L-Verse: Bidirectional Generation Between Image and Text

Taehoon Kim* Gwangmo Song Sihaeng Lee Sangyun Kim Yewon Seo
Soonyoung Lee Seung Hwan Kim Honglak Lee Kyunghoon Bae
LG AI Research

Abstract

Far beyond learning long-range interactions of natural language, transformers are becoming the de-facto standard for many vision tasks with their power and scalability. Especially with cross-modal tasks between image and text, vector quantized variational autoencoders (VQ-VAEs) are widely used to make a raw RGB image into a sequence of feature vectors. To better leverage the correlation between image and text, we propose L-Verse, a novel architecture consisting of feature-augmented variational autoencoder (AugVAE) and bidirectional auto-regressive transformer (BiART) for image-to-text and text-to-image generation. Our AugVAE shows the state-of-the-art reconstruction performance on ImageNet1K validation set, along with the robustness to unseen images in the wild. Unlike other models, BiART can distinguish between image (or text) as a conditional reference and a generation target. L-Verse can be directly used for image-to-text or text-to-image generation without any finetuning. In quantitative and qualitative experiments, L-Verse shows impressive results against previous methods in both image-to-text and text-to-image generation on MS-COCO Captions. We furthermore assess the scalability of L-Verse architecture on Conceptual Captions and present the initial result of bidirectional vision-language representation learning on general domain.

1. Introduction

Image-to-text and text-to-image generation and can be summarized as a task of learning cross-modal representations of image and text. Recent studies [7, 10, 11, 32] on vision-language tasks have highly improved the performance of each target task, in particular with various transformer architectures [3, 4, 9, 45]. Initially designed to understand natural language, the dot-product multi-head attention mechanism [45] effectively learns long-range interactions of sequential data. To leverage transformer architectures [45] also in vision domains, an input image is factorized into a sequence of latent feature vectors.

To encode an image into a sequence of latent feature vectors, vector quantized variational autoencoder (VQ-VAE) [44] can be used to learn a discrete latent representation with quantized embedding vectors from the visual codebook. VQ-VAE is a simple and powerful representation learning method to make image sequential and is widely used in conditional image generation tasks with auto-regressive pairs like RNNs [33, 44] or transformers [10–12, 32]. Improving the reconstruction quality of VQ-VAE is also an active area of research [12, 32, 33].

Combining an auto-regressive transformer [3] with a feature extractor like VQ-VAEs or other deep convolutional neural networks (CNNs) is becoming a popular approach for various vision-language tasks. However, training a model for unidirectional image-to-text [7] or text-to-image [10, 32] generation task still requires a large amount of data.

Figure 1. Examples of L-Verse on zero-shot text-to-image generation (256 × 256 pixels) on Conceptual Captions (top) and image-to-text generation on MS-COCO Captions (bottom). Trained in bidirectional manner, L-verse can both generate well-conditioned synthetic images and detailed captions without any finetuning.

*Correspondence to: taehoon.kim@lgresearch.ai
2. Related Work

Adapting transformer architectures \cite{Vaswani2017, JMLR:v18:17-558, Devlin2019} for various vision-language tasks has been an active research area in the recent years. Since an image is a matrix of RGB pixel values, it should be first factorized into a sequence of feature vectors. Recent auto-regressive transformer based generative models \cite{Ramesh2021, 10.1145/3319520.3394844, 10.1145/3307650.3317688} utilize different variants of VQ-VAE \cite{DBLP:journals/corr/McDermottBB17} to compress and reconstruct images. In this section, we introduce the main concept of VQ-VAE and its variants. We also explain how VQ-VAE or other CNN architectures are combined with auto-regressive transformers to solve image-to-text or text-to-image generation tasks.

2.1. Vector Quantized Variational Autoencoder

Vector quantized variational autoencoder, VQ-VAE \cite{DBLP:journals/corr/McDermottBB17}, is a set of an encoder $E$, a decoder $G$, and a visual codebook $Z$ for learning discrete representations of images. The CNN encoder $E$ factorizes the continuous representation of an image into a series of discrete vectors $z_q$, each selected from visual codebook $Z$. The CNN decoder $G$ is used to reconstruct any $z_q$ sampled from $Z$. Razavi \textit{et al.} \cite{Razavi2020} extend this approach to use hierarchical feature representation and apply exponential-moving-average (EMA) weight update to codebook $Z$. To better optimize the training of VQ-VAE, Ramesh \textit{et al.} \cite{Ramesh2021} use the gumbel-softmax relaxation \cite{JMLR:v18:17-558, Esser2021}. Esser \textit{et al.} \cite{10.1145/3307650.3317688} further improve the quality of image reconstruction with additional CNN discriminator, originated from generative adversarial network (GAN) \cite{Goodfellow2014}.

