Knowledge-Based Systems 227 (2021) 107200

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Image generation from text with entity information fusion

Deyu Zhou [a,*], Kai Sun [a], Mingqi Hu [a], Yulan He [b]

[a] School of Computer Science and Engineering, Key Laboratory of Computer Network and Information Integration, Ministry of Education, Southeast University, China

[b] Department of Computer Science, University of Warwick, UK

ARTICLE INFO

Article history:
Received 11 January 2021
Received in revised form 21 May 2021
Accepted 2 June 2021
Available online 6 June 2021

MSC:
00-01
99-00

Keywords:
Image generation from text
Entity information fusion
End-to-end frameworks
Entity Matching Score

ABSTRACT

Image generation from text is the task of generating new images from a textual unit such as word, phase, clause and sentence. It has attracted great attention in both the community of natural language processing and computer vision. Current approaches usually employ an end-to-end framework to tackle the problem. However, we find that the entity information, including categories and attributes of the images, are ignored by most approaches. Such information is crucial for guaranteeing semantic alignment and generating image accurately. For two pictures of the same category, the emphasis of the corresponding text description may be different, but the images generated by these two sentences should have some similarities and the generation process can learn from each other. Therefore, we propose two novel end-to-end frameworks to incorporate entity information in the process of image generation. For the first framework, an image representation is generated from entity labels using the variational inference mechanism and then fused with the representation generated from the corresponding sentence. Instead of fusing the images in high-dimensional space, images are inferred and fused in the latent space (the low-dimensional space) in the second framework, where computationally intensive upsampling modules are shared. Moreover, a novel metric (Entity Matching Score) is proposed to measure the degree of consistency of the generated image with its corresponding text description and the effectiveness of the metric has been proved by the generated samples in our experiments. Experimental results show that both the proposed frameworks outperform some state-of-the-art approaches significantly on two benchmark datasets.

1. Introduction

Image generation from text, aiming at generating images based on natural language texts, is an interesting but challenging task. It can find applications in computer aided design, text illustration and data augmentation. As a hot research topic, it has attracted a lot of attentions in recent years [1–3]. However, due to the challenges faced in language understanding and image generation, it is far from being solved.

There are many related methods making progress on this task and they are usually based on two kinds of generative model, Variational Auto-Encoder (VAE) [4] and Generative Adversarial Networks (GAN) [5]. Yan et al. [6] tried to encode the text by variational inference and generate images in decoder and Razavi et al. [7] leveraged VAE with the deep autoregressive model and then estimate the density of pixel space by probability factorization directly. However, these VAE-based methods always have the problem of blurred generated images. Since the introduction of GAN [5], methods for text to image generation have shown some promising results. These methods learn to generate an image from a global sentence embedding in generator and judge the authenticity and relevance of the image by discriminator. StackGAN [2] designed a multi-stage model to generate higher resolution images step by step and this framework was followed by later work. Recently, AttnGAN [3] was proposed by leveraging attention over text for generating fine-grained images. Although the attention mechanism can be used to learn mappings between words of an input sentence and the corresponding parts of an output image, it does not guarantee to learn the correct alignment between entity words in texts and objects in images due to the lack of supervision information.

In this paper, we explore a different way to accurately capture entity information. The term '*entity*' here refers to a class of instances or the specific attributes in real images. For example, the sentence '*This little blue bird is almost completely blue with black primary and secondaries*' describes a bird, but has a more finely grained entity label '*Indigo Bunting*', which has more fine-grained information. If we know '*Indigo Bunting*' is a kind of little blue bird and learn from the images in the same category, then generating image from this sentence will be easier than from scratch. We

* Corresponding author.
E-mail addresses: d.zhou@seu.edu.com (D. Zhou), sunkai@seu.edu.com (K. Sun), mingq@seu.edu.com (M. Hu), yulan.he@warwick.ac.uk (Y. He).

https://doi.org/10.1016/j.knosys.2021.107200
0950-7051/

know that images with same entity often have certain similarity, and this information can directly affect the overall image style or the specific objects in the image. To this end, we propose a novel image generation method to better incorporate the entity information of the text which in turn allows generating images more accurately. Specifically, an end-to-end structure which combines the entity-level network and the sentence-level network is proposed to learn entity representation and global semantics simultaneously. The entity representation and global semantics are then fused to enable the generation of better semantically-aligned images.

The main contributions of this paper are summarized as follows:

- We propose two end-to-end architectures to incorporate entity information. The first one uses a class-conditional generation framework to generate the entity image and then fuses both the images generated from the entity and the sentence in the image space. The second one infers the latent semantics from the entity and sentence separately and then fuses them in the latent space to generate the final image. The CapsuleNet [8] is also introduced into the generator and discriminator to further enhance entity representation learning.
- We propose a novel evaluation metric, Entity Matching Score (EMS), to evaluate the degree of a model capturing entity representation. It can also be used to measure the consistency between the generated image and the text.
- Qualitative and quantitative experiments are conducted on two datasets, and results show that the proposed approach significantly outperforms the existing state-of-the-art models. The methods have obvious advantages in the acquisition of entity based on our EMS metric. The detailed ablation study also shows the effectiveness of entity information fusion.

2. Related work

Image generation from text has attracted more and more interests in recent years. This task was usually formulated as the conditional image generation problem and approaches to text-to-image generation can be divided into three main categories, conditional VAE-based methods, autoregressive-based methods and conditional GAN-based methods.

Conditional Variational Auto-Encoder (CVAE) [6] was built on the VAE [4] in which the generative process is conditional on some observed variable. In the text-to-image generation task, the condition would be the input text. But CVAE approaches suffer from the same limitation of VAE that the generated images are usually blurry.

The autoregressive-based model is to directly estimate the density of pixel space by probability factorization and maximum likelihood estimation. The representative approaches are PixelRNN [9] and PixelCNN [10]. Recently, Razavi et al. [7] proposed a new method, VQ-VAE, which models the density on a low-dimension discrete latent space and then generate images by a decoder. VQ-VAE leverages VAE with the deep autoregressive model and it can generate high-resolution photo-realistic images.

