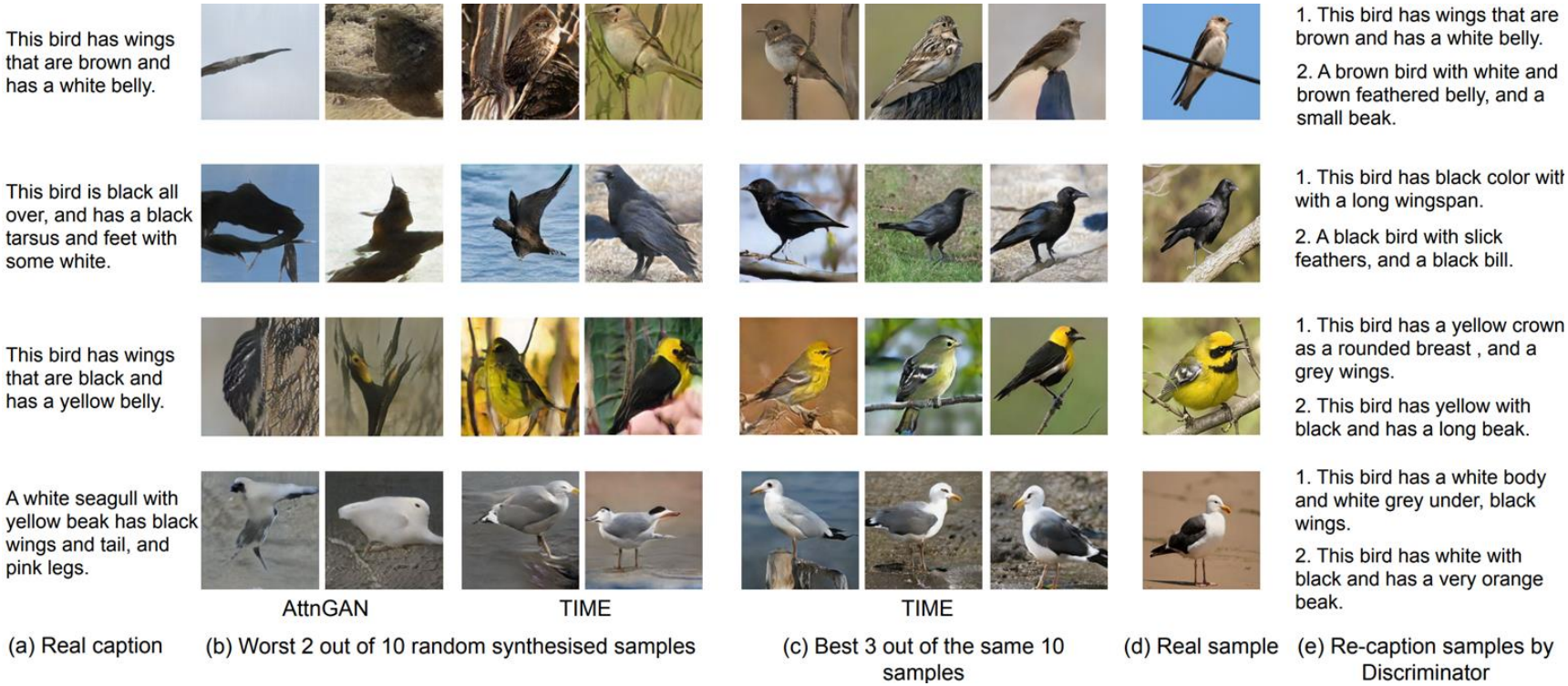
TIME: Text and Image Mutual-Translation Adversarial Networks

Bingchen Liu, Kunpeng Song, Yizhe Zhu,
Gerard de Melo, Ahmed Elgammal
Rutgers University

Text-to-image and image-to-text (image captioning)



Text descriptions → G → Generated images → D → Generated image captions

Contents

1. Motivation
    a. Prior works on text-to-image synthesis
    b. Mutual translation between the language and visual domains
2. Methods
    a. Achieve both text-to-image and image-captioning within one GAN
    b. Annealing the adversarial loss to balance the G&D training
    c. Positional encoding on image feature (haven been done in many works)
    d. Some technique model tricks (a more controllable generator)
3. Experimental results
    a. Transformer structure configuration
    b. FID & Text-image consistency, etc
    c. Qualitative results