2.2. Image-to-Text Generation

As the \textit{dot-product multi-head attention} \cite{Vaswani2017} was initially designed for language tasks, transformers have achieved new state-of-the-art results in generating natural and detailed captions corresponding to an input image. Previous works \cite{Xu2015, Chen2016} utilize region features extracted using Faster R-CNN \cite{Ren2015} to generate captions for each image. While visual semantics of each region improves the quality, objects outside detection target classes (80 classes for MS-COCO Detection \cite{COCO}) get ignored.

2.3. Text-to-Image Generation

Generative adversarial networks (GANs) \cite{Goodfellow2014, Mirza2014, Xu2017, 10.1145/3307650.3313385} have been traditionally used for text-conditional image generation tasks. GAN based models focus on finding better modeling assumptions for specific data domains like CUB-200 \cite{Wah2011} or MS-COCO Captions \cite{COCO}. Ramesh \textit{et al.} \cite{Ramesh2021} first trained a 12-billion parameter transformer \cite{10.1145/3319520.3394844} on 250-million image-text pairs for text-to-image generation in the general domain. Ding \textit{et al.} \cite{10.1145/3307650.3317688} proposed a 4-billion parameter transformer, CogView, with stable training techniques and finetuning strategies for various downstream tasks.
3. Method

3.1. Preliminary

Ramesh et al. [32] proposed a two-stage training procedure for text-to-image generation with an auto-regressive transformer [3]:

- **Stage 1**: Train a discrete variational autoencoder (dVAE) [32] to compress each 256 $\times$ 256 RGB into a 32 $\times$ 32 grid of image tokens with each element of 8192 ($d_z$) possible values.

- **Stage 2**: Concatenate up to 256 BPE-encoded text tokens with the 32 $\times$ 32 = 1024 image tokens, and train an auto-regressive transformer [3] to model the joint distribution over text and image tokens.

The approach maximizes the evidence lower bound [20,36] on the joint likelihood of the model distribution over the image $x$, the caption $y$, and the tokens $z$ for the encoded RGB image. From the factorization $p_{\theta,\psi}(x,y,z) = p_\theta(x|y,z)p_\psi(y,z)$, the lower bound is yielded as

$$\ln p_{\theta,\psi}(x,y) \geq \mathbb{E}_{z \sim q_{\phi}(z|x)} \left( \ln p_\theta(x|y,z) - D_{KL}(q_{\phi}(y,z|x), p_\psi(y,z)) \right)$$

(1)

where:

- $q_\phi$ denotes the distribution over the 32 $\times$ 32 encoded tokens generated by dVAE encoder from the image $x$.

- $p_\theta$ denotes the distribution over the reconstructed image $\hat{x}$ from dVAE decoder.

- $p_\psi$ denotes the joint distribution over the text and image tokens modeled by the transformer.

In Stage 1, dVAE (or other VQ-VAE variants) learns to minimize the reconstruction loss between $x$ and $\hat{x}$. In Stage 2, an auto-regressive transformer optimizes two negative log-likelihood (NLL) losses: (i) for caption $y$ and (ii) for encoded image tokens $z$.

3.2. Proposed Approach: L-Verse Framework

Inspired by DALL-E [32], we propose two major improvements for high-fidelity image reconstruction and bidirectional image-text generation:

- We improve the diversity of a visual codebook $Z$ with cross-level feature augmentation. We first train multi-level (hierarchical) VQ-VAE (blue in Figure 2) and apply weight-sharing to vector quantizers [33,44] in each feature-level. The hierarchical VQ-VAE is then fine-tuned to a VQ-VAE with codebook size $N = 32 \times 32$.