Compared with other methods, approaches built on conditional Generative Adversarial Networks (CGAN) [11] are still the mainstream methods for text to image generation and show promising results. Reed et al. [1] firstly proposed an effective conditional GAN framework for text-to-image generation. Zhang et al. [2] proposed StackGAN which generates high-resolution images from the low-resolution images with multi-stages. However, all of these methods are conditioned on the global sentence embeddings and do not attempt to disentangle features encoding entities and their relations described in text.

Instead of generating images from natural language sentences, the class-conditional image generation [10,12] task aims to generate images given an image class label. Odena et al. [12] proposed AC-GAN, which used an auxiliary classifier in the discriminator to impose the class-conditional constraint. Miyato and Koyama [13] proposed a projection discriminator to solve the mode-collapse problem caused by AC-GAN and further improved the generation performance. Instead of improving the discriminator, Zhang et al. [14] applied the self-attention mechanism in the generator to generate more globally coherent images. Hu et al. [15] proposed a variational conditional generator framework to learn latent attributes in the class category. Hinz et al. [16] attempted to generate images from entities in the COCO dataset [17]. However, to the best of our knowledge, there is no work incorporating the entity representation into the text-to-image generation framework.

More recently, AttnGAN [3] was proposed to learn the mappings between words and regions in images through the attention on text at different stages. However, the attention mechanism between specific word and sub-region does not guarantee to learn the correct alignment between entity words and image objects. The following work of AttnGAN mainly focuses on the network structure, such as the introduction of cycle structure [18] and siamese network [19]. Compared with these methods which try to change the network structure, our method pays more attention to the introduction of entity information.

There are also some researches focus on the layout of the image. Hong et al. [20] proposed a pipeline framework, which firstly generates the semantic layout of objects from text and then generate images. Hinz et al. [21] try to fuse the layout encoding, spatial label and image caption to generate images in COCO dataset. However, a lot of supervised information such as the masks of objects and the spatial location is needed and the data pre-processing is also very cumbersome.

In this paper, we propose two novel end-to-end text-to-image frameworks to fuse entity information and only extra entity labels are needed.

3. Methodology

In this section, two end-to-end frameworks with entity representation learning are proposed for text-to-image generation. The first framework generates an entity image from an entity label and then fuses the entity image with the image generated from the corresponding sentence. The second framework infers the latent space from a given entity label and a sentence separately, and then generates images from the fused representations. The main difference between them is that the entity information was fused in the high-dimensional image space in the first framework and the second one fuses entity in the low-dimensional latent space.

In our experiments, we implement our two frameworks based on two representative baseline model. In this section, we first introduce two baseline models used in our experiments for easy understanding, then we give detailed descriptions for two proposed frameworks respectively.

3.1. The baseline models used to implement the frameworks

As a generative model, GAN can generate images by random noise, and the discriminator will restrict the generated image to be close to the real image. Conditional GAN(CGAN) [11] adds a

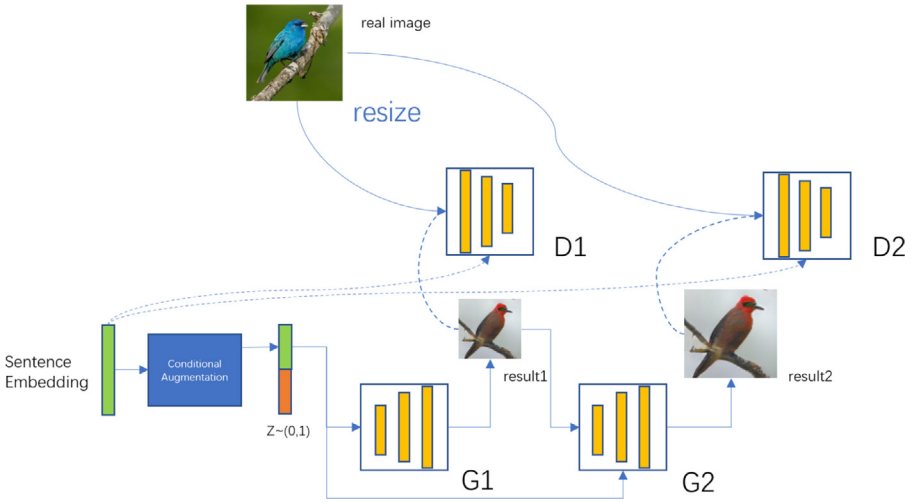

Fig. 1. The structure of Stacked Generative Adversarial Networks (StackGAN).

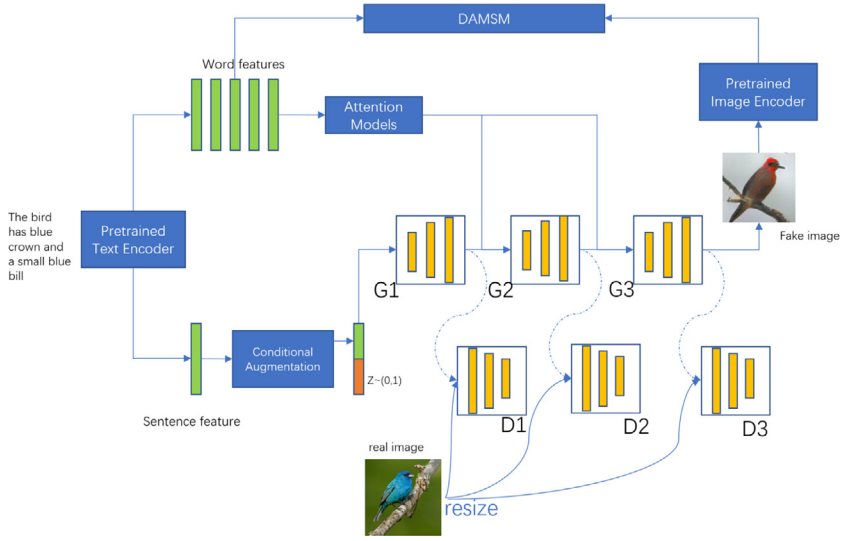

Fig. 2. The structure of Attentional Generative Adversarial Network (AttnGAN).

