{Male} + $E_{\text{Black Hair}}$ - E_{Smiling}
- E_{Male} - $E_{\text{Black Hair}}$ + E_{Smiling}

Figure 2: Illustration of concept conjunction and negation. All the images are generated through the conjunction and negation of energy functions. For example, images in the central part is the conjunction of male, black hair, and smiling energy function.

Concept Conjunction In concept conjunction, given separate independent concepts such as a particular gender, hair style, and facial expression, we wish to construct an output with the specified gender, hair style, and facial expression – the combination of each concept. Since the likelihood of an output given a set of specific concepts is equal to the product of the likelihood of each individual concept, we have Equation 4, which is also known as the product of experts (Hinton, 2002)

$$p(x|c_1 \text{ and } c_2, \dots, \text{ and } c_i) = \prod_i p(x|c_i) \propto e^{\sum_i E(x|c_i)}. \tag{4}$$

We can thus apply Equation 3 to the distribution that is the sum of the energies of each concept to obtain Equation 5 to sample from the joint concept space.

$$\tilde{\mathbf{x}}^k = \tilde{\mathbf{x}}^{k-1} - \frac{\lambda}{2}\nabla_{\mathbf{x}} \sum_i E_\theta(\tilde{\mathbf{x}}^{k-1}|c_i) + \omega^k,\ \omega^k \sim \mathcal{N}(0, \lambda) \tag{5}$$

Concept Disjunction In concept disjunction, given separate concepts such as the color red and the color blue, we wish to construct an output that is either red or blue – either of the given concepts. Thus, we wish to construct a new distribution which is sharply peaked when any of the chosen concepts are true. A natural choice of such a distribution is the sum of the likelihood of each concept (Equation 6) – which will be sharp whenever any of the chosen concepts are true.

$$p(x|c_1 \text{ or } c_2, \dots \text{ or } c_i) \propto \sum_i p(x|c_i) \propto \sum_i e^{-E(x|c_i)} \propto e^{\text{logsumexp}(-E(x|c_i))}, \tag{6}$$

where $\text{logsumexp}(f_1, \dots, f_N) = \log \sum_i \exp(f_i)$. We can thus apply Equation 3 to the distribution that is logsumexp of the negative energies of each concept to obtain Equation 7 to sample from the additive concept space.

$$\tilde{\mathbf{x}}^k = \tilde{\mathbf{x}}^{k-1} + \frac{\lambda}{2}\nabla_{\mathbf{x}}\text{logsumexp}(-E(x|c_i)) + \omega^k,\ \omega^k \sim \mathcal{N}(0, \lambda) \tag{7}$$

Concept Negation In concept negation, we wish to construct an output that does not contain the concept. Given a color red, we wish to construct an output that is of a different color, such as blue. Thus, we want to construct a distribution that places high likelihood to data that is outside a given concept. One way to generate such a distribution is to construct a probability distribution parameterized by an EBM where the energy is a negative scalar multiplies the energy of the target concept. However, an important issue is that negation is always defined with respect to another data distribution – the opposite of alive may be dead, but not inanimate. Negation without a data distribution is not integrable and leads to a generation of chaotic textures that, although indeed is the absence of concept desired, also does not capture the essence of negation. Thus in our experiments with negation, we jointly combine negation with another conditional model to ground the negation and obtain the probability distribution in Equation 8.