Related works

Text to image models using GAN:
DM-GAN, OP-GAN, CP-GAN, DF-GAN, Dual-Attn GAN, MirrorGAN, SD-GAN, AttnGAN, and StackGAN, etc

Dynamic Memory module

Semantic Object Accuracy

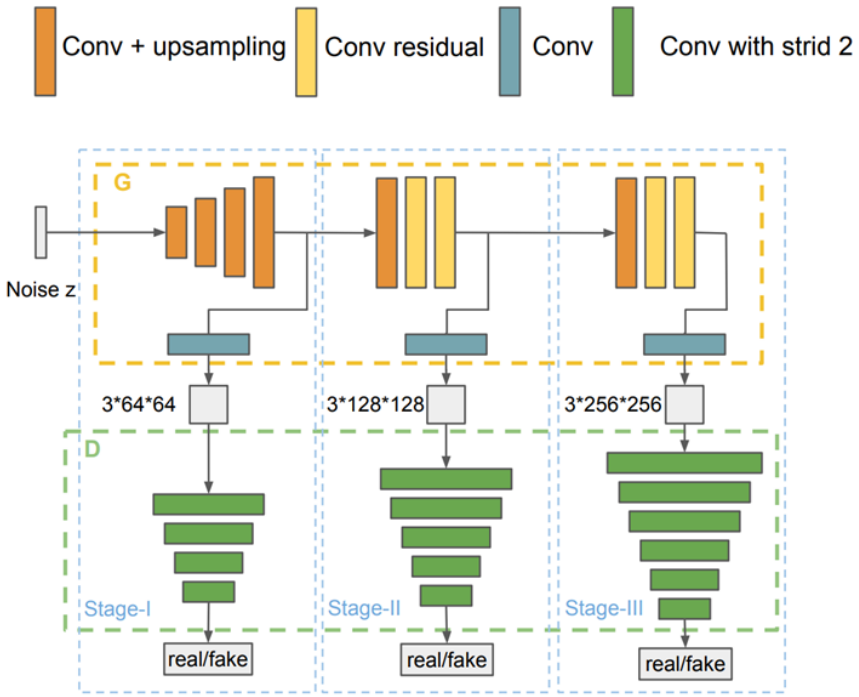
Semantic Disentangling

Content Parsing

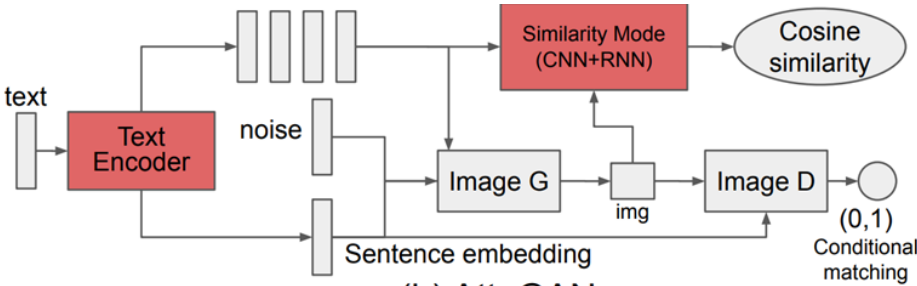