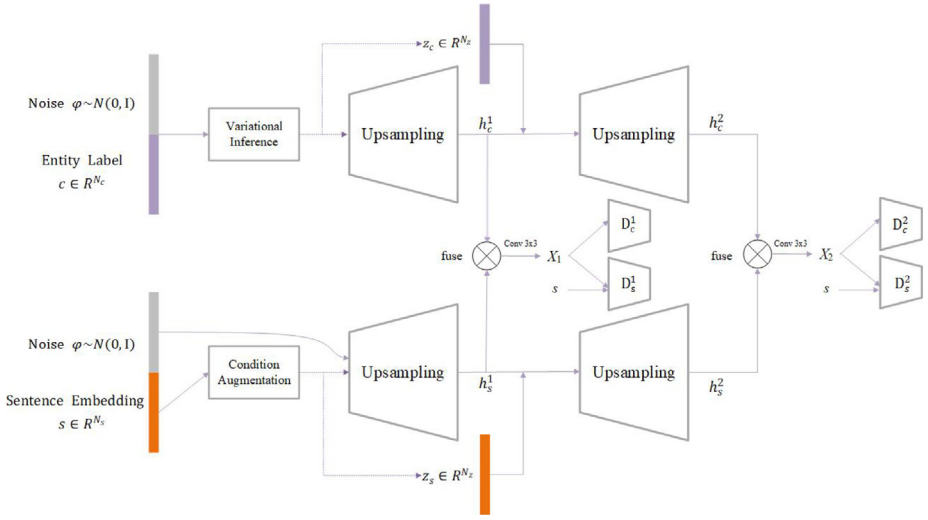

Fig. 3. The text-to-image generation framework based on image space information fusion.

class y to the input of generator and discriminator as label, and then the training loss function becomes:

$$L(D, G) = \mathbb{E}_{\mathbf{x},\mathbf{y}\sim\ p_{data}(\mathbf{x},\mathbf{y})}[\log(D(\mathbf{x}, \mathbf{y}))]+ \mathbb{E}_{\mathbf{z}\sim\ p_{\mathbf{z}}(\mathbf{z}),\mathbf{y}\sim\ p_{data}(\mathbf{y})}[\log(1 - D(G(\mathbf{z}, \mathbf{y}), \mathbf{y}))]. \quad (1)$$

It means that GAN has different objective functions for different labels, so we can use the labels to control the generated images. The basic idea of text to image generation is that using text as condition in CGAN and Reed et al. [1] has proved the feasibility of this idea.

However, experiments show that CGAN cannot generate images with high resolution directly and specific word features were ignored in the sentence embedding. StackGAN [22] and AttnGAN [3] introduced hierarchy generation and attention mechanism in this task and made great progress in terms of the generation quality.

Stacked Generative Adversarial Networks (StackGAN)

StackGAN was proposed to generate images with high resolution by a hierarchical generator and the design was widely used in the following work. The structure of StackGAN can be divided into two stages. As shown in Fig. 1, the first stage is a standard conditional GAN. The input is the sentence embedding pretrained by the text description and Conditioning Augmentation (CA) module is used to yield more training pairs to enhance the robustness of the model.

In the first stage, the generated low resolution 64 × 64 image and real data are used for adversarial training to obtain a coarse-grained generation model; the second stage takes the results of the first stage and original text embedding as input, and the second generator will generate high-resolution 128 × 128 images. In addition, both discriminators will use the sentence embedding to determine the relevance between the text and generated image.

StackGAN++ [2] was an improved version put forward by the same team of StackGAN and there has been some changes in training methods, including transforming the two-stage training process into the end-to-end training and set the generated image size and the number of stages to be adjustable. In our following implementations, we choose the training strategy from StackGAN++ and set the size of the generated image to be 128 × 128.

Attentional Generative Adversarial Network (AttnGAN)

In StackGAN, we use the pretrained sentence embedding and the word features has been ignored. AttnGAN focuses on the related words in natural language description and synthesizes the features of different sub-regions and word features of the image by attention mechanism. The structure of the AttnGAN is shown in Fig. 2, and the discriminators are omitted here for the same setting as StackGAN.

As shown in Fig. 2, the input has been replaced with the original text and the pretrained text encoder will extract feature from text in sentence level and word level. The sentence feature is the main input similar to StackGAN and the word features add specific details iteratively in high level generator with attention mechanism. In AttnGAN, a deep attention multimodal similarity model is also added to calculate the fine-grained image-text matching loss and make great progress in the experiments.

However, AttnGAN has not taken the specific entity information into account either. In our work, we propose two frameworks to fuse the entity information with the text feature and implement our frameworks based on StackGAN and AttnGAN.