- We use segment embedding to indicate whether each token is given as a conditional reference ([REF]) or a generation target ([GEN]). For example, [REF] is added to each text token and [GEN] is added to each image token for text-to-image generation.
Figure 3. Comparison of input images (top), reconstructions from multi-level (hierarchical) feature-augmented variational autoencoder (AugVAE-ML) (middle), and reconstructions from single-level feature-augmented variational autoencoder (AugVAE-SL) (bottom) on Imagenet1K validation set. The resolution of each image is 256 × 256 pixels.

Following subsections describe the training and sampling procedure of L-Verse in detail. The overview of L-Verse framework with actual reconstruction and generation examples are shown in Figure 2.

3.3. Feature-Augmented Variational Autoencoder

Razavi et al. [33] states that increasing the number of latent feature map adds extra details to the reconstruction. However, increasing the number of latent map also increases the total codebook size $N$, from $32 \times 32 = 1024$ [44] to $32 \times 32 + 64 \times 64 = 5120$ [33].

For the high-quality image reconstruction at low-cost, we choose to use the single $32 \times 32$ latent map and augment the visual codebook $Z$ instead. From the example in Figure 4, similar patterns in various patch sizes can appear both in one image (blue) and across different images (red). As the distance between similar patterns gets closer after vector quantization (VQ) [44], extracting patches from different latent maps and storing them in one place removes duplicates and fills the codebook with unique 8192 ($d_z$) possible values.

We optimize the encoder - vector quantizer - decoder architecture of VQ-VAE [44] for cross-level feature augmentation:

- We define the encoder as $z = E(x, f, d_{out})$, where $x$ is an $n \times n \times d_{in}$ tensor and $f$ is a downsampling factor. $E(f, d_{out})$ downsamples a tensor $x$ into an $\frac{n}{f} \times \frac{n}{f} \times d_{out}$ tensor $z$.

- We define the decoder as $\hat{x} = G(\hat{z}, f, d_{out})$, where $\hat{z}$ is an $n \times n \times d_{in}$ tensor and $f$ is a upsampling factor. $G(f, d_{out})$ upsamples an $n \times n \times d_{in}$ tensor $\hat{z}$ into an $nf \times nf \times d_{out}$ tensor $\hat{x}$.

Hierarchical AugVAE (AugVAE-ML) consists of one $E(4, 256)$ and three $E(2, 256)$, four $VQ(8192)$ with shared weights, and three $G(2, 256)$ and one $G(4, 3)$. As shown in Figure 2 with blue dotted and connected lines, $E(4, 256)$ first downsamples a $256 \times 256 \times 3$ RGB image into a $64 \times 64 \times 256$ latent feature tensor. Each $E(2, 256)$ downsamples the previous tensor by 2. In total, four latent feature tensors ($64 \times 64 \times 256$, $32 \times 32 \times 256$, $16 \times 16 \times 256$, and $8 \times 8 \times 256$) are extracted. These four tensors are quantized with a $VQ(8192)$ for each latent map. During the training of AugVAE-ML, each codebook with 8192 values gains diversity via weight sharing. Each $G(2, 256)$ upsamples the concatenation of previous tensor (if exists) and $\hat{z}$ of each level by 2. $G(4, 3)$ reconstructs the original input from the last latent tensor and the quantized vector.

To reduce the overall codebook size $N$, we finetune the AugVAE-ML into a single-level AugVAE (AugVAE-SL) of $32 \times 32$ latent map. We remove encoders and decoders with $16 \times 16$ and $8 \times 8$ latent map and replace the concatenation before each decoder with a $1 \times 1$ convolution to expand the last channel of previous latent tensor by 2. This modification to AugVAE-ML effectively stabilizes the finetuning process. The final architecture of AugVAE is depicted in Figure 2 with blue connected lines. As shown in Figure 3, AugVAEs can compress and reconstruct images with high-fidelity. Implementation details of AugVAE architecture and training hyperparameters are provided in Appendix A.
3.4. Bidirectional Auto-Regressive Transformer

With the masked dot-product multi-head attention, the conventional auto-regressive transformer [3] can only understand a given sequence from left to right. Bidirectional generation between text and image doesn’t require a transformer to be fully-bidirectional: learning how to distinguish an image $\rightarrow$ text sequence and a text $\rightarrow$ image sequence is enough.

