LayoutVAE: Stochastic Scene Layout Generation From a Label Set

Akash Abdu Jyothi\textsuperscript{1,3}, Thibaut Durand\textsuperscript{1,3}, Jiawei He\textsuperscript{1,3}, Leonid Sigal\textsuperscript{2,3}, Greg Mori\textsuperscript{1,3}
\textsuperscript{1}Simon Fraser University \textsuperscript{2}University of British Columbia \textsuperscript{3}Borealis AI
\{aabdujyo, tdurand, jha203\}@sfu.ca lsigal@cs.ubc.ca mori@cs.sfu.ca

Abstract

Recently there is an increasing interest in scene generation within the research community. However, models used for generating scene layouts from textual description largely ignore plausible visual variations within the structure dictated by the text. We propose LayoutVAE, a variational autoencoder based framework for generating stochastic scene layouts. LayoutVAE is a versatile modeling framework that allows for generating full image layouts given a label set, or per label layouts for an existing image given a new label. In addition, it is also capable of detecting unusual layouts, potentially providing a way to evaluate layout generation problem. Extensive experiments on MNIST-Layouts and challenging COCO 2017 Panoptic dataset verifies the effectiveness of our proposed framework.

1. Introduction

Scene generation, which usually consists of realistic generation of multiple objects under a semantic layout, remains one of the core frontiers of computer vision. Despite the rapid progress and recent successes in object generation (e.g., celebrity face, animals, etc.)\cite{1, 9, 13} and scene generation\cite{4, 11, 12, 19, 22, 30, 31}, little attention has been paid to frameworks designed for stochastic semantic layout generation. Having a robust model for layout generation will not only allow us to generate reliable scene layouts, but also provide priors and means to infer latent relationships between objects, advancing progress in the scene understanding domain.

A plausible semantic layout calls for reasonable \textit{spatial} and \textit{count} relationships (relationships between the number of instances of different labels) between objects in a scene\cite{5, 27}. For example, a person would either ride (\textit{on top of}) a horse, or stand \textit{next to} a horse, but seldom would he be \textit{under} a horse. Another example would be that the number of ties would very likely be smaller than or equal to the number of people in an image. The first example shows an instance of a plausible spatial relationship and the second shows a case of a generic count relationship. Such intrinsic relationships buried in high-dimensional visual data are usually learned implicitly by mapping the textual description to visual data. However, since the textual description can always be treated as an abstraction of the visual data, the process becomes a \textit{one-to-many} mapping. In other words, given the same text information as a condition, a good model should be able to generate multiple plausible images all of which satisfy the semantic description.

Previous works focused on a popular simplified instance of the problem described above: scene generation based on sentence description\cite{6, 11, 12, 19, 24, 25, 30}. A typical sentence description includes partial information on both the background and objects, along with details of the objects’ appearances and scene layout. These frameworks rely heavily on the extra relational information provided by the sentence. As a result, although these methods manage to generate realistic scenes, they tend to ignore learning the intrinsic relationships between the objects, prohibiting the wide adoption of such models where weaker descriptions are provided.

In this work, we consider a more sophisticated problem: scene generation based on a label set description. A label
set, as a much weaker description, only provides the set of labels present in the image (without any additional relationship description), requiring the model to learn spatial and count relationships from visual data.

Furthermore, the ambiguity of this type of label set supervision calls for diverse scene generation. For example, given the label set person, surfboard, sea, a corresponding scene could have multiple instances of each label (under plausible count relationships), positioned at various locations (under plausible spatial relationships). For instance in the COCO dataset [21], there are 869 images in the training set that have the label set person, sea and surfboard. Figure 1 shows examples of multiple plausible images with this label set.

We propose LayoutVAE, a stochastic model capable of generating scene layouts given a label set. The proposed framework can be easily embedded into existing scene generation models that take scene layout as input, such as [10, 31], providing them plausible and diverse layouts. Our main contributions are as follows.

- We propose a new model for stochastic scene layout generation given a label set input. Our model has two components, one to model the distributions of count relationships between objects and another to model the distributions of spatial relationships between objects.