3.2. Framework1: Entity Information Fusion in Image Space (EIF-IS)

As illustrated in Fig. 3, the first framework is a twin network based on a multi-stage generation structure [2]. The whole framework is composed of an entity-level network and a sentence-level network. The sentence-level network is similar to the standard StackGAN, which extract features from sentence embedding and transform it to the image space by a upsampling module. We follow the practice of VCGAN [15], which employing a variational inference mechanism to enhance the entity representation in the entity-level network and a similar upsample operation will extract the entity feature from entity labels. The details of the framework is described below:

Entity Information Fusion in Image Space

We use h_c to denote the image features of the entity-level network and h_s to denote the image features of the sentence-level network. Essentially, h_c encodes the entity information and h_s encodes more fine-grained information such as specific attributes and relations. To fuse the entity information encoded in h_c^i and the global information encoded in h_s^i (i for the ith stage), the element-wise product operation is used to obtain the final representation, $h_i = h_c^i \odot h_s^i$. Then h_i is passed to a size-invariant convolution layer to output the final image of this stage, $x = \text{conv}_{3\times3}(h_i)$. The generation process can be described as below:

$$\begin{aligned} z_c &= Encoder_c(\varphi, c), \\ z_s &= Encoder_s(s), \\ h_c^1 &= Upsampling_c(z_c), \\ h_s^1 &= Upsampling_s(z_s, \varphi), \\ h_1 &= h_c^1 \odot h_s^1, \\ X_1 &= Conv_{3\times3}(h_1); \end{aligned} \quad (2)$$

$$\begin{aligned} h_c^2 &= Upsampling_c(h_c^1, z_c), \\ h_s^2 &= Upsampling_s(h_s^1, z_s), \\ h_2 &= h_c^2 \odot h_s^2, \\ X_2 &= Conv_{3\times3}(h_2). \end{aligned} \quad (3)$$

As shown in the above, a two-stage learning framework for sentence-level image generation based on image space representation fusion is given. The model can be stacked back to generate higher resolution images. $Encoder_c$ and $Encoder_s$ represent word-level and sentence-level hidden variable inferences. Two levels of image features are obtained through upsampling and element-wise multiplication fusion. Finally, a simple size-invariant convolution decoder outputs the target image of the stage. In this frame, we design two discriminators of different levels to judge semantic relevance and entity relevance respectively.

Sentence-level Discriminator

Similar to the general discriminator, the sentence-level discriminator D_m in the framework is used to determine whether the input image matches a given text description condition. s indicates a sentence embedding. The objective function is defined as follows:

$$\begin{aligned} \mathcal{L}_m = &- \mathbb{E}_{x_r\sim p_{\text{data}}}\left[\log D_m\left(x_r, s\right)\right] \\ &- \mathbb{E}_{x_f\sim p_G}\left[\log\left(1 - D_m\left(x_f, s\right)\right)\right]. \end{aligned} \quad (4)$$

For image x, an intermediate representation is produced through the down-sampling operation of the discriminator. This image representation and the sentence embedding are connected as the input of the classification network to determine whether the image and the sentence match. The discriminator losses of all stages are added up as the final total loss to train the generator.

Entity-level Discriminator

The original discriminator judges whether an input image is real or fake and whether it matches a given text description. To ensure that the final image contains the desired entities, an entity-level discriminator D_c is proposed to learn the entity-level

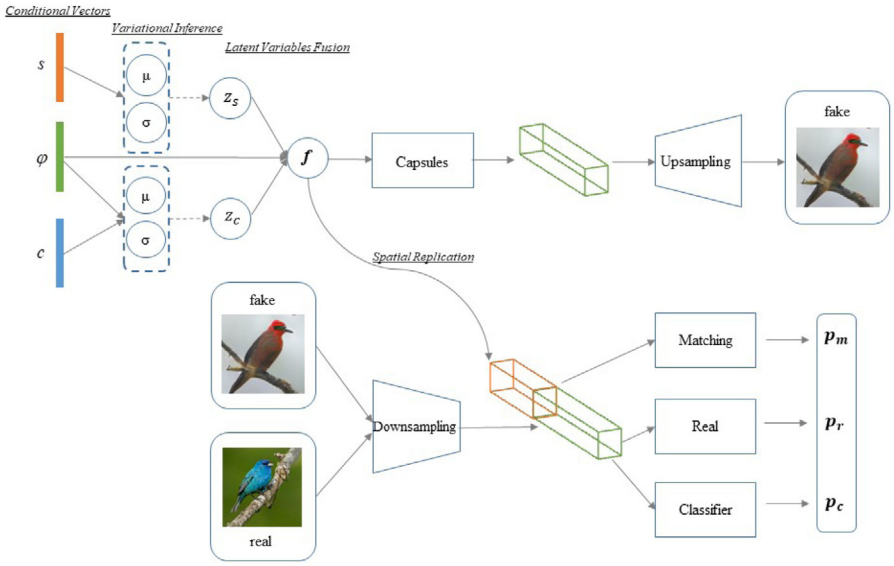

Fig. 4. The text-to-image generation framework based on latent space information fusion. s denotes the sentence embedding and c denotes the entity label.

network. The discriminator outputs both a probability distribution over the generated entity image e, $p(e|x)$, and a probability distribution over the entity classes c, $p(c|x)$. The objective function of the entity-level discriminator D_c includes two parts:

$$\begin{aligned} L_e &= \mathbb{E}[\log p(e = real|x_r) + \log p(e = fake|x_f)], \\ L_c &= \mathbb{E}[\log p(c = c_x|x_r) + \log p(c = others|x_f)]. \end{aligned} \tag{5}$$

where L_e is the standard GAN objective [5], which maximizes the log-likelihood for the binary classification task and is equivalent to minimizing the Jensen–Shannon Divergence between the true data distribution and the model distribution. In L_c, an additional class, '*others*', is introduced to label the generated image x_f, which means not belonging to any of the known entity classes.

In the case of an image containing multiple entities such as those in the COCO dataset [17], we use *Margin Loss* [8] to optimize each entity category separately. It is defined as follows:

$$\begin{aligned} L_k =& T_k \max\left(0, m^+ - \|\mathbf{v}_k\|\right)^2 \\ &+ \lambda\,(1 - T_k) \max\left(0, \|\mathbf{v}_k\| - m^-\right)^2, \end{aligned} \tag{6}$$

where L_k represents the k th output neuron, and takes $m^+ = 0.9, m^- = 0.1$. When the entity k appears, $T_k = 1$. λ is used to weaken the influence of categories that do not appear in the initial learning on model optimization, usually taken as $lambda = 0.5$. The total loss is a simple sum of all capsule losses.