We just tell our bidirectional auto-regressive transformer (BiART) whether the given text (or image) is a conditional reference ([REF]) or a generation target ([GEN]). We feed BiART with an extra sequence of segment indexes for each token. A learnable embedding vector is assigned to each segment index ([REF]) and ([GEN]) and added to the input sequence. This simple idea enables the training and sampling of bidirectional image-text generation with BiART.

For training, we feed the input sequence in text $\rightarrow$ image or image $\rightarrow$ text order alternately for each iteration. In each iteration, BiART optimizes two negative log-likelihood (NLL) losses: (i) for the conditional reference $x$ indexed as [REF] and (ii) for the generation target $x$ indexed as [GEN]. When converges, BiART performs image-to-text (dotted red line in Figure 2) and text-to-image (connected red line in Figure 2) generations without any finetuning.

3.5. Training Details

Architecture Overview We first train 100-million parameter AugVAE-SL on ImageNet1K [8]. From results in Figure 3, 5 and Table 1, our AugVAE-SL shows impressive reconstruction results with both in-domain and out-of-domain images. We use ImageNet1K-trained AugVAE-SL as encoder and decoder of L-Verse and pair encoded tokens with corresponding text tokens. BiART in L-Verse is 500-million parameter GPT [3] transformer. While DALL-E [32] andCogView [10] use a sparse-transformer [4] with custom attention masks for fast training and sampling, we use a GPT-style [3] full-transformer to model the bidirectional cross-modal representation between image and text. We use 64 BPE-encoded [38] text tokens with 49808 possibilities and 1024 encoded image tokens with 8192 possibilities. More details are provided in Appendix B.

Mixed Precision Training To save computational cost and sampling time, BiART is trained with FP16 (O2) mixed-precision training without inefficient stabilization methods like PB-relaxation [10] or Sandwich-LayerNorm [10]. These techniques are designed to eliminate the overflow in forward pass, but computationally inefficient. We instead inference AugVAE in FP32 to prevent the underflow caused by the vector quantizer.

3.6. Sampling Details

Image Sampling Similar to Ramesh et al. [32], we rerank samples drawn from BiART using a pretrained contrastive model, CLIP [31]. CLIP assigns a score (clip-score) based on how well the image and text match with each other. For text-to-image generation, we make 64 samples from trained L-Verse model and calculate the clip-score to select a Top 1 image. We repeat this process $k$ times with different random seeds to sample $k$ images in total.

Text Sampling Our L-Verse auto-regressively generates a sequence of tokens. To generate an RGB image, 1024 $(32 \times 32)$ tokens should be generated one-by-one. However, the length of text may vary depending on its reference image. For this reason, generating full 64 tokens doesn’t always guarantee the quality of sampled text. In worst case, the result caption can be just a repeated sequence of same sentence and [PAD] tokens. From the statistics of MS-COCO Captions [24], each caption contains average 16 words. We first sample 32 text tokens for each reference image and split the result caption by the full stop (.) token. We only use the first split to calculate the clip-score for reranking. This process dramatically saves computation time to generate 64 samples and select Top 1.

From machine evaluation metrics in Table 2, truncated captions from 32 tokens achieve new state-of-the-art in all metrics except CIDEr [46] among the peers trained only on MS-COCO Captions. L-Verse also shows comparable performance to OSCAR [23], which is pretrained on 6.5-million image-text pairs. While full 64 token captions score 181.6 in CIDEr and 28.9 in SPICE [1], we figured out that scores are high just because each caption has more meaningful words. In our inner-group examination between full and truncated captions, we have agreed that each truncated version is more concise and accurate. We further investigate the quality of L-Verse generated captions with human evaluation, in comparison with human labeled ground-truths.
Figure 5. Qualitative evaluation on the reconstruction performance of different VQVAEs with unseen image domains. For all settings, we use ImageNet1K trained models without any finetuning. Images are resized to 256 × 256 with LANCZOS [6] filter. Cross-level feature augmentation allows AugV AE-SL to express out-of-domain unseen images in high-fidelity. Please zoom in for the detailed comparison.