- We propose a new synthetic dataset, MNIST-Layouts, that captures the stochastic nature of scene layout generation problem.

- We experimentally validate our model using MNIST-Layouts and the COCO [21] dataset which contains complex real world scene layouts. We analyze our model and show that it can be used to detect unlikely scene layouts.

2. Related Work

Sentence-conditioned image generation. A variety of models have proposed to generate an image given a sentence. Reed et al. [25] use a GAN [7] that is conditioned on a text encoding for generating images. Zhang et al. [30] propose a GAN based image generation framework where the image is progressively generated in two stages at increasing resolutions. Reed et al. [24] perform image generation with sentence input along with additional information in the form of keypoints or bounding boxes.

Hong et al. [11] break down the process of generating an image from a sentence into multiple steps. The input sentence is first used to predict the objects that are present in the scene, followed by prediction of bounding boxes, then semantic segmentation masks, and finally the image. While scene layout generation in this work predicts probability distributions for bounding box layout, it fails to model the stochasticity intrinsic in predicting each bounding box. Gupta et al. [8] use an approach similar to [11] to predict layouts for generating videos from scripts. Johnson et al. [12] uses the scene graph generated from the input sentence as input to the image generation model. Given a scene graph, their model can generate only one scene layout.

Deng et al. [6] propose PNP-Net, a VAE framework to generate image of an abstract scene from a text based program that fully describes it. While PNP-Net is a stochastic model for generation, it was tested on synthetic datasets with only a small number of classes. Furthermore, it tries to encode the entire image into a single latent code whereas, in LayoutVAE, we break down just the layout generation step into two stages with multiple steps in each stage. Based on these reasons, it is unclear whether PNP-Net can scale up to real world image datasets with a large number of classes. Tao et al. [29] propose a GAN based model with attention for sentence to image generation. The more recent work from Li et al. [19] follow a multi-stage approach similar to [11] to generate an image from a sentence, with the key novelty of using attention mechanisms to create more realistic objects in the image.

Layout generation in other contexts. Chang et al. [3] propose a method for 3D indoor scene generation based on text description by placing objects from a 3D object library, and is later improved in [2] by learning the grounding of more detailed text descriptions with 3D objects. Wang et al. [28] use a convolutional network that iteratively generates a 3D room scene by adding one object at a time. Qi et al. [23] propose a spatial And-Or graph to represent indoor scenes, from which new scenes can be sampled. Different from most other works, they use human affordances and activity information with respect to objects in the scene to model probable spatial layouts. Li et al. [18] propose a VAE based framework that encodes object and layout information of indoor 3D scenes in a latent code. During generation, the latent code is recursively decoded to obtain details of individual objects and their layout.

More recently, Li et al. [17] proposed LayoutGAN, a GAN based model that generates layouts of graphic elements (rectangles, triangles etc.). While this work is close to ours in terms of the problem focus, LayoutGAN generates label sets based on input noise, and it cannot generate layout for a given set of labels.

Placing objects in scenes. Lee et al. [16] propose a conditional GAN model for the problem of adding the segmentation mask of a new object to the semantic segmentation of an image. Lin et al. [20] address a similar problem of adding an RGB mask of an object into a background image.
Figure 2: Model. LayoutV AE is composed of two models: CountV AE which predicts the number of objects for each category and BBoxV AE which predicts the bounding box of each object. A graphical model is given in Appendix A of the supplementary.

3. Background

In this section, we first define the problem of scene layout generation from a label set, and then provide an overview of the base models that LayoutV AE is built upon.

3.1. Problem Setup

We are interested in modeling contextual relationships between objects in scenes, and furthermore generation of diverse yet plausible scene layouts given a label set as input. The problem can be formulated as follows.