$$p(x|\text{not}(c_1), c_2) \propto e^{\alpha E(x|c_1) - E(x|c_2)} \tag{8}$$

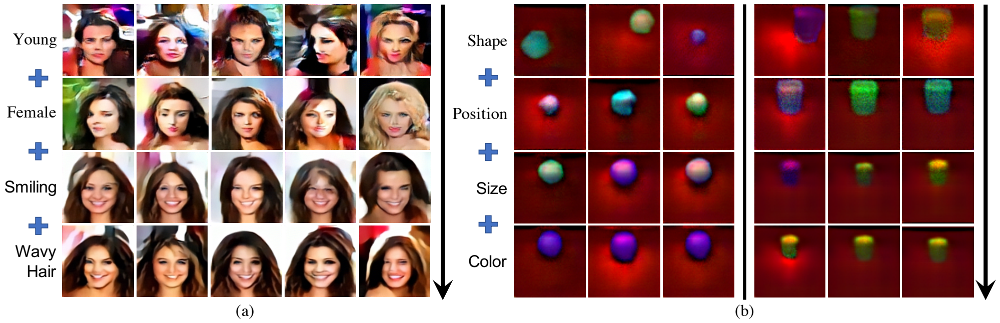


Figure 3: Examples of combination of different attributes on CelebA (a) and Mujoco scenes (b) via summation of energies. Each row adds an additional energy function attribute. For example, (a) images on the first row are only conditioned on young while images on the last row are conditioned on young, female, smiling and wavy hair; (b) images on the first column are only conditioned on shape while images on the last column are conditioned on shape, position, size and color. The left part of (b) is the generation results of sphere and the right part is cylinder.

This allows us to apply Equation 3 to the distribution to obtain Equation 9 to sample from the negated concept space.

$$\tilde{\mathbf{x}}^k = \tilde{\mathbf{x}}^{k-1} - \frac{\lambda}{2}\nabla_{\mathbf{x}}(\alpha E(x|c_1) - E(x|c_2)) + \omega^k,\ \omega^k \sim \mathcal{N}(0, \lambda) \tag{9}$$

We note that the combinations of conjunctions, disjunctions and negations can specify more complex operators such as implication.

Concept Inference In concept inference, we wish to infer the underlying concept through which a given input is generated. Given several example inputs of an underlying concept, we wish to combine the data to make an informed estimation of the underlying concept. Assuming each input is independent of each other, the overall likelihood of the inputs is equivalent to the product of likelihood of each input under a concept and thus is the conjunction of likelihood of each individual data point

$$p(x_1, x_2, \ldots, x_n|c_1) \propto e^{\sum_i E(x_i|c_1)}. \tag{10}$$

We can then obtain maximum a posteriori (MAP) estimates of concepts by minimizing the energy of the above expression.

3 Experiments

We perform empirical studies to answer the following questions: Can EBMs exhibit concept compositionality, such as concept negation, conjunction, and disjunction, in generating images? Can we take advantage of concept combinations to learn new concepts in a continual manner? Does explicit factor decomposition enable better generalization? Can we perform inference across multiple inputs?

3.1 Setup

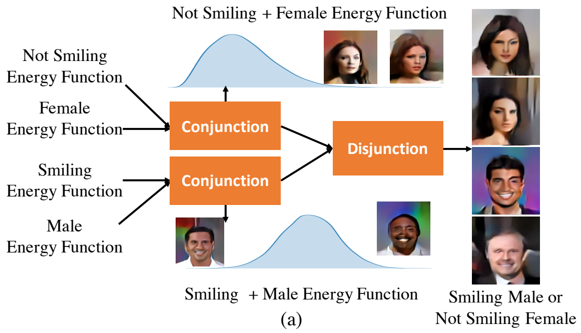
We perform experiments on 64x64 different object scenes rendered in Mujoco (Todorov et al., 2012) and the 128x128 CelebA dataset. For scenes rendered in Mujoco, we generate a central object of shape either sphere, cylinder, or box of varying size and height, with some number of (specified) additional background objects. Images are generated with varying lighting and camera positions.

We use the ImageNet32x32 architecture and ImageNet128x128 architecture from (Du & Mordatch, 2019) with the Swish activation (Ramachandran et al., 2017) on Mujoco and CelebA datasets. Models are trained on Mujoco datasets for up to 1 day on 1 GPU and for 1 day on 8 GPUs for CelebA.

3.2 Compositional Generation

We show EBMs are able to generate images that exhibit the versions of compositionality described in the methods section.