Deep Fusion

Mirror Redescription

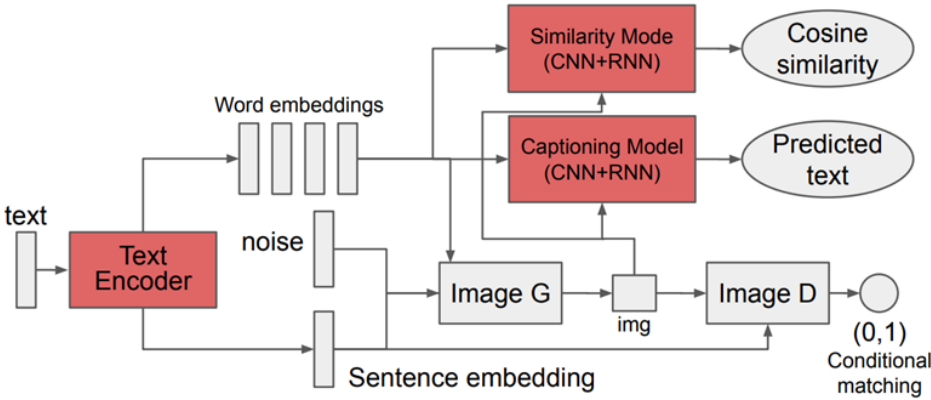
Related works



(a) StackGAN architecture

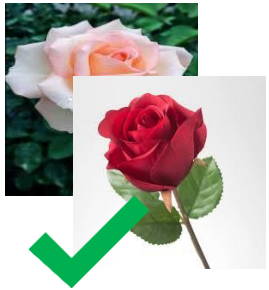


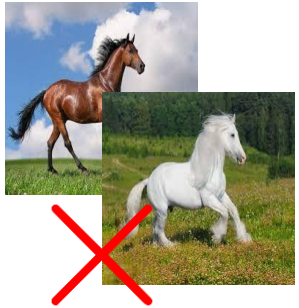
(b) AttnGAN



(c) MirrorGAN

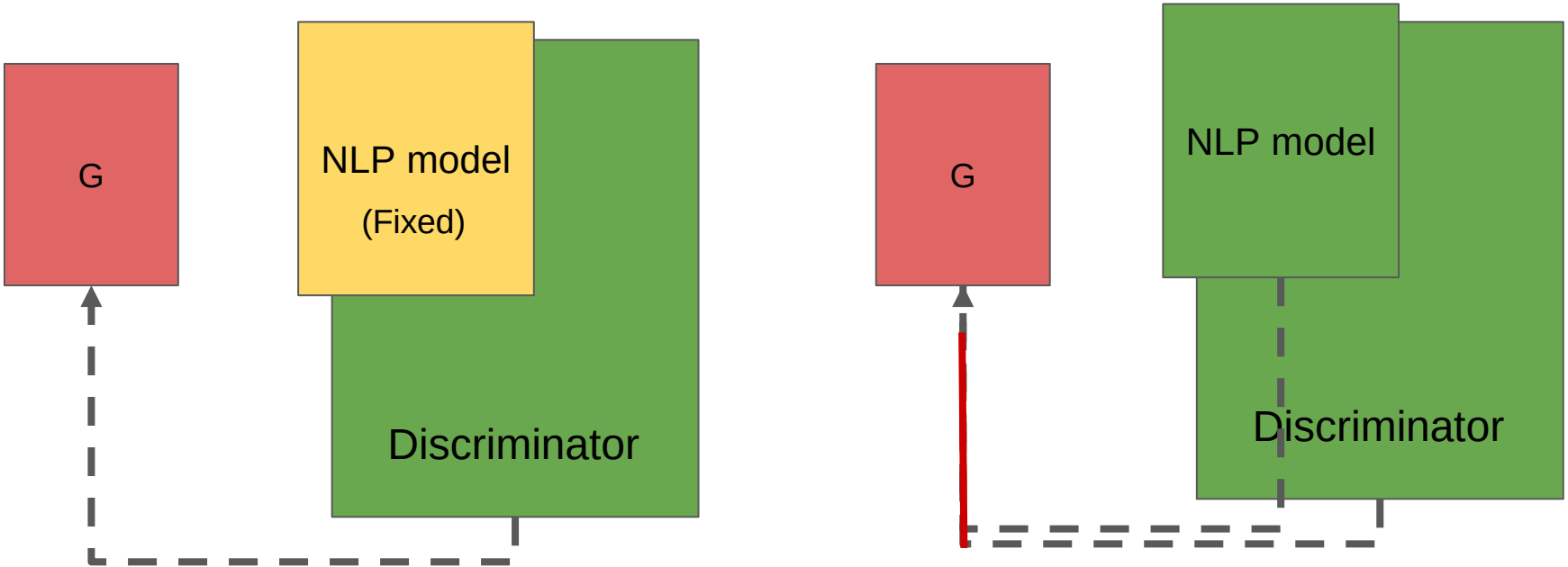
The Need for Mutual-Translation: Saving Time