Given a collection of $M$ object categories, we represent the label set corresponding to an image in the dataset as $L \subseteq \{1, 2, 3, ..., M\}$ which indicates the categories present in the image. Note that here we use the word “object” in its very general form: car, cat, person, sky and water are objects. For each label $k \in L$, let $n_k$ be the number of objects of that label in the image and $B_k = \{b_{k,1}, b_{k,2}, ..., b_{k,n_k}\}$ be the set of bounding boxes. $D_{k,i} = [x_{k,i}, y_{k,i}, w_{k,i}, h_{k,i}]$ represents the top-left coordinates, width and height of the $i$-th bounding box of category $k$. We train a generative model to predict diverse yet plausible sets of $\{B_k : k \in L\}$ given the label set $L$ as input.

3.2. Base Models

Variational Autoencoders. A variational autoencoder (VAE) [15] describes an instance of a family of generative models $p_\theta(x, z) = p_\theta(x|z)p_\theta(z)$ with a complex likelihood function $p_\theta(x|z)$ and an amortized inference network $q_\phi(z|x)$ to approximate the true posterior $p_\theta(z|x)$. Here $x$ represents observable data examples, $z$ the latent codes, $\theta$ the generative model parameters, and $\phi$ the inference network parameters. To prevent the latent variable $z$ from just copying $x$, we force $q_\phi(z|x)$ to be close to the prior distribution $p_\theta(z)$ using a KL-divergence term. Usually in VAE models, $p_\theta(z)$ is a fixed Gaussian $\mathcal{N}(0, I)$. Both the generative and the inference networks are realized as nonlinear neural networks. An evidence lower bound (ELBO) $\mathcal{L}(x; \theta, \phi)$ on the generative data likelihood $\log p(x)$ is used to jointly optimize $\theta$ and $\phi$:

$$\mathcal{L}(x; \theta, \phi) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - KL(q_\phi(z|x) || p_\theta(z))$$

(1)

Conditional VAEs. A conditional VAE (CVAE) [26] defines an extension of VAE that conditions on an auxiliary description $c$ of the data. The auxiliary conditional variable makes it possible to infer a conditional posterior $q_\phi(z|x, c)$ as well as perform generation $p_\theta(x|z, c)$ based on a given description $c$. The ELBO is thus updated as:

$$\mathcal{L}^{CVAE}(x, c; \theta, \phi) = \mathbb{E}_{q_\phi(z|x, c)}[\log p_\theta(x|z, c)] - KL(q_\phi(z|x, c) || p_\theta(z|c))$$

(2)

In CVAE models, the prior of the latent variables $z$ is modulated by the auxiliary input $c$.

4. LayoutV AE for Stochastic Scene Layout Generation

In this section, we present LayoutV AE and describe the scene layout generation process given a label set. As discussed in section 1, this task is challenging and solving it requires answering the following two questions: what is the number of objects for each category? and what is the location and size for each object? LayoutV AE is naturally decomposed into two models: one to predict the count for each given label, named CountV AE, and another to predict the location and size of each object, named BBoxV AE. The overall structure of the proposed LayoutV AE is shown in
The number of objects (count) for each label is first predicted by CountVAE, then BBoxVAE predicts the bounding box for each object. The two-step approach with stochastic models naturally allows LayoutVAE to generate diverse layouts. In addition, it provides the flexibility to handle various types of input as it allows us to use each module independently. For example, BBoxVAE can be used to generate a layout if counts are available, or add a single module independently. For example, BBoxVAE can be used to handle various types of input as it allows us to use each category that was predicted before and the counts

\[ \hat{n}_m \sim p_{\varphi}(\hat{n}_k | \hat{z}_k^c, c_k^c) \]

Finally, the count is sampled from the predicted Poisson count distribution:

\[ \hat{n}_k \sim p_{\varphi}(\hat{n}_k | \hat{z}_k^c, c_k^c) \]

This label count is further used in the conditioning variable for the future steps of CountVAE.