Concept Conjunction We find that in Figure 3 (a) that EBMs are able to combine independent concepts of age, gender, smile, and wavy hair with each additional attribute allowing more precise generation. Similarily, we find in Figure 3 (b) that EBMs are able to combine independent concepts of shape, position, size, and color together to generate more precise generations.


Not Smiling + Female Energy Function
Not Smiling Energy Function
Female Energy Function
Smiling Energy Function
Male Energy Function
Conjunction
Conjunction
Disjunction
Smiling + Male Energy Function
(a)
Smiling Male or Not Smiling Female

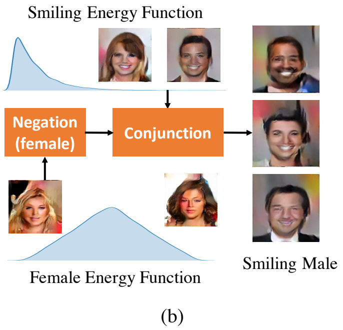


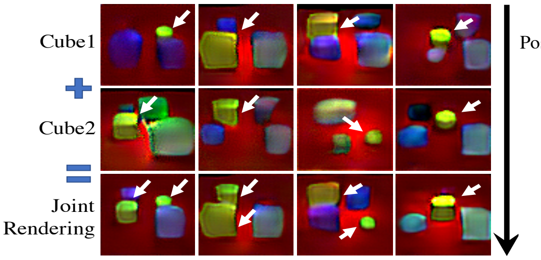
Figure 4: (a) Examples of concept disjunction on joint attributes (represented by conjunction of energy) of not smiling+female and smiling+male. EBMs are able to reliably support concept disjunction (generation of either one concept or the other) even when the concept itself is compound. (b) Examples of concept negation on the attributes of smiling female. When negating the female energy in combination with the smiling energy function, we are able to generate photos of males that are smiling.

Concept Disjunction We also find that EBMs are able to combine concepts additively (generate images that are concept A or concept B) as shown in Figure 4 (a). By constructing sampling using logsumexp, EBMs are able to either sample an image that is not smiling female or smiling male, where both not smiling female and smiling male are specified through the conjunction of energies of the two different concepts. This result also shows that concept disjunction can be chained on top of other operators such as concept conjunction.

Concept Negation We further generate concepts that are the opposite of the trained concept in Figure 4 (b), where we find that negating female, in combination with smiling leads to generation of a smiling male. Furthermore, we note that the ability of concept conjunction, disjunction, and negation allows us to flexibly specify any set of pairwise concepts.

Multiple Object Combination Finally, we explore the use of an EBM to model single object-based concepts. To investigate this, we constructed a dataset consisting of a central green cube with size and position annotations, in conjunction with large amount background clutter objects (which are not green), in which we trained a conditional EBM.

Despite the fact that the training dataset does not have any other green cubes, we find that adding two conditional EBMs conditioned on two different position and sizes, allowing us to selectively generate two different cubes in Figure 5. Furthermore, we find that such generation is able to satisfy the constraints of the dataset. For example, when two conditional cubes are too close, the conditionals EBMs are able to default to generating one cube.



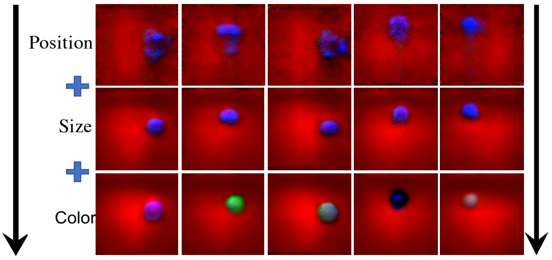
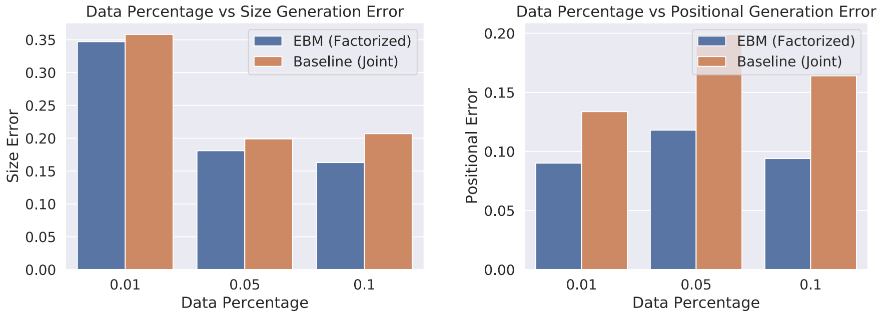