3.3. Framework2: Entity Information Fusion in Latent Space (EIF-LS)

Although the entity information can be injected into the model learning through image space fusion in the first framework, the number of the network parameters is huge. Thus we proposed a simplified framework to fuse entity knowledge in the low-dimensional latent space, where computationally intensive upsampling modules are shared. In addition to this, to further improve the generation performance, the Capsule network [8] is also used in both the generator and the entity-level discriminator framework. The framework is illustrated in Fig. 4.

Entity Information Fusion in Latent Space

The latent variable z_c is inferred from the entity class label and z_s is inferred from the sentence embedding. The dimensions of both latent variables are fixed to 128. We proposed three ways to fuse the entity representation and the global representation: (a) *Summation*: $z = z_c + z_s$; (b) *Product*: $z = z_c \odot z_s$; (c) *Concatenation*: $z = [z_c, z_s]$. We choose the best fusion method empirically. The generation process based on latent space representation fusion is described as follows:

$$\begin{aligned} z_c &= Encoder_c(\varphi, c), \\ z_s &= Encoder_s(s), \\ z &= fuse(z_c, z_s, \varphi), \\ h &= Upsampling(Capsule(z)), \\ X &= Conv_{3\times3}(h). \end{aligned} \tag{7}$$

Similar to the generation process based on image space fusion, the one-stage image hidden representation h and the fused conditional representation z is sent to the next stage for generation. The *fuse* function here is one of the three fusion methods described above (noise is also added). In this frame, we design a novel capsule layer to enhance the learning ability of entity representation and the experiment has proved our idea.

Generator with Capsules

The basic structures of the generator are based on the Deep Convolution GAN (DCGAN) [23]. The noise vector or conditional vector will be first mapped to a long, narrow three-dimensional image feature space (e.g., $1024 \times 4 \times 4$) by a fully-connected layer, which will be enlarged in size and compressed in channels to output the image by upsampling operations. However, the conditional vector usually resides in a low-dimensional space such as 128 dimensions. The mapping from low dimensions of 128 to very high dimensions $1024 \times 4 \times 4$ is difficult to learn by a fully-connected layer.

Thus, we design a novel *Linear Capsule Layer* to replace the original fully-connected layer to enhance the learning ability of entity representation. A capsule is a new type of neuron whose activity vector represents the instantiation parameters of a specific type of entity such as an object or an object part [8]. It is suitable for modeling the latent space in the proposed framework and the following ablation study will confirm this. In our experiments, we use one capsule layer with 1024 bottom-shared capsules with the input size of 16×8 and the output size of 1024×16 which is reshaped to $1024 \times 4 \times 4$.

Discriminator with Capsules

The activity vector in capsules represents the probability of the corresponding entity [8]. Therefore, we take the capsule layer

instead of the last fully-connected layer for classification. An m-dimensional vector represents a particular category of entities. The output of the capsule layer is of the size $N \times m$, where N is the number of all entities. The output will be normalized by rows to produce N-dimension probability vector. At last, a softmax layer is applied to the vector to output the final N-category probability distribution. We use the capsules to improve the classification performance of the entity-level discriminator and constrain the entity-level generator better.

Also, taking the advantage of multi-task learning, we share the downsampling operations (i.e., convolutions) of both entity-level and sentence-level discriminators. There are three heads for the shared hidden image features, which correspond to the p_m, p_r, p_c in Fig. 4 separately. The first one is the *matching head*, which reads the fused conditional embeddings and image features (concatenated along the depth dimension), then outputs the matching score of the image and the conditional input. The second one is the *real/fake head* for judging the fidelity of the image and the third one is the *classifier head* for judging if the image contains the desired input entities.

3.4. Training

Both frameworks are implemented with conditional GAN based methods, and the Generator G is jointly trained with the Discriminator D. The final loss functions are given as follows:

$$\begin{aligned}\mathcal{L}_D &= -(\mathbb{E}_{x\sim P}[\log D(x)_c] + \mathbb{E}_{x\sim Q}[\log D(x)_c]) \\ &- (\mathbb{E}_{x\sim P}[\log D(x,s)_m] + \mathbb{E}_{x\sim Q}[\log D(x,s)_m]) \\ &- (\mathbb{E}_{x\sim P}[\log D(x)_r] + \mathbb{E}_{x\sim Q}[\log(1-D(x)_r)]);\end{aligned} \tag{8}$$

$$\begin{aligned}\mathcal{L}_G &= -\mathbb{E}_{x\sim Q}[\log D(x)_r] - \mathbb{E}_{x\sim Q}[\log D(x)_c] \\ &- \mathbb{E}_{x\sim Q}[\log D(x,s)_m] + KL(q(z_c|c,\varphi)\|p(z)) \\ &+ KL(q(z_s|s)\|p(z)).\end{aligned} \tag{9}$$

where P is the true data distribution and Q is the generated data distribution. $D(x)_r$ denotes the probability of image x being real, $D(x)_c$ denotes the probability over the correct entity class label and $D(x,s)_m$ denotes the matching probability between the image and the conditional input.

Two KL divergence terms are added to $\mathcal{L}_G$ in Eq. (9), as the regularization loss for constraining the latent variable z_c and z_s. We assume that the latent posterior q is a diagonal Gaussian with mean μ and standard deviation σ and the prior $p(z)$ is a standard Gaussian with zero mean and unit variance.

4. Experiments

We conduct experiments to evaluate the performance of the proposed two frameworks in comparison with existing models. We also perform ablation study to gain more insights into our proposed frameworks.

4.1. Datasets and evaluation

Datasets. The experiments are conducted on two datasets, CUB [24] and COCO [17], which are widely used in the text-to-image generation task. The statistics of the two datasets are presented in Table 1. The COCO dataset is more challenging since the images contained involve multiple entities and more complex layouts.

Evaluation. Following the evaluation setup in the previous text-to-image methods, Inception Score (IS) [25] is used to evaluate the quality and variety and the generated images. IS is defined below:

$$\mathrm{IS} = \exp\left(\mathbb{E}_{x\sim Q}[KL(p(y|x)\|p(y))]\right), \tag{10}$$

Table 1
Statistics of the datasets.