<table>
<thead>
<tr>
<th>Model</th>
<th>Codebook Size N</th>
<th>dZ</th>
<th>FID</th>
</tr>
</thead>
<tbody>
<tr>
<td>DALL-E [32]</td>
<td>32 × 32</td>
<td>8192</td>
<td>32.01</td>
</tr>
<tr>
<td>VQGAN [12]</td>
<td>16 × 16</td>
<td>1024</td>
<td>7.94</td>
</tr>
<tr>
<td>VQGAN [12]</td>
<td>16 × 16</td>
<td>16384</td>
<td>4.98</td>
</tr>
<tr>
<td>AugV AE-SL</td>
<td>32 × 32</td>
<td>8192</td>
<td>3.28</td>
</tr>
<tr>
<td>VQVAE-2 [33]</td>
<td>64 × 64 &amp; 32 × 32</td>
<td>512</td>
<td>~10</td>
</tr>
<tr>
<td>VQGAN [12]</td>
<td>64 × 64 &amp; 32 × 32</td>
<td>512</td>
<td>1.45</td>
</tr>
<tr>
<td>AugV AE-ML</td>
<td>64 × 64 ~ 8 × 8</td>
<td>8192</td>
<td>1.04</td>
</tr>
</tbody>
</table>


4. Experiments

In this section, we demonstrate the performance of proposed L-Verse in every aspect with both quantitative and qualitative experiments. We mainly discuss reconstruction performance on ImageNet1K [8] and out-of-domain unseen images, image-to-text generation (image captioning) results on MS-COCO Captions [24], text-to-image generation results on MS-COCO Captions. For MS-COCO, we trained L-Verse on MS-COCO Captions 2014 Karpathy splits for fair evaluation with previous methods. We also include results of L-Verse trained on Conceptual Captions [39] to further discuss the scalability of L-Verse architecture for zero-shot text-to-image generation. The FID can change depending on calculation tools. For fair comparison, we compute the Reconstruction FID with torch-fidelity [29], caption evaluation metrics in Table 2 with nlg-eval [40], and FIDs in Table 3 with the DM-GAN code [53], available at https://github.com/MinfengZhu/DM-GAN.

4.1. Image Reconstruction

As Esser et al. [12] stated, the reconstruction Fréchet Inception Distance (FID) [16] of a VQ-VAE provide a lower bound on the achievable FID of the generative model trained on it. From the results on ImageNet1K validation set in Table 1, our AugV AE-ML trained with novel cross-level feature augmentation achieves FID of 1.04, meaning AugV AE-ML can compress and reconstruct image without nearly any information loss. Reconstruction examples on Figure 3 also demonstrates AugV AE-ML’s qualitative performance. Finetuned from AugV AE-ML, our AugV AE-SL also achieves new state-of-the-art FID of 3.28 among its single-level peers.

In a more difficult setting, we evaluate AugV AE-SL on reconstructing out-of-domain unseen images. From the examples in Figure 5, AugV AE-SL trained on ImageNet1K shows impressive reconstruction fidelity for all validation input images without extra finetuning. From this result, we believe that our AugV AE-SL can work as a new “imagenet-backbone” for various vision tasks. Detailed examination with more examples for each dataset in Figure 5 can be found in Appendix C.
<table>
<thead>
<tr>
<th>Model</th>
<th>B-4</th>
<th>M</th>
<th>R</th>
<th>C</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCST [35]</td>
<td>34.2</td>
<td>26.7</td>
<td>55.7</td>
<td>114.0</td>
<td>-</td>
</tr>
<tr>
<td>Up-Down [2]</td>
<td>36.3</td>
<td>27.7</td>
<td>56.9</td>
<td>120.1</td>
<td>21.4</td>
</tr>
<tr>
<td>RFNet [41]</td>
<td>36.5</td>
<td>27.7</td>
<td>57.3</td>
<td>121.9</td>
<td>21.2</td>
</tr>
<tr>
<td>Up-Down+HIP [50]</td>
<td>38.2</td>
<td>28.4</td>
<td>58.3</td>
<td>127.2</td>
<td>21.9</td>
</tr>
<tr>
<td>GCN-LSTM [28]</td>
<td>38.2</td>
<td>28.5</td>
<td>58.3</td>
<td>127.6</td>
<td>22.0</td>
</tr>
<tr>
<td>SGAE [49]</td>
<td>38.4</td>
<td>28.4</td>
<td>58.6</td>
<td>127.8</td>
<td>22.1</td>
</tr>
<tr>
<td>ORT [15]</td>
<td>38.6</td>
<td>28.7</td>
<td>58.4</td>
<td>128.3</td>
<td>22.6</td>
</tr>
<tr>
<td>AOANet [17]</td>
<td>38.9</td>
<td>29.2</td>
<td>58.8</td>
<td>129.8</td>
<td>22.4</td>
</tr>
<tr>
<td>M² Transformer [7]</td>
<td>39.1</td>
<td>29.2</td>
<td>58.6</td>
<td>131.7</td>
<td>22.6</td>
</tr>
<tr>
<td>L-verse</td>
<td>39.9</td>
<td>31.4</td>
<td>60.4</td>
<td>102.2</td>
<td>23.3</td>
</tr>
<tr>
<td>L-verse (OSCAR)</td>
<td>40.5</td>
<td>29.7</td>
<td>-</td>
<td>137.6</td>
<td>22.8</td>
</tr>
<tr>
<td>L-verse (OSCAR)</td>
<td>41.7</td>
<td>30.6</td>
<td>-</td>
<td>140.0</td>
<td>24.5</td>
</tr>
</tbody>
</table>