Figure 5: Object compositionality with EBMs. An EBM is trained to generate a green cube of specified size and shape in a scene with other obstacles. At test time, we sample from the conjunction of two different EBMs conditioned on different position/size attributes (shown in panels cube 1 and cube 2), which generates cubes at both locations. This generation further exhibits global coherence, by merging both cubes when they are too close (right-most column).

Figure 6: Examples of continual learning of concepts. A positional energy function is learned on set of cubes of one color. A shape energy function is then learned on a set of shapes of a fixed color. Finally, a color energy function learned different colored shapes. We able to continually learn and generate different color shapes at different positions, even though position is only learned on cubes of fixed color, and shape is only learned on shapes of fixed color.

3.3 Continual Learning

An important ability humans are endowed with is the ability to both continually learn new concepts, and to extrapolate existing concepts in combination with previously learned concepts. We evaluate



(a) (b)

Figure 7: Illustration of generation of size/position concepts as a function of data percentage. By learning a composable representation of underlying concepts, EBMs are able to extrapolate better with less data, and exhibit both lower size and positional error.

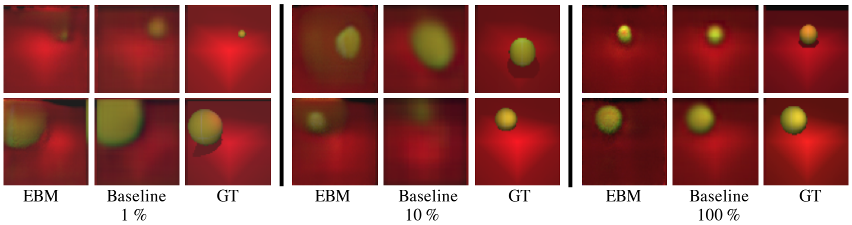


Figure 8: Illustration of generations from extrapolations of concepts of size and position. Models only see all possible sizes of spheres at (1%, 10%, 100% respectively) of the right most positions in images and are asked generated novel size/position images at remaining image locations.

to what extent compositionality in EBMs allow us to exhibit these properties through the continual learning protocol described below:

1. We first train an EBM for position based generation by training it on a dataset of cubes at various positions of a fixed color.
2. Next we train an EBM for shape based generation, by training the model in combination with the positional model to generate images on a dataset of different shapes (through summation) at different positions, but with the position based EBM fixed.
3. Finally we train an EBM for color based generation, by training the model in combination with both positional and shape models to generate images on a dataset of different shapes at different positions and colors (through summation). Again we fix both position and shape EBMs, and only train the color based generation.

We show in Figure 6 that this allow us to **extrapolate** our learned models for position and shape to generate different position shapes of various colors. The first column of Figure 6 shows the generations of a positional model, while the second column shows the generation of both positional and shape models, and the third column columns shows the generation of position, shape, and color models. Even though the positional model has only seen cubes of a particular color at a particular position, and color model shapes of a particular color, the third column illustrates that the composition of all three models is able to allow the generation of different colored shapes at various positions.