### Algorithm 1 CountVAE: Loss computation on the image

**Input:** Label set \( L \), instance count \( \{ n_k : k \in L \} \)

1. \( L^{\text{count}} = 0 \)
2. for \( k \in L \) do
3. Compute the conditioning input \( c_k^c \) for category \( k \) (Equation 3)
4. Compute the variational lower bound \( \mathcal{L}^c(n_k, c_k^c; \theta^c, \phi^c) \) (Equation 7)
5. \( \mathcal{L}^{\text{count}} = \mathcal{L}^{\text{count}} + \mathcal{L}^c(n_k, c_k^c; \theta^c, \phi^c) \)
6. end for

**Output:** \( \mathcal{L}^{\text{count}} / ||L|| \)

where \( \mu_{\varphi} \) (resp. \( \mu_{\theta} \)) and \( \sigma^2_{\varphi} \) (resp. \( \sigma^2_{\theta} \)) are functions that estimate the mean and the variance of the approximate posterior (resp. prior). Details of the model architecture are given in subsection B.1 of the supplementary.
4.2. BBoxVAE

Given the label set \( \mathbb{L} \) and the object count per category \( \{ n_m : m \in \mathbb{L} \} \), the BBoxVAE predicts the distribution of coordinates for the bounding boxes autoregressively. We follow the same predefined label order as CountVAE in the label space, and order the bounding boxes from left to right for each label. All bounding boxes for a given label are predicted before moving on to the next label. BBoxVAE is a conditional VAE that models the \( j \)-th bounding box \( \mathbf{b}_{k,j} \) for the label \( k \) given the label set \( \mathbb{L} \) along with the count for each label, current label \( k \) and coordinate and label information of all the bounding boxes that were predicted earlier. Previous predictions include all the bounding boxes for previous labels as well as the ones previously predicted for the current label: \( \mathbb{B}_{k,j}^{\text{prev}} = \{ \mathbf{b}_{m,:} : m < k \} \cup \{ \mathbf{b}_{k,i} : i < j \} \)

The conditioning input of the BBoxVAE is:

\[
e_b^{k,j} = \left( \{ n_m : m \in \mathbb{L} \}, k, \mathbb{B}_{k,j}^{\text{prev}} \right)
\]

We use the notation of exponent \( b \) to indicate that it is related to the BBoxVAE. We model bounding box coordinates and size information using a quadrivariate Gaussian distribution as shown in Equation 12. We omit the subscript \( k, j \) for all variables in the equation for brevity.

\[
p_{\theta^b}(x, y, w, h|\mathbf{z}^b, c^b) = \mathcal{N}(x, y, w, h|\mu^b(\mathbf{z}^b, c^b), \sigma^b(\mathbf{z}^b, c^b))
\]

where \( \theta^b \) represents the generative model parameters of BBoxVAE, \( \mathbf{z}^b \) denotes a sampled latent variable and \( \mu^b \) and \( \sigma^b \) are functions that estimate the mean and the variance respectively for the Gaussian distribution. \( x \) (resp. \( y \)) is the \( x \)- (resp. \( y \)-) coordinate of the top left corner and \( w \) (resp. \( h \)) is the width (resp. height) of the bounding box. These variables are normalized between 0 and 1 to be independent of the dimensions of the image. Details of the model architecture are given in subsection B.2 of the supplementary.

5. Experiments

We implemented LayoutVAE in PyTorch, and used Adam optimizer [14] for training the models. CountVAE was trained for 50 epochs at a learning rate of \( 10^{-5} \) with batch size 32. BBoxVAE was trained for 150 epochs at a learning rate of \( 10^{-4} \) with batch size 32. We chose the best model during training based on the validation loss, and evaluate that model’s performance on the test set. We present quantitative and qualitative results that showcase the usefulness of LayoutVAE model.

5.1. Datasets

MNIST-Layouts. We created this dataset by placing multiple MNIST digits on a \( 128 \times 128 \) canvas based on predefined rules detailed in subsection C.1 in the supplementary material. We use the digits \( \{ 1, 2, 3, 4 \} \) as the global set of labels, thus limiting the maximum number of labels per image to 4. The dataset consists of 5000 training images, and 1000 images each for validation and testing. Figure 3 shows some examples of scene layouts from the training set.
and takes as input the same conditioning information used in the VAE models. As in LayoutVAE, we have two separate sub-models – CountMLP for predicting count distribution given the corresponding conditioning information, and BBoxMLP for predicting bounding box distribution given the corresponding conditioning information.