| Table 2. Comparison with state-of-the-arts on MS-COCO Captions Karpathy test split. We mainly compare results with models trained only on MS-COCO. Results from OSCAR (which requires additional fine-tuning) is given as a reference. |

4.2. Image-to-Text Generation

We evaluate the image-to-text generation (image captioning) performance of L-Verse with (i) machine evaluation metrics against previous MS-COCO trained state-of-the-arts and (ii) human evaluation against corresponding ground-truth (reference) captions.

**Machine Evaluation** We first compare the performance of our model with MS-COCO trained image captioning models in Table 2. We also include OSCAR [23], which is finetuned from a pretrained model with 6.5-million image-text pairs, to assess the scalability of our model with larger dataset. With proposed sampling method in Section 3.6, L-Verse surpasses all the other methods in terms of BLEU-4, METEOR, ROUGE, and SPICE without any object detection framework or other extra information. L-Verse also shows comparable performance to OSCAR, showing that pretraining L-Verse on a larger set of image-text pairs is a promising direction for future work.

**Human Evaluation** Without caption truncation, L-Verse achieves the highest score in CIDEr and SPICE. As we stated in Section 3.6, machine evaluation metrics don’t always guarantee the qualitative performance of generated captions. We further conduct a human evaluation similar to the one used in Li et al. [22]. We directly evaluate L-Verse generated captions with human-labeled ground-truth captions, which is the theoretical upper-bound of L-Verse in image-to-text generation. We randomly sample 500 sets of images, corresponding ground-truth caption (GT), and L-Verse generated caption (Pred) from MS-COCO 2014 mini-val split for the evaluation pool. 150 anonymous people participated for the evaluation. For each participant, we show randomly sampled 50 sets of image, GT, and Pred from the pool and ask to choose the best caption for each set. To cope with tie situation, we also allow each participant to choose “Both captions well describe the image”. We provide more details on human evaluation in Appendix D. Results in Figure 7 show that L-Verse can generate a detailed explanation of a given image, receiving 30.4% of votes (Pred + Both) in total. Examples in Figure 6 also demonstrate that L-Verse doesn’t miss the detail of each image.
Table 3. Fréchet Inception Distance (FID) on a subset of 30,000 captions sampled from MS-COCO Captions validation set. We mainly compare results with models trained only on MS-COCO. In the bottom part of the table, we provide results from DALL-E, Cogview, and L-Verse-CC (which are trained from much larger datasets) as references.