3.4 Cross Product Generalization

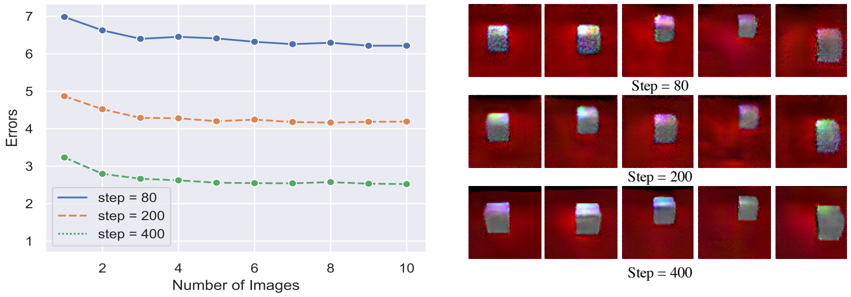
Humans are further endowed with the ability to extrapolate novel combinations of concepts when only a limited number of demonstrations of different concepts learned. For example, when finding a new toy in a shop, we can already anticipate changes in both the size and shape of the object.

We evaluate the extent to which EBMs, which allow us to factorize generation into different concepts, can help us extrapolate. To test this, we construct a sphere dataset consisting of sphere of all sizes at a specified percentage of the rightmost positions and large spheres remaining positions, with size/position annotations. At test time, we then evaluate the ability to construct spheres of various sizes at positions not seen at the train time. Such a setup requires a model to be able to extrapolate the learned position and size latents to generate these new images.

To train an EBM on this task, we first train an EBM conditioned on the position on with large sphere images at each position, as well as an EBM conditioned on size at each position where all different

Table 1: Position prediction error on different test datasets. "Test" has the same data distribution with training set. Other datasets change one or more environmental parameters, e.g. color, size, type, and light, which are unseen in the training set. "Avg" is the average error of "Color", "Light", "Size", and "Type". "Steps"indicates the number of negative sampling steps used to train the EBM. EBMs are able to generalize better. Larger number of negative sampling steps significantly decrease overall EBM error.

Model	Steps	Color	Light	Size	Type	Avg	Test
EBM	80	11.172	8.458	13.201	7.107	9.985	5.582
EBM	200	10.899	6.307	8.431	6.304	7.985	3.903
EBM	400	**4.084**	**4.033**	**6.853**	**3.694**	**4.666**	**2.917**
Resnet	-	20.002	5.881	10.378	6.310	10.643	3.635



(a)

(b)

Figure 9: The influence of multiple observations on EBMs. Multiple images are generated under different lighting conditions and camera heights. (a) The position prediction error decreases when the number of input images increases independent of negative training steps used to train models. (b) Examples of generated images with varying number of negative sampling steps. Large number of steps leads to more realistic images.

sizes of spheres are available. We then finetune the summation of the EBMs to generate objects in all positions/sizes in the training dataset. We compare with a baseline model trained on conditioning both latents together, optimized for MSE loss (with the same architecture/number parameters as the EBMs) when generating new combinations.

To evaluate the performance of generations, we train a discriminatory model to regress both the position and size of a generated sphere image. We plot histogram of differences in regressed size and position from a generation compared to conditioned latents in Figure 7. We find that EBMs are able to extrapolate both position and size better than a baseline joint model. We note that both models obtain less positional generation error at 1% data as opposed to 5% or 10% of data. This result is due the make-up of the data – with 1% data, only 1% of the rightmost sphere positions have different size annotations, so failed extrapolation causes models to generate large spheres at the conditioned position. Once there are more different size sphere annotations from data, models either collapse to an existing size or position, leading to a higher error. Qualitatively, Figure 8 illustrates that by learning a separate conceptual model for each latent, at a low percentage of data, the EBM is still able to combine concepts, while under full data, both models perform well.

3.5 Compositional Concept Inference

In this section, we show that EBMs are able to not only infer concepts from images but infer concepts in a compositional manner.