Datasets	CUB	COCO
Training images	8,855	Over 80 k
Test images	2,933	Over 40 k
Class labels	200	80
Captions per image	10	5

where $p(y|x)$ is the conditional class distribution, and $p(y) = \int_x p(y|x)p(x)$ is the marginal class distribution. The higher IS value indicates that the generated images contain clearer and more recognizable objects. In our experiments, we use the pretrained Inception model provided in [2] to evaluate the performance of our approaches.

One limitation of IS is that it cannot evaluate the semantic consistency. A recently proposed metric, R-precision [3], is used for measuring consistency. However, it depends on pretrained encoders for generating embeddings to calculate the similarity. But the encoders employed in recent works are actually different [3, 18].

Since text and images reside in two different embedding spaces, it is hard to define the similarity between text and image. We propose an entity-based metric, called *Entity Matching Score (EMS)*, to evaluate the coarse-grained alignment between a given piece of text and an image. Given a sentence and the class label of its main entity, c_i, a pretrained classifier will output $P_{c_i}(x)$, the probability that the generated image x contains an object belonging to the class c_i. The final EMS score is the aggregated probabilities over N images can it can be calculated as follows:

$$\mathrm{EMS} = -\frac{1}{N}\sum_{i=1}^{N}\log P_{c_i}(x_i). \tag{11}$$

The approach with lower EMS is better. In our experiments, we use the same pretrained Inception net (the same as the one in IS) as the classifier and the N can be set as 30,000 on CUB.

Baselines. We choose three recently proposed models as the baselines: StackGAN [2], AttnGAN [3] and MirrorGAN [18]. In the following experiments, we first implement two frameworks based on StackGAN on CUB and COCO dataset, then we have done further research based on AttnGAN on CUB dataset.

For all the experiments, the dimension of the latent variable and the noise variable is fixed to 128. The entity class labels are encoded as one-hot or binary representations (for multiple attributes in CUB or categories in COCO). We choose Adam solver with hyper-parameters set to $\beta_1 = 0.5, \beta_2 = 0.999$ and the learning rate $\alpha = 0.0002$. The balanced update frequencies (1:1) of discriminator and generator are employed. We use the char-CNN-RNN text encoder provided by [26] to encode each sentence into a 1024-dimensional text embedding. The batch normalization [27] is used in both the generator and the discriminator, and spectral normalization [28] is used in the discriminator to make the training more stable.

4.2. Implementation based on StackGAN

To show the effectiveness of our proposed frameworks, we first conducted experiments with a two-stage network on CUB and COCO based on StackGAN. The entity label we use in the experiment is the category of the image. In the experiment, we followed VC-GAN [15] to use variational inference to encodes the entity information and use pretrained sentence embedding [1] as the text information. And these two kinds of information was then fused in the image space (EIF-IS) or latent space (EIF-LS). The two-stage GAN generates 64 × 64 and 128 × 128 images

Fig. 5. The frameworks based on StackGAN. The generated 128 × 128 samples from CUB test set.

Fig. 6. The frameworks based on StackGAN. The generated 128 × 128 samples from COCO test set.

respectively, and we compare and show the results of the second stage with the baseline method.

We verified our framework in experiments and compared the results with StackGAN which are shown in Table 2. Here, EIF-IS denotes our proposed first framework, Entity Information Fusion in Image Space, while EIF-LS denotes the proposed second framework, Entity Information Fusion in Latent Space. From Table 2, it can be observed that EIF-IS and EIF-LS achieve similar Inception scores on CUB, which improve over StackGAN by 11.6%. On the more complex dataset, COCO, EIF-LS performs better than EIF-IS and outperforms StackGAN by achieves more higher score, which have a 14.4%. improvement compared with the baseline. The results show that our proposed method can make a significant improvement on image quality and produce the clear entity and the sampled results will confirm this.

The samples comparison between the proposed approaches and the baseline is shown in Figs. 5 and 6. From Fig. 5, all three models have successfully generated target images related to input text. Compared with the baseline, the images generated by the proposed method are clearer on the vision, and the entity details are more abundant and prominent. COCO is a more challenging dataset. From Fig. 6, the three models have roughly captured the main information in the text, but the details are still insufficient. Among them, the image generated by the baseline is blurry, and some of the components are difficult to distinguish. The main

Table 2
Performance comparison with StackGAN on CUB and COCO with 128 × 128 resolution.

Methods		StackGAN	EIF-IS	EIF-LS
IS ↑	CUB	3.35 ± .02	**3.74 ± .03**	3.73 ± .05
	COCO	7.34 ± .17	7.46 ± .30	**8.40 ± .28**

Table 3
Specific attributes and the corresponding parts in CUB dataset.

Parts	Beak	Belly	Tail	Wing	Body
Attributes	HasBillShape HasBillColor HasBillLength	HasBellyPattern HasBellyColor	HasUpperTailColor HasUnderTailColor HasTailPattern HasTailShape	HasWingPattern HasWingColor HasWingShape	HasUnderPartsColor HasUpperPartsColor HasPrimaryColor
Parts	Back	Breast	Fore-head	Bird (all parts)	Throat
Attributes	HasBackColor HasBackPattern	HasBreastPattern HasBreastColor	HasFore-headColor	HasSize HasShape	HasThroatColor
Parts	Head	Leg	Crown	Nape	Eye
Attributes	HasHeadPattern	HasLegPattern	HasCrownColor	HasNapeColor	HasEyeColor

Table 4
Comparison with the state-of-the-art models on CUB with 256 × 256 resolution. N in EMS metric is set as 30,000 on CUB.

Methods		AttnGAN	MirrorGAN	EIF-LS
CUB	IS ↑	4.36 ± .03	4.56 ± .05	**4.79 ± .05**
	EMS ↓	4.01	3.77	**3.13**

entities in the text are also missing. The proposed method is more visually clear, the layout is more natural, and it has a better match with the text conditions, and the main entities are also better captured.