<table>
<thead>
<tr>
<th>Model</th>
<th>FID-0</th>
<th>FID-1</th>
<th>FID-2</th>
<th>FID-4</th>
<th>FID-8</th>
</tr>
</thead>
<tbody>
<tr>
<td>AttnGAN [48]</td>
<td>35.2</td>
<td>44.0</td>
<td>72.0</td>
<td>108.0</td>
<td>100.0</td>
</tr>
<tr>
<td>DM-GAN [53]</td>
<td>26.0</td>
<td>39.0</td>
<td>73.0</td>
<td>119.0</td>
<td>112.3</td>
</tr>
<tr>
<td>DF-GAN [43]</td>
<td>26.0</td>
<td>33.8</td>
<td>55.9</td>
<td>91.0</td>
<td>97.0</td>
</tr>
<tr>
<td>L-Verse</td>
<td>45.8</td>
<td>41.9</td>
<td>35.5</td>
<td>30.2</td>
<td>29.8</td>
</tr>
<tr>
<td>L-Verse-CC</td>
<td>37.2</td>
<td>31.6</td>
<td>25.7</td>
<td>21.4</td>
<td>21.1</td>
</tr>
<tr>
<td>†DALL-E [32]</td>
<td>27.5</td>
<td>28.0</td>
<td>45.5</td>
<td>83.5</td>
<td>85.0</td>
</tr>
<tr>
<td>†CogView [10]</td>
<td>27.1</td>
<td>19.4</td>
<td>13.9</td>
<td>19.4</td>
<td>23.6</td>
</tr>
</tbody>
</table>

* FID-k: FID of images blurred by radius k Gaussian filter.
* L-Verse trained on Conceptual Captions.
† Models trained on over 30 million image-text pairs.

4.3. Text-to-Image Generation

Following Ramesh et al. [32] and Ding et al. [10], we evaluate the text-to-image generation performance of L-Verse by comparing it to prior approaches. We compute FIDs in Table 3 after applying a Gaussian filter with varying radii to both validation images and samples from L-Verse. We use the image sampling process explained in Section 3.6. Generated samples with corresponding captions from MS-COCO are provided in Appendix E.

According to Ramesh et al. [32], training a transformer on tokens from a VQ-VAE encoder disadvantages model since it generates an image in low-frequency domain. Trained on same MS-COCO training set, L-Verse achieves best FID among previous approaches by a large margin with a slight blur of radius 2. The gap tends to increase as the blur radius is increased. We also compare L-Verse-CC, L-Verse trained on Conceptual Captions [39], with DALL-E [32] and CogView [10]. Considering the size of training data, L-Verse shows comparable text-to-image generation performance to other large-scale transformers as blur radius increases.

It is interesting that L-Verse shows decreasing FID with increasing blur radius, while other models show increasing FID. We hypothesize that L-Verse focuses on objects in the reference text, showing lower FIDs when high-frequency details are lost. This finding also corresponds with image-to-text generation results in Section 4.2. We also provide initial zero-shot text-to-image generation results with L-Verse in Figure 8. Trained on Conceptual Captions [39], L-Verse generates detailed images with objects in reference texts. We believe that L-Verse will also be able to generate realistic images in zero-shot fashion when trained with sufficient data and scale.

5. Conclusion

This paper presents L-Verse, a novel framework for bidirectional generation between image and text. Our feature-augmented variational autoencoder (AugVAE) achieves new state-of-the-art reconstruction FID and shows its potential as an universal backbone encoder-decoder for generative models. We also enable bidirectional training of auto-regressive transformer with segment embedding. Proposed bidirectional auto-regressive transformer (BiART) learns both image-to-text and text-to-image as a whole. Experimental results demonstrate that our L-Verse framework shows remarkable performance in both image-to-text and text-to-image generation.

Acknowledgments

We would first like to thank Yountae Jung and LG AI Research AI Platform Sector for providing technical support and computing resources for training and evaluation. We would also like to thank Jinsu Yoo, Daechul Ahn, Janghyeon Lee, Yeonsik Jo, and other members of LG AI Research Vision Lab for helpful discussions and feedback on the paper. Finally, we would like to thank all members of LG AI Research, without whom this work would not have been possible.
References

pcial: Semantic propositional image cap- 
ation evaluation. In Proceedings of the European Conference 
7

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien 
Bottom-up and top-down attention for image captioning 
and visual question answering. In Proceedings of the IEEE/CVF 
Conference on Computer Vision and Pattern Recognition, 
2018. 7

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Sub- 
biah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakan- 
tan, Pranav Shyam, Girish Sastry, Amanda Askell, Sand- 
hini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom 
Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, 
Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric 
Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack 
Clark, Christopher Berner, Sam McCandlish, Alec Radford, 
Ilya Sutskever, and Dario Amodei. Language models are 
few-shot learners. In Advances in Neural Information Pro- 
cessing Systems, 2020. 1, 2, 3, 5, 12