**BBoxLSTM.** This model is analogous to the bounding box predictor used in Hong et al. [11]. It consists of an LSTM that takes in the label set embedding (as opposed to sentence embedding in [11]) at each step and predicts the label and bounding box distributions for bounding boxes in the image one by one. At each step, the LSTM output is first used to generate the label distribution. The bounding box distribution is then generated conditioned on the label.

**BLSTM.** We also use bidirectional LSTM in an analogous fashion as LayoutVAE. We have CountBLSTM and BBoxBLSTM to predict count distribution and bounding box distribution respectively, where we input the respective conditioning information at each step for the BLSTM. The conditioning information is similar as in the VAE models except that we do not give the pooled embedding of previous counts/bounding boxes information as we now have a bidirectional recurrent model. Note that having a bidirectional model can possibly alleviate the need to explicitly define the order in the label and the bounding box coordinates spaces.

**sg2im [12].** This model generates a scene layout based on a scene graph. We compare LayoutVAE with this model because label set (in this case, with multiple copies of labels as per the ground truth count of each label) can be seen as the simplest scene graph without any relations. We also explored another strategy where a scene graph is randomly generated for the label set, but we found that model performance worsened in that setting. For this experiment, we use the code and the pretrained model on COCO provided by the authors at https://github.com/google/sq2im to predict the bounding boxes.

Note that these models are limited in their ability to model complex distributions and generate diverse layouts, whereas LayoutVAE can do so using the stochastic latent variable.

### 5.2. Baseline Models

To our knowledge, the problem of generating scene layouts (with both thing and stuff categories, and multiple bounding boxes per category covering almost the entire image) from a label set is new and has not been studied. There is no existing model for this task so we adapt existing models that generate scene layout from more complex inputs e.g. sentence, scene graphs. Note that we choose the same distributions as in LayoutVAE for modeling the outputs (counts or bounding box information) to have a fair comparison between all the models.

**Autoregressive MLP (AR-MLP).** This model is analogous to our proposed VAE based model, except that it has a multi-layer perceptron instead of a VAE. It runs autoregressively and takes as input the same conditioning information used
5.3. Quantitative Evaluation

To compare the models, we compute average negative log-likelihood (NLL) of the test samples’ count or bounding box coordinates under the respective predicted distributions. We compute Monte Carlo estimate of NLL for the VAE models by drawing 1000 samples from the conditional prior at each step of generation. LayoutVAE and the baselines are trained and evaluated by teacher forcing i.e. we provide the ground truth context and count/bounding box history at each step of computation.

Comparison with baseline models. Table 1 (accuracy metrics) and Table 2 (likelihood) show count and bounding box generation performances for all the models. First, we observe that LayoutVAE model is significantly better than all the baselines. In particular the large improvement with respect to the autoregressive MLP baseline shows the relevance of stochastic model for this task. It is also interesting to note that autoregressive MLP model performs better than the recurrent models. Finally, we observe that sg2im model [12] is not able to predict accurate bounding boxes without a scene graph with relationships. LayoutVAE and the baselines show similar performance for count prediction in MNIST-Layouts because estimating count distribution over the 4 labels of MNIST-Layouts is much easier than in the real world data from COCO.

Ablation study. In Table 3, we analyze the importance of context representation and history in the conditioning information. We observe that both history and context are useful because they increase the log-likelihood but their effects vary for count and bounding box models. The context is more critical than the history for the CountVAE whereas the history is more critical than the context for the BBoxVAE.

### Table 3: Ablation study. Test NLL results for CountVAE and BBoxVAE by prior sampling on COCO dataset. The history is the previous counts for CountVAE and the previous bounding box information for BBoxVAE. The context is the label set for CountVAE and the label set with counts for BBoxVAE.

<table>
<thead>
<tr>
<th>History</th>
<th>Context</th>
<th>NLL CountVAE</th>
<th>NLL BBoxVAE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.592</td>
<td>4.17</td>
</tr>
<tr>
<td>✓</td>
<td></td>
<td>0.570</td>
<td>3.78</td>
</tr>
<tr>
<td>✓ ✓</td>
<td></td>
<td>0.581</td>
<td>2.93</td>
</tr>
<tr>
<td>✓ ✓</td>
<td>✓</td>
<td>0.562</td>
<td>2.72</td>
</tr>
</tbody>
</table>

Table 4: Analysis of label order. NLL for BBoxVAE for different label set order on COCO dataset.