Therefore, from the quantitative and qualitative experimental results, the proposed method significantly exceeds the current baseline model in visual quality and text consistency, and can well highlight the main entity targets in the text. In addition, for the two fusion generation methods based on entity representation learning, in view of the better performance and higher time efficiency of the EIF-LS model, we chooses the EIF-LS model to continue compare with two state-of-the-art models.

4.3. Implementation based on AttnGAN

We also implement our EIF-LS based on AttnGAN and compare the results with the state-of-the-art models such as AttnGAN and MirrorGAN on CUB dataset. In this part, we use the same three-stage generator as AttnGAN, and generate 256 × 256 images in the last stage. Two kinds of entity label were used in our experiment. One is using the former category information as baseline, and the other is introducing attribute information [24], such as the color, shape and size of different parts of a bird. There are total 28 attributes of different parts and 312-dimensional binary vector was provided in the dataset and we are the first to try to use these features, which is generally served for image classification, to enhance the image generation. The specific attributes and the corresponding parts are shown in the Table 3. From Table 3, we can find that these attributes are much more fine-grained than the separate category label. Obviously, the second entity label can provide richer entity information than the first one.

In the baseline method, we process the entity information with the same method as the implementation based on StackGAN, and we use pretrained text encoder to replace the pretrained sentence embedding for calculating the attention between each word and the sub-region of the image. While we choose the second kind of entity label, we replace the category label with a 312-dimensional attribute vector, and the parts of text processing are consistent. We compare quantitive results of the two different entity labels in Table 4, and show images generated by the last stage of better model in Fig. 7.

From Table 4, it can be observed that MirrorGAN gives superior results compared to AttnGAN. But our EIF-LS built on AttnGAN achieves the highest IS and EMS scores, outperforming MirrorGAN.

The samples of the proposed method based on AttnGAN is been shown in Fig. 7. The results show that AttnGAN-based method can generate higher resolution images with more details and our work also acts on improving image quality and consistency. From Fig. 7 and Table 4, we can find that methods with lower EMS score show better relevance of entities in text and image space and it can confirm the effectiveness of the EMS metric.

To further illustrate the effect of entity information fusion, we conduct experiments using different types of entity information in CUB dataset. The results are shown in Table 5. Compared with using the label of "category", attribute labels provides a more fine-grained constraints for text to image generation. The results show that if more entity knowledge can be reasonably incorporated in the training, the quality of generated images will be improved correspondingly. This ablation study further demonstrates the necessity of incorporating entity information.

4.4. Ablation study on EIF-LS framework

To gain better insights into our proposed frameworks, we conduct an ablation study on the more optimal framework, EIF-LS, on CUB.

Different Ways of Latent Space Fusion

We first evaluate the different ways of latent space fusion and show the results in Table 6. We can see the concatenation of the latent variable representing the entity representation and that of sentence embedding (EIF-LS-Concat) gives the best results overall, and it significantly outperforms the baseline on EMS. We also notice that fusing latent variables through summation (EIF-LS-Sum) or element-wise production (EIF-LS-Product) is less effective and both perform worse compared to StackGAN on IS. However, the generation consistency is still improved on EMS. We suspect that latent variable fusing by summation or element-wise production may lose some information but the entity representation is helpful to the consistency.

Capsule Network in Generator

In order to learn the entity representation and global semantics better, we introduce the capsules in both the generator and discriminator networks as presented in Section 3.3. The results of the Generator with Capsules (with G-Cap) are shown in Table 7. Compared Table 7 with Table 6, we can see that the scores of the generator with capsules are better than the one without across all latent space fusion methods on both metrics. This shows that the generator with capsules is an effective combination compared with the generator with fully connected layer. Also, EIF-LS-Sum and EIF-LS-Product with G-Cap are comparable with StackGAN and EIF-LS-Concat with G-Cap significantly outperforms StackGAN with 11.6% improvement on IS and 17.4% improvement on EMS.

Fig. 7. The frameworks based on AttnGAN. The generated 256 × 256 samples from CUB test set.

Table 5
Results of different entity information on CUB with 256 × 256 resolution. *N* in EMS metric is set as 30,000 on CUB.

Methods		AttnGAN	EIF-LS w/category	EIF-LS w/attribute	EIF-LS w/category + attribute
CUB	IS ↑	4.36 ± .03	4.67 ± .05	4.71 ± .05	**4.79 ± .05**
	EMS ↓	4.01	3.39	3.39	**3.13**

Table 6
Results of different latent space fusion on CUB.

Methods		StackGAN	EIF-LS-Sum	EIF-LS-Product	EIF-LS-Concat
CUB	IS ↑	3.35 ± .02	3.05 ± .03	3.10 ± .05	**3.48 ± .02**
	EMS ↓	4.95	4.60	4.71	**4.53**

Table 7
Results of generator with capsules on CUB.

Methods		StackGAN	EIF-LS-Sum w/G-Cap	EIF-LS-Product w/G-Cap	EIF-LS-Concat w/G-Cap
CUB	IS ↑	3.35 ± .02	3.39 ± .03	3.27 ± .03	**3.73 ± .05**
	EMS ↓	4.95	4.52	4.51	**4.09**

Table 8
Results of discriminator with capsules on CUB.

Methods		StackGAN	EIF-LS-Concat w/D-Cap	EIF-LS-Concat w/G-Cap;w/D-Cap
CUB	Inception Score ↑	3.35 ± .02	**3.71 ± .02**	3.45 ± .03
	EMS ↓	4.95	**4.50**	4.52

Capsule Network in Discriminator

Similarly, we conduct experiments on the discriminator with capsules using the concatenation of latent variables and show the results in Table 8. We can see that the combination of EIF-LS-Concat and Discriminator with Capsules gives a similar result on IS but poor performance on EMS compared to EIF-LS-Concat with G-Cap. This may be because capsule network increases the complexity of the model and increases the risk of overfitting of the entity discriminator on the CUB. However, putting everything together (EIF-LS-Concat + G-Cap + D-Cap) results in worse performance. We speculate that this is partly due to the difficulty in training GAN-like models when model complexity increases.