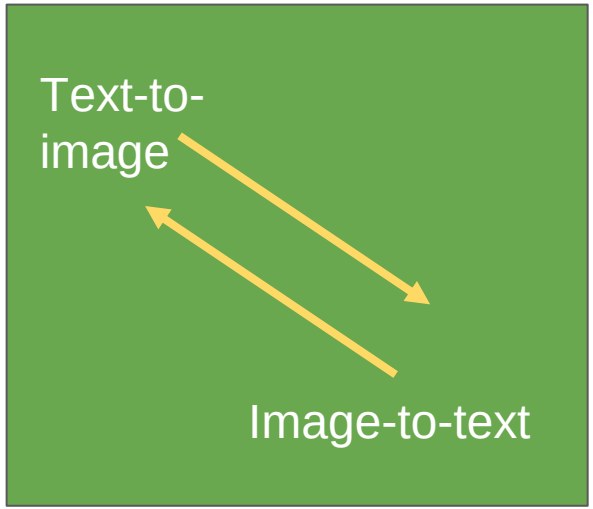


No pre-trained Language Model for the dataset

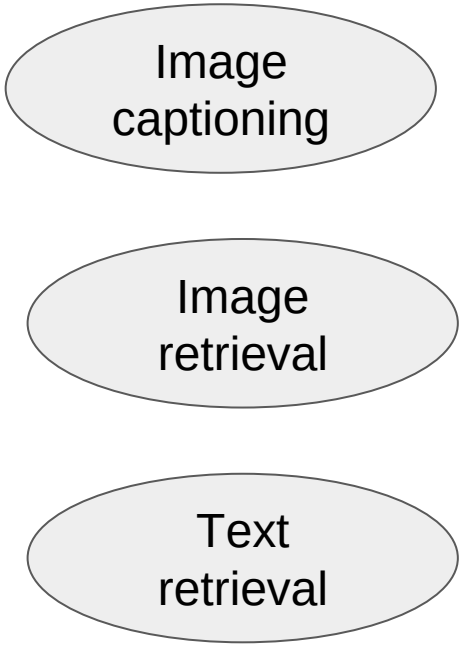
The Need for Mutual-Translation: Better Performance



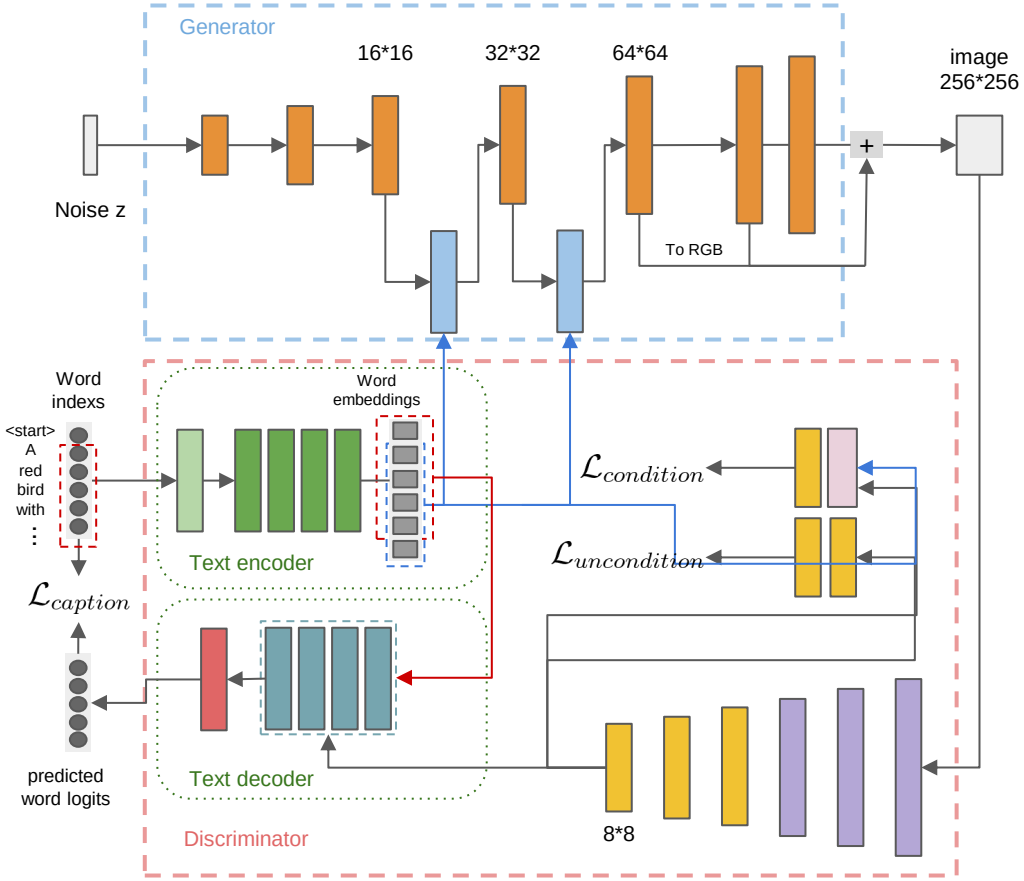
The Need for Mutual-Translation: Good Pre-Training