<table>
<thead>
<tr>
<th>things before stuffs</th>
<th>stuffs before things</th>
<th>random</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.72</td>
<td>2.71</td>
<td>3.22</td>
</tr>
</tbody>
</table>

### Figure 4: Example where likelihood under LayoutVAE increases when flipped upside down. We can see that the flipped layout is equally plausible for this example.

### Figure 5: Image Generation from a label set. We show diverse realistic layouts generated by LayoutVAE for the input label set person, surfboard and sea. We generate images from the layouts by using the generative model provided by Zhao et al. [31]

Despite these different behaviours, we note that history and context are complementary and increase the performances for both CountVAE and BBoxVAE. This experiment validates the importance of both context and history in the conditioning information for both CountVAE and BBoxVAE.

Analysis of the label set order. We performed experiments by changing the order of labels in the label set. We consider three variants for these experiments — a fixed order with all the things before stuffs (default choice for all the other experiments), reverse order of the above with all the stuffs before things, and finally we randomly assigned the order of labels for each image. Table 4 shows the results of the experiment on BBoxVAE. We found that fixed order (things first or stuffs first) gave similar results, while randomizing the order of labels across images resulted in a significant reduction in performance.

Detecting unlikely layouts in COCO. To test the ability of the model to differentiate between plausible and unlikely layouts, we perform an experiment where we flip the image layouts in the test set upside down, and evaluate the NLL (in this case, by importance sampling using 1000 samples from the posterior) of both the normal layout and the flipped layout under BBoxVAE. We found that the NLL got worse for 92.58% of image layouts in the test set when flipped upside down. The average NLL for original layouts is 2.26 while that for the flipped layouts is 4.33. We note that there are instances where flipping does not necessarily lead to an
Figure 6: Stochastic layout generation for MNIST-Layouts. We show 5 sample layouts generated by LayoutVAE for each input set of labels. We fill in the generated bounding boxes with randomly sampled MNIST digits of appropriate labels. We can see that LayoutVAE generates bounding boxes following all the rules that we defined for MNIST-Layouts (e.g. large 2s in the middle, small 3s at the bottom etc. with the appropriate count values). The complete set of rules is given in subsection C.1 in the supplementary material.

Figure 7: Bounding box prediction using LayoutVAE. We show the steps of bounding box generation for a test set sample. At each step, we obtain diverse bounding box predictions for the input label (written in subcaptions) given the ground truth layout up to the previous step (shown in cyan bounding boxes). We show bounding box heatmap using 100 samples from the prior distribution (top row for each step, red means high probability), and bounding boxes using 20 samples from the prior distribution. The predicted layouts are overlaid on the test image for clarity.

anomalous layout, which explains why likelihood worsened only for 92.58% of the test set layouts. Figure 4 shows such an example where flipping led to an increased likelihood under the model. Additional results are shown in subsection C.3 in the supplementary material.

5.4. Qualitative Evaluation

Scene layout samples. We present qualitative results for diverse layout generation on COCO(Figure 5) datasets and MNIST-Layouts(Figure 6). The advantages of modeling scene layout generation using a probabilistic model is evident from these results: LayoutVAE learns the rules of object layout in the scene and is capable of generating diverse layouts with different counts of objects. More examples can be found in subsection C.4 in the supplementary material.

Per step bounding box prediction for COCO. Figure 7 shows steps of bounding box generation for a test set of labels. For each step, we show the probability map and 20 plausible bounding boxes. We observe that given some bounding boxes, the model is able to predict where the next object could be in the image. More examples for bounding box generation are provided in subsection C.5 in the supplementary material.