5. Conclusion

In this paper, we have proposed two novel frameworks for incorporating entity information into existing text-to-image architectures. In our work, the entity information and sentence semantics are learned simultaneously to complement and strengthen each other. The main difference between the two is the time to fuse the entity representation and global semantics and the experiments show that fusion in low dimensional hidden space can generate images with higher quality. We have also explored using capsules to further exploit the learning ability of entity representation and improve the performance. Experimental results show that the proposed approach significantly outperforms the baselines. In the future, we plan to investigate explicitly encoding entities and spatial information to better guide the image generation process.

CRediT authorship contribution statement

Deyu Zhou: Conceptualization, Supervision, Writing - review & editing. **Kai Sun:** Methodology, Code, Writing - original draft. **Mingqi Hu:** Methodology, Code, Writing - original draft. **Yulan He:** Writing - review & editing.

Acknowledgements

We are grateful to the reviewers for their valuable comments and constructive suggestions. This work was funded by the National Key Research and Development Program of China (2016YFC1306704), the National Natural Science Foundation of China (61772132), the EPSRC (grant no. EP/T017112/1, EP/V048597/1) and a Turing AI Fellowship funded by the UK Research and Innovation (UKRI) (grant no. EP/V020579/1).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, Honglak Lee, Generative adversarial text to image synthesis, in: Proceedings of the 33rd International Conference on Machine Learning, vol. 48, PMLR, 2016, pp. 1060–1069.

[2] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, Dimitris N Metaxas, Stackgan++: Realistic image synthesis with stacked generative adversarial networks, IEEE Trans. Pattern Anal. Mach. Intell. 41 (8) (2018) 1947–1962.

[3] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, Xiaodong He 0001, AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 1316–1324.

[4] Diederik P. Kingma, Max Welling, Auto-encoding variational Bayes, in: 2nd International Conference on Learning Representations, ICLR 2014, 2014.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.

[6] Xinchen Yan, Jimei Yang, Kihyuk Sohn, Honglak Lee, Attribute2image: Conditional image generation from visual attributes, in: European Conference on Computer Vision, Springer, 2016, pp. 776–791.

[7] Ali Razavi, Aäron van den Oord, Oriol Vinyals, Generating diverse high-resolution images with VQ-VAE, in: DGS@ICLR, OpenReview.net, 2019.

[8] Sara Sabour, Nicholas Frosst, Geoffrey E. Hinton, Dynamic routing between capsules, 2017, arXiv preprint arXiv:1710.09829.

[9] Aäron van den Oord, Nal Kalchbrenner, Koray Kavukcuoglu, Pixel recurrent neural networks, 2016, http://arxiv.org/abs/1601.06759.

[10] Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al., Conditional image generation with pixelcnn decoders, in: Advances in Neural Information Processing Systems, 2016, pp. 4790–4798.

[11] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1125–1134.

[12] Augustus Odena, Christopher Olah, Jonathon Shlens, Conditional image synthesis with auxiliary classifier GANs, in: Proceedings of the 34th International Conference on Machine Learning, vol. 70, PMLR, 2017, pp. 2642–2651.

[13] Takeru Miyato, Masanori Koyama, cGANs with projection discriminator, in: International Conference on Learning Representations, 2018.

[14] Han Zhang, Ian Goodfellow, Dimitris Metaxas, Augustus Odena, Self-attention generative adversarial networks, in: International conference on machine learning, PMLR, 2019, pp. 7354–7363.

[15] Mingqi Hu, Deyu Zhou, Yulan He, Variational conditional GAN for fine-grained controllable image generation, in: Asian Conference on Machine Learning, 2019.

[16] Tobias Hinz, Stefan Heinrich, Stefan Wermter, Generating multiple objects at spatially distinct locations, 2019, arXiv preprint arXiv:1901.00686.

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, C. Lawrence Zitnick, Microsoft coco: Common objects in context, in: European Conference on Computer Vision, Springer, 2014, pp. 740–755.

[18] Tingting Qiao, Jing Zhang, Duanqing Xu, Dacheng Tao, MirrorGAN: Learning text-to-image generation by redescription, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1505–1514.

[19] Guojun Yin, Bin Liu, Lu Sheng, Nenghai Yu, Xiaogang Wang, Jing Shao, Semantics disentangling for text-to-image generation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2327–2336.

[20] Seunghoon Hong, Dingdong Yang, Jongwook Choi, Honglak Lee, Inferring semantic layout for hierarchical text-to-image synthesis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7986–7994.

[21] Tobias Hinz, Stefan Heinrich, Stefan Wermter, Semantic object accuracy for generative text-to-image synthesis, 2019, arXiv preprint arXiv:1910.13321.

[22] Zhang Han, Xu Tao, Hongsheng Li, StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017.

[23] Alec Radford, Luke Metz, Soumith Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, 2015, arXiv preprint arXiv:1511.06434.

[24] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, Serge Belongie, The caltech-ucsd birds-200-2011 dataset, 2011.

[25] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, Improved techniques for training gans, in: Advances in Neural Information Processing Systems, 2016, pp. 2234–2242.

[26] Scott Reed, Zeynep Akata, Bernt Schiele, Honglak Lee, Learning deep representations of fine-grained visual descriptions, in: IEEE Computer Vision and Pattern Recognition, 2016.

[27] Sergey Ioffe, Christian Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: ICML, in: JMLR Workshop and Conference Proceedings, vol. 37, JMLR.org, 2015, pp. 448–456.

[28] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, Yuichi Yoshida, Spectral normalization for generative adversarial networks, in: International Conference on Learning Representations, 2018.