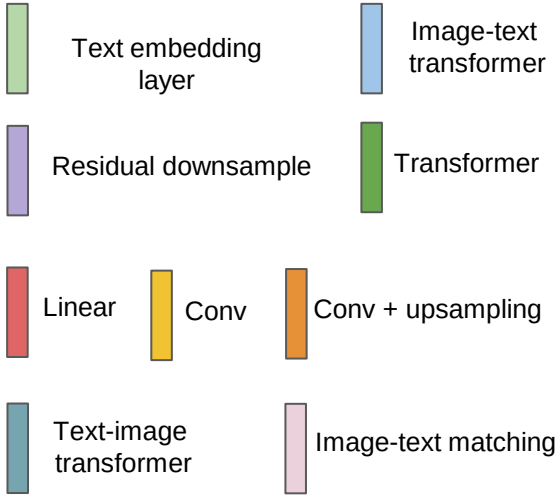


Mutual-translation model

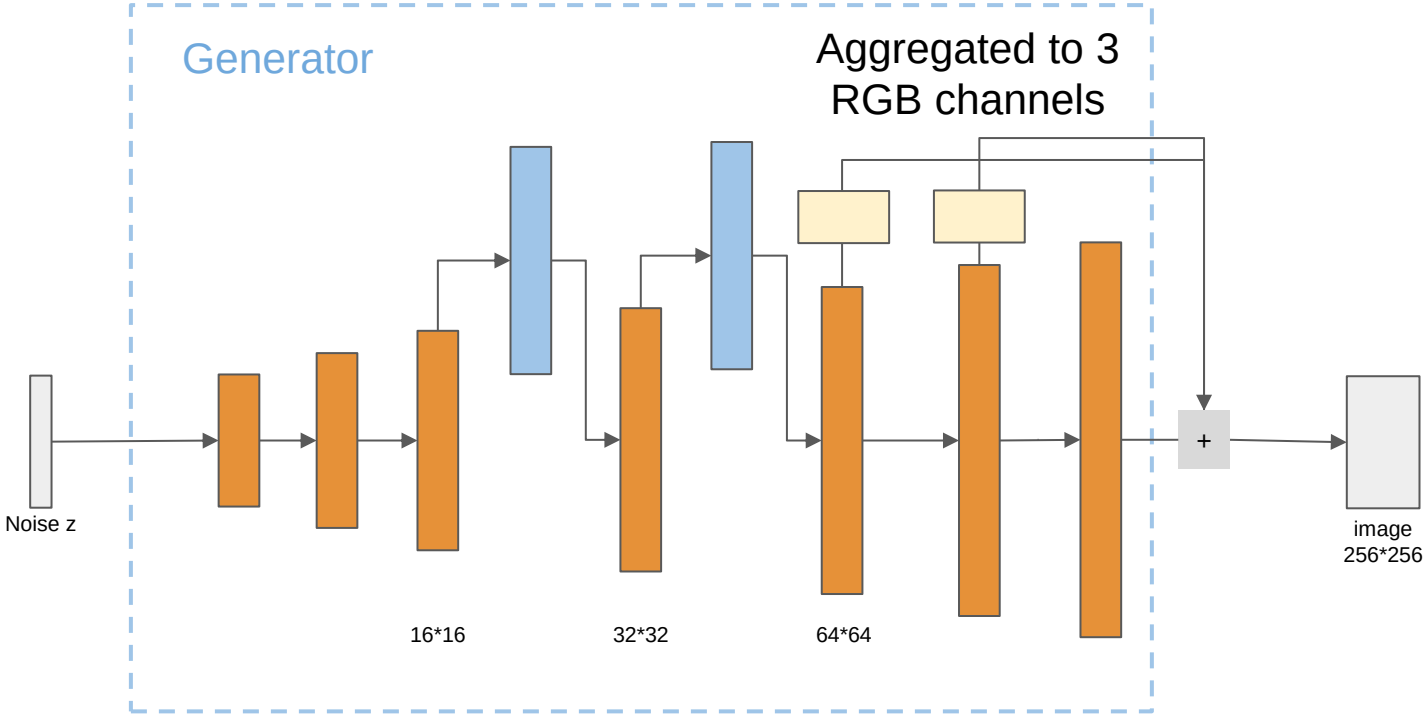


Methods: Model Overview

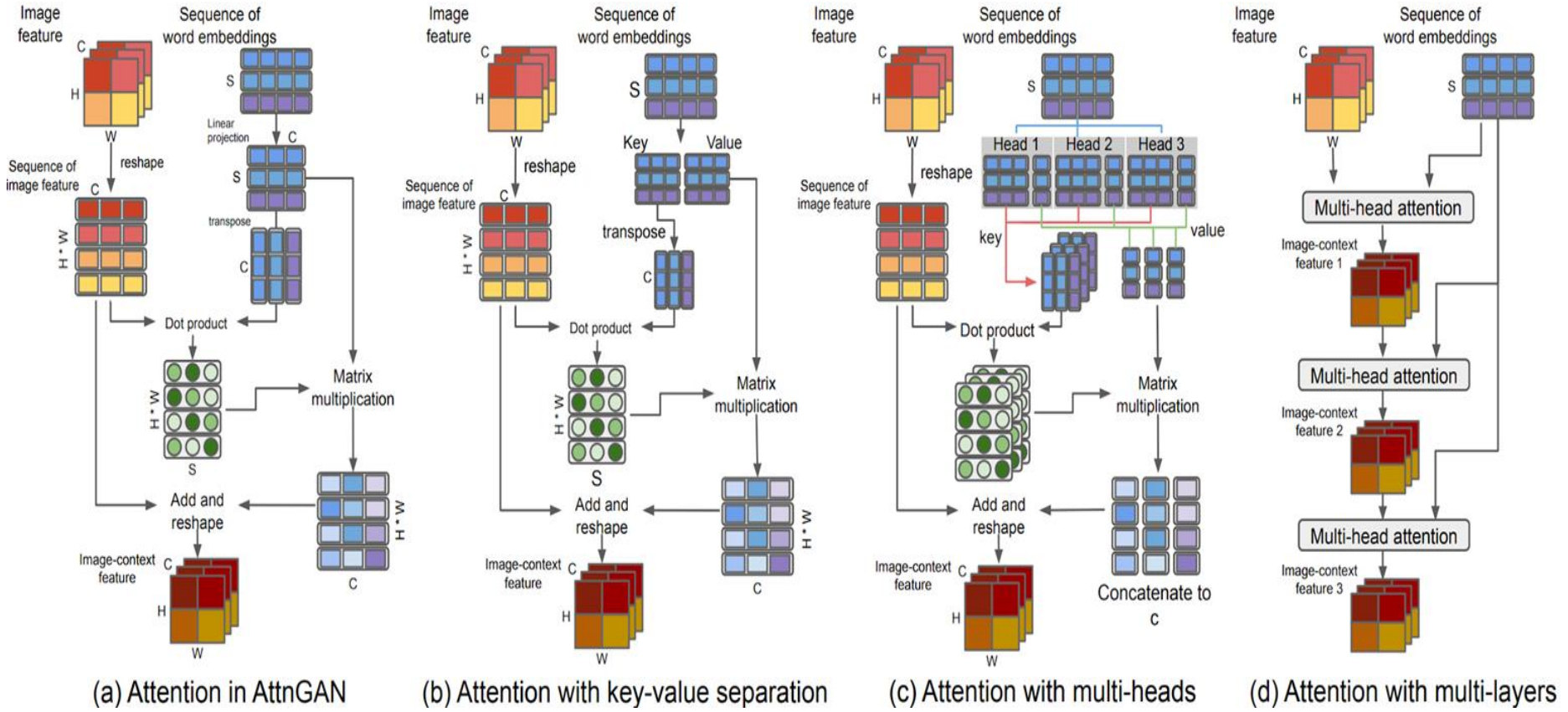




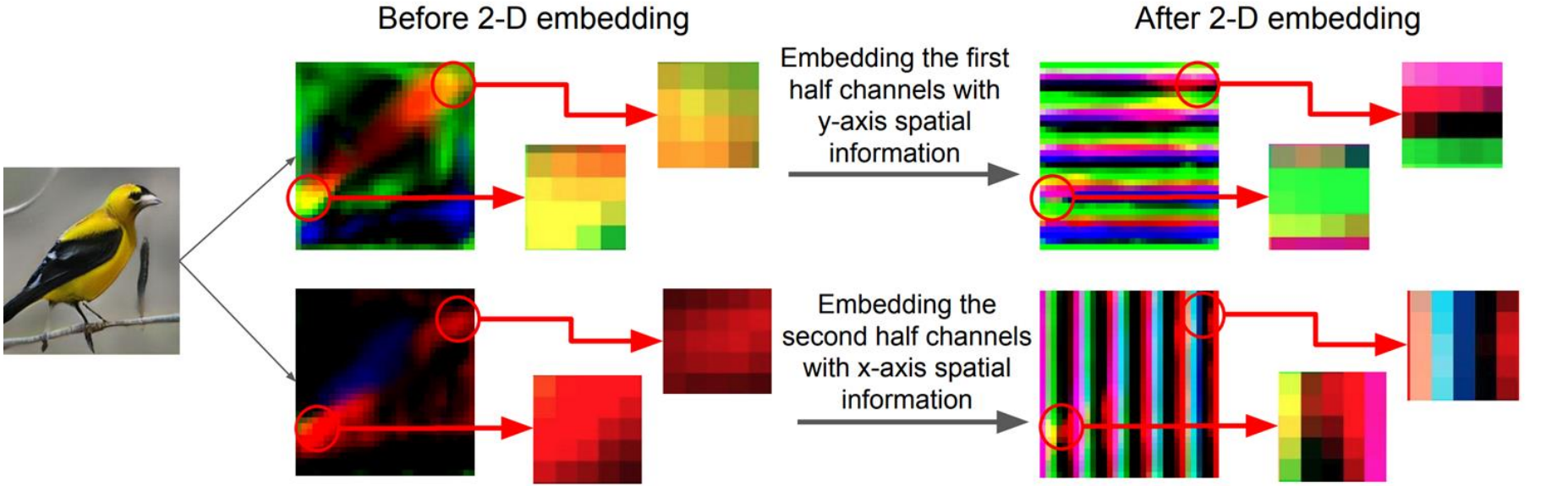
Methods: Removing the Stacked Generator



Methods: Transformer for Image Feature-Map



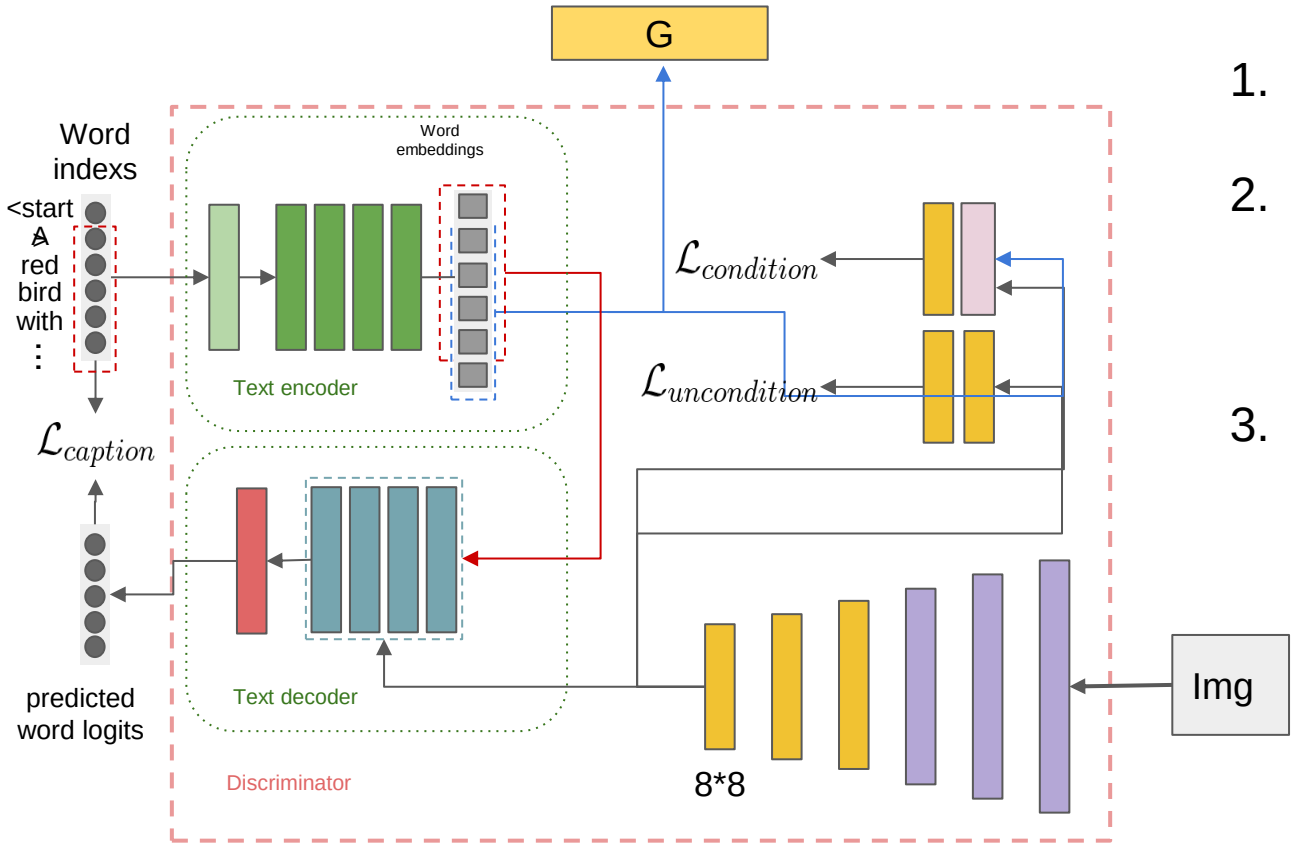
Methods: 2-D Positional Encoding



(a) Features at faraway regions can have similar values, thus hard to be distinguished by attention

(b) Features at faraway regions no longer have similar values after positional embedding, thus will be treated differently by attention.

Methods: Discriminator Objectives



1. Vanilla GAN loss
2. Image captioning loss for the text encoder/decoder
3. Image and text matching conditional loss

Methods: Annealing Image--Text Matching Loss