Concept Inference By minimizing the energy of a concept given an image, EBMs can be adapted to infer concept labels. To evaluate this, we generate a large dataset of cubes/spheres at various different locations in Mujoco, with random lighting small camera perturbations, and train an EBM to infer the position of cubes/spheres. We benchmark the inference ability of an EBM by computing the mean absolute error $(|\hat{x} - x| + |\hat{y} - y|)/2$, where (x, y) is the predicted object position and $(\hat{x}, \hat{y})$ is the ground truth position.

We compare EBMs with a baseline ResNet model (with the same architecture as the EBM) trained directly on position regression. Table 1 shows the comparisons of EBMs with different number of Langevin sampling steps and the ResNet model. We test the performance on several different datasets. "Color" refers to a test dataset where the object colors never been seen in the training dataset. "Light" means a test dataset using different light sources. "Size" means a test dataset where the object sizes are not covered in the training dataset and "Type" dataset consists cylinder images while the training images are spheres or cubes. EBMs with larger Langevin sampling steps outperform the ResNet benchmark, generalizing significantly better to distribution shift. Furthermore, larger numbers of Langevin sampling steps have better generation quality as shown in Figure 9 (b). This suggests that many figures in this paper (which are trained with 40 sampling steps) can likewise see a large boost in the generation quality.

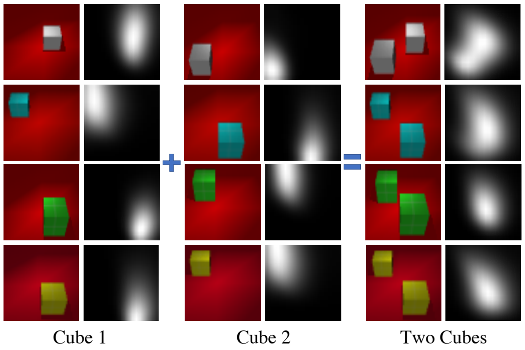

Cube 1 Cube 2 Two Cubes

Figure 10: Energy based model trained on single cubes and tested on two cubes. The RGB images and Grey images are the input images and energy map of energy at each position respectively. The energy maps of two cubes is the addition of energy map of two single cubes. The combined energy maps match the input images by highlighting two region or a union region.

Compositional Concept Inference We next examine the compositionality of inference in EBMs. We verify EBMs can effectively use the information from multiple observations by measuring the mean absolute error of position regression when given different views (from a variation of camera heights and lighting) of the predictions, using the MAP inference based on Equation 10. We display results in Figure 9 (a) and find that multiple observations reduce the position prediction errors.

Emergent Compositional Inference We also investigate the emergent compositional ability of EBMs by testing EBMs trained on position regression on single object images to scenes of two objects. We plot energy maps over possible positional labels in Figure 10 as well as individual energy maps over each cube. We find that EBMs have implicit compositional inference, with the joint positional energy map matching the summation of individual positional energy maps of each object.

4 Related Work

Our work draws on results in energy based models - see (LeCun et al., 2006) for a comprehensive review. A number of methods have been used for inference and sampling in EBMs, from Gibbs Sampling (Hinton et al., 2006), Langevin Dynamics (Du & Mordatch, 2019), Path Integral methods (Du et al., 2019) and learned samplers (Kim & Bengio, 2016). In this work, we show that MCMC sampling on EBMs through Langevin Dynamics can generate plausible natural images.

Compositionality has been incorporated in representation learning (see (Andreas, 2019) for a summary) and in generative modeling. One approach to compositionality has focused on learning disentangled factors of variation (Higgins et al., 2017; Kulkarni et al., 2015; Vedantam et al., 2018). Such an approach allows the combinatorial specification of outputs, but does not allow the addition of new factors. A different approach to compositionality includes learning various different pixel/segmentation masks for each concept (Greff et al., 2019; Gregor et al., 2015). However such a factorization may have difficulty capturing the global structure of an image, and in many cases different concepts can not be explicitly factored as attention masks.