6. Conclusion

We propose LayoutVAE for generating stochastic scene layouts given a label set as input. It comprises of two components to model the distribution of spatial and count relationships between objects in a scene. We compared it with other existing methods or analogues thereof, and showed significant performance improvement. Qualitative visualizations show that LayoutVAE can learn intrinsic relationships between objects in real world scenes. Furthermore, LayoutVAE can be used to detect abnormal layouts.
References


[30] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks. In IEEE International Conference on Computer Vision (ICCV), 2017. 1, 2

A. Graphical Model

For each of CountVAE and BBoxVAE, we have a conditional VAE that is autoregressive (Figure 15), where the conditioning variable contains all the information required for each step of generation. By explicitly designing the conditioning variable this way, we forgo using past latent codes for generation.

![Graphical model for LayoutVAE](image)

Figure 15: Graphical model for LayoutVAE. \(c\) denotes context (label set for CountVAE, label set with counts for BBoxVAE) and \(c_k\) denotes the conditioning information for the CVAE. Dashed arrow denotes inference of the approximate posterior.

B. Model Architecture

B.1. CountVAE

We represent the label set \(L\) by a multi-label vector \(s \in \{0, 1\}^M\), where \(s_k = 1\) (resp. 0) means the \(k\)-th category is present (resp. absent). For each step of CountVAE, the current label \(k\) is represented by a one-hot vector denoted as \(l^k \in \{0, 1\}^M\). We represent the count information \(n_k\) of category \(k\) as a \(M\) dimensional one-hot vector \(y^k\) with the non-zero location filled with the count value. This representation of count captures its label information as well. We perform pooling over a set of previously predicted label counts by summing up the vectors. The conditioning input is

\[
c^c_k = FC\left([\text{MLP}(s), \text{MLP}(l^k), \text{MLP}\left(\sum_{m<k}y^m\right)]\right)
\]

where \([\cdot, \cdot]\) denotes concatenation, FC a fully connected layer and MLP a generic multi-layer perceptron. Figure 16 shows the architecture in detail.

B.2. BBoxVAE

We represent label and count information pair for each category \(\{n_m : m \in L\}\) as \(y^m\) using the same strategy as in CountVAE. Pooled representation of the label set along with counts is obtained by summing up these vectors to obtain a multi-label vector. For each step of BBoxVAE, the current label \(k\) is represented by a one-hot vector denoted as \(l^k\). We use LSTM for pooling previously predicted bounding boxes \(B^{prev}_{k,j}\). We represent each bounding box as a vector of size 4, and we concatenate \(M\) dimensional label vector to add label information to it. We pass \(M+4\) dimensional vectors of successive bounding boxes through an LSTM and use the final step output as the pooled representation. Figure 18 shows the pooling operation, and Figure 17 shows the detailed architecture for each module in BBoxVAE. The conditioning input is

\[
c^b_{k,j} = FC\left([\text{MLP}\left(\sum_{m \in L}y^m\right), \text{MLP}(l^k), \text{MLP}(B^{prev}_{k,j})]\right).
\]

BBoxVAE predicts the mean for the quadrivariate Gaussian (Equation 12), while covariance is assumed to be a diagonal matrix with each value of standard deviation equal to 0.02.
Figure 16: CountVAE Architecture.
Figure 17: BBoxVAE Architecture.
Figure 18: **LSTM pooling for bounding boxes.** We use an LSTM to pool the set of previously predicted bounding boxes to be used in the conditioning information for BBoxVAE.
C. Experiments

C.1. MNIST-Layouts dataset

Table 9 shows the rules used to generate the dataset. To generate the layouts, we adapted the code provided at https://github.com/aakhundov/tf-attend-infer-repeat.

<table>
<thead>
<tr>
<th>Label</th>
<th>Count</th>
<th>Location</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3,4</td>
<td>top</td>
<td>medium</td>
</tr>
<tr>
<td>2</td>
<td>2,3</td>
<td>middle</td>
<td>large</td>
</tr>
<tr>
<td>3</td>
<td>1,2</td>
<td>bottom</td>
<td>small-medium</td>
</tr>
<tr>
<td>4 (2 present)</td>
<td>count(2)+3,6</td>
<td>around a 2</td>
<td>small</td>
</tr>
<tr>
<td>4 (2 absent)</td>
<td>2</td>
<td>bottom-right</td>
<td>small</td>
</tr>
</tbody>
</table>

Table 9: Rules for generating MNIST-Layouts dataset. Given a label set, we use uniform distribution over the possible count values to generate count. We then sample over a uniform distribution over the location and size ranges (precise details skipped in the table for brevity) to generate bounding boxes for each label instance. When label 4 is present in the input label along with label 2 (4\textsuperscript{th} row in the table), we randomly choose an instance of 2 and place all the 4s around that.