A small bird that is grayish brown with a striking blue color on the head, wing and breast area.

This bird has a yellow crown, a rounded breast, and grey wings.

Epoch 20 Epoch 50 Epoch 100 Epoch 150 Epoch 200

$$
\begin{aligned}
s_{\text{pivot}} &= \text{detach}(\mathbb{E}[D_{\text{c}}(I_{\text{match}}, Enc(T^{\text{real}}))]); \\
\mathcal{L}_{\text{cond}} &= \mathbb{E}[\min(0, 1 - D_{\text{c}}(I_{\text{match}}, Enc(T^{\text{real}})))] \\
&+ \mathbb{E}[\min(0, 1 + D_{\text{c}}(I_{\text{mismatch}}, Enc(T^{\text{real}})))] \\
&+ \mathbb{E}[\min(0, -s_{\text{pivot}} \times p + D_{\text{c}}(I_{\text{fake}}, Enc(T^{\text{real}})))].
\end{aligned}
$$

Methods: Generator Loss

$$\mathcal{L}_{\mathrm{caption-g}} = -\sum_{\mathrm{k}=1} \log(P_{\mathrm{k}}(T_{\mathrm{k}}^{\mathrm{real}}, D_{\mathrm{f}}(G(z, Enc(T^{\mathrm{real}}))))); \tag{5}$$