In contrast, our approach towards compositionality focuses on composing separate learned probability distribution of concepts. Such an approach allows viewing factors of variation as constraints (Mnih & Hinton, 2005). (Hinton, 1999) shows that product of EBMs allows for conjunction of different concepts. In our work we illustrate additional logical compositions and corresponding performance on realistic datasets.

Our work is motivated by the goal of continual lifelong learning - see (Parisi et al., 2018) for a thorough review. Many methods are focused on how to overcome catashtophic forgetting (Kirkpatrick et al., 2017; Li & Hoiem, 2017), but do not support dynamically growing capacity. Progressive growing of the models (Rusu et al., 2016) has been considered, but is implemented at the level of the model architecture, whereas our method is agnostic to the models. Meta and few-shot learning (Reed et al., 2017; Bartunov & Vetrov, 2018) is another approach, but focuses on learning to model images rather than factors.

5 Conclusion

We have presented work demonstrating the potential of EBMs for both compositional generation and inference and hope to inspire future work in this direction.

References

Jacob Andreas. Measuring compositionality in representation learning. *arXiv preprint arXiv:1902.07181*, 2019. 8

Sergey Bartunov and Dmitry Vetrov. Few-shot generative modelling with generative matching networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 670–678, 2018. 8

Yilun Du and Igor Mordatch. Implicit generation and generalization in energy-based models. *arXiv preprint arXiv:1903.08689*, 2019. 2, 4, 8

Yilun Du, Toru Lin, and Igor Mordatch. Model based planning with energy based models. *CoRL*, 2019. 8

SM Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S Morcos, Marta Garnelo, Avraham Ruderman, Andrei A Rusu, Ivo Danihelka, Karol Gregor, et al. Neural scene representation and rendering. *Science*, 360(6394):1204–1210, 2018. 2

Jerry A Fodor and Ernest Lepore. *The compositionality papers*. Oxford University Press, 2002. 1

Klaus Greff, Raphaël Lopez Kaufmann, Rishab Kabra, Nick Watters, Chris Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference. *arXiv preprint arXiv:1903.00450*, 2019. 1, 8

Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*, 2015. 8

Irina Higgins, Loic Matthey, Arka Pal, Christopher P Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. Beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017. 8

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. *ICLR*, 2018. 1

Geoffrey E Hinton. Products of experts. *International Conference on Artificial Neural Networks*, 1999. 8

Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002. 1, 2, 3

Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Comput.*, 18(7):1527–1554, 2006. 8

Taesup Kim and Yoshua Bengio. Deep directed generative models with energy-based probability estimation. *arXiv preprint arXiv:1606.03439*, 2016. 8

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114 (13):3521–3526, 2017. 8

Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. Deep convolutional inverse graphics network. In *NIPS*, 2015. 8

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. 1

Yann LeCun, Sumit Chopra, and Raia Hadsell. A tutorial on energy-based learning. 2006. 8

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017. 8

Andriy Mnih and Geoffrey Hinton. Learning nonlinear constraints with contrastive backpropagation. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 2, pp. 1302–1307. IEEE, 2005. 8

German Ignacio Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *CoRR*, abs/1802.07569, 2018. URL http://arxiv.org/abs/1802.07569. 8

Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017. 4

Scott Reed, Yutian Chen, Thomas Paine, Aäron van den Oord, SM Eslami, Danilo Rezende, Oriol Vinyals, and Nando de Freitas. Few-shot autoregressive density estimation: Towards learning to learn distributions. *arXiv preprint arXiv:1710.10304*, 2017. 2, 8

Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016. 8

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017. 1

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033. IEEE, 2012. 4

Sjoerd van Steenkiste, Karol Kurach, and Sylvain Gelly. A case for object compositionality in deep generative models of images. *arXiv preprint arXiv:1810.10340*, 2018. 1

Ramakrishna Vedantam, Ian Fischer, Jonathan Huang, and Kevin Murphy. Generative models of visually grounded imagination. In *ICLR*, 2018. 1, 8