C.2. Analysis of latent code size

In Table 10, we analyze the dimension of the latent space for both CountVAE and BBoxVAE. For each dimension we report the NLL performance on COCO dataset. We observe that both models have good performance when the latent code size is between 32 and 128, and the models are not sensitive to this hyperparameter. Increasing the latent space beyond 128 does not improve the performances.

<table>
<thead>
<tr>
<th>Latent Code Size</th>
<th>CountVAE</th>
<th>BBoxVAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.569</td>
<td>4.13</td>
</tr>
<tr>
<td>4</td>
<td>0.568</td>
<td>3.11</td>
</tr>
<tr>
<td>8</td>
<td>0.565</td>
<td>2.91</td>
</tr>
<tr>
<td>16</td>
<td>0.564</td>
<td>2.72</td>
</tr>
<tr>
<td>32</td>
<td>0.562</td>
<td>2.69</td>
</tr>
<tr>
<td>64</td>
<td>0.563</td>
<td>2.70</td>
</tr>
<tr>
<td>128</td>
<td>0.562</td>
<td>2.70</td>
</tr>
</tbody>
</table>

Table 10: Effect of latent code size. Average NLL over COCO test set for CountVAE and BBoxVAE while varying the size of the latent code.
C.3. Detecting unlikely layouts

In this section, we present more examples from our experiment on detecting unlikely layouts by flipping the original layout and computing likelihood under LayoutVAE. In Figure 19, we present some typical examples where likelihood under LayoutVAE (BBoxVAE, to be precise, since CountVAE gives the same result for original and flipped layouts as the label counts remain the same) decreases when flipped upside down. This behaviour was observed for 92.58% samples in the test set. In Figure 20, we present some examples of unusual layouts where likelihood under LayoutVAE increases when flipped upside down.

Figure 19: Some examples where likelihood under BBoxVAE decreases when flipped upside down. We show the test image, layout for the image and the flipped layout. Negative log likelihood (NLL) of the layout under BBoxVAE is shown along with each layout. We can see that the flipped layout is highly unlikely in these examples.

Figure 20: Some examples where likelihood under BBoxVAE increases when flipped upside down. We show the test image, layout for the image and the flipped layout. Negative log likelihood (NLL) of the layout under BBoxVAE is shown along with each layout. We can see that the flipped layout is equally or sometimes more plausible in these examples.
C.4. Examples for Layout Generation

Figure 21 shows examples of diverse layouts generated using LayoutVAE.

Figure 21: Layout generation using LayoutVAE. We show 5 randomly sampled layouts for each input label set.
C.5. Examples for Bounding Box Generation

Figure 22 presents examples that showcase the ability of LayoutVAE to use conditioning information to predict plausible bounding boxes. We present additional examples of stochastic bounding box generation for test samples that have labels *person*, *surfboard* and *sea* in Figure 23 and Figure 24.

Figure 22: **Importance of conditioning information.** We present two examples here both of which have *person* as their first label. We see that bounding boxes for the first label *person* (first column) are close to the center in the first example, whereas they are smaller and close to the left side of the scene in the second example. This is because BBoxVAE gets count of *persons* (2 for the first example, and 12 for the second examples) in its conditioning information, which prompts the model to predict bounding boxes appropriately.
Figure 23: Test examples with labels person, surfboard, sea. We show steps of bounding box generation for test set samples in the same manner as in Figure 7 from the main paper.
Figure 24: An example with more objects in the scene. We show steps of bounding box generation for test set samples in the same manner as in Figure 7 from the main paper.