$$\mathcal{L}_{\mathrm{uncond-g}} = -\mathbb{E}[\log(D_{\mathrm{u}}(G(z, Enc(T^{\mathrm{real}}))))]; \tag{6}$$

$$\mathcal{L}_{\mathrm{cond-g}} = -\mathbb{E}[D_{\mathrm{c}}(G(z, Enc(T^{\mathrm{real}})), Enc(T^{\mathrm{real}}))]. \tag{7}$$

Methods: More Controllable Generator by Dropping Sentence Level Embeddings

AttnGAN results

This is a small bird with white belly and a **short** beak.

This is a small bird with white belly and a **long** beak.

z1

A yellow bird with a yellow belly and **brown** wings.

A yellow bird with a yellow belly and **black** wings.

z2

The bird has a beautiful red body and grey **wings**.

The bird has a beautiful red body and grey **head**.

z3

Methods: More Controllable Generator

This bird has wings that are black and has a red belly.

This bird has wings that are brown and has a yellow belly.

This bird has wings that are blue and has a white belly.

This bird is blue with green and red and has a very short beak.

This bird has a red crown with black wings and a long black beak.

z1 z2 z3 z4

Synthesis of TIME with short sentences

This is a green bird with a brown crown and a white and black belly.

This small bird has black eyerings and cheek patches, with a white breast.

The bird has a white belly, brown back and a small bill, with black patches on the back.

A small bird with a small beak compared to its head size, that is covered in red, brown and white.

A brown bird with white and brown feathered belly and breast, and small pointed yellow bill.

z5 z6 z7 z8

Synthesis of TIME with long sentences

Methods: Why the Joint Training works?