Generating long sequences with sparse transformers, 2019. 
1, 2, 5

Stargan v2: Diverse image synthesis for multiple domains. 
In Proceedings of the IEEE/CVF Conference on Computer 
Vision and Pattern Recognition, 2020. 12

[6] M. A. Clark, Chulwoo Jung, and Christoph Lehner. Multi- 
grid lanczos. EPJ Web of Conferences, 175, 2018. 6, 12

[7] Marcella Cornia, Matteo Stefaniini, Lorenzo Baraldi, and 
Rita Cucchiara. Meshed-memory transformer for image cap- 
tioning. In Proceedings of the IEEE/CVF Conference on 
Computer Vision and Pattern Recognition, 2020. 1, 2, 7

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei- 
Fei. ImageNet: A Large-Scale Hierarchical Image Database. 
In Proceedings of the IEEE/CVF Conference on Computer 
Vision and Pattern Recognition, 2009. 2, 5, 6, 12

Toutanova. BERT: Pre-training of deep bidirectional trans- 
formers for language understanding. In Proceedings of the 
Conference of the North American Chapter of the Associa- 
tion for Computational Linguistics: Human Language Tech- 
nologies, 2019. 1, 2

[10] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zhang, 
Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, 
Hongxia Yang, and Jie Tang. Cogview: Mastering text-to- 
image generation via transformers. In Advances in Neural 
Information Processing Systems, 2021. 2, 5, 8, 13

[11] Patrick Esser, Robin Rombach, Andreas Blattmann, and 
Björn Ommer. Imagebert: Bidirectional context with mul- 
tinomial diffusion for autoregressive image synthesis. In 
Advances in Neural Information Processing Systems, 2021. 1

[12] Patrick Esser, Robin Rombach, and Björn Ommer. Taming 
transformers for high-resolution image synthesis. In Pro- 
cedings of the IEEE/CVF Conference on Computer Vision 
and Pattern Recognition, 2021. 1, 2, 6, 12

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing 
Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and 
Yoshua Bengio. Generative adversarial nets. In Advances in 
Neural Information Processing Systems, 2014. 2

Deep residual learning for image recognition. In Proceed- 
ings of the IEEE/CVF Conference on Computer Vision and 
Pattern Recognition, 2016. 12

[15] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao 
Soares. Image captioning: Transforming objects into words. In 

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, 
Bernhard Nessler, and Sepp Hochreiter. Gans trained by a 
two time-scale update rule converge to a local nash equilib- 
rium. In Advances in Neural Information Processing Sys- 
tems, 2017. 6

[17] Lun Huang, Wennmin Wang, Jie Chen, and Xiao-Yong Wei. 
Attention on attention for image captioning. In Proceed- 
7

[18] Eric Jang, Shixiang Gu, and Ben Poole. Categorical repa- 
parameterization with gumbel-softmax. In Proceedings of 
the International Conference on Learning Representations, 
2017. 2

generator architecture for generative adversarial networks. 
In Proceedings of the IEEE/CVF Conference on Computer 
Vision and Pattern Recognition, 2019. 12

[20] Diederik P Kingma and Max Welling. Auto-encoding vari- 
tional bayes. In Proceedings of the International Conference 
Learning Representations, 2014. 3

[21] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uij- 
lings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan 
Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, 
and Vittorio Ferrari. The open images dataset v4. Interna- 
tional Journal of Computer Vision, 128(7), 2020. 12

[22] Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng 
Gao, and Dan Jurafsky. Deep reinforcement learning for di- 
alogue generation. In Proceedings of the Conference on Em- 
pirical Methods in Natural Language Processing, 2016. 7

[23] Xiujuan Li, Xi Yin, Chunyu Li, Pengchuan Zhang, Xiaowei 
Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu 
Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics 
aligned pre-training for vision-language tasks. In Proceed- 
2, 5, 7, 13

Bourdou, Ross Girshick, James Hays, Pietro Perona, Deva 
Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft 
coco: Common objects in context. In Proceedings of the Eu- 
ropean Conference on Computer Vision, 2014. 2, 5, 6, 12, 
13

Deep learning face attributes in the wild. In Proceedings


[51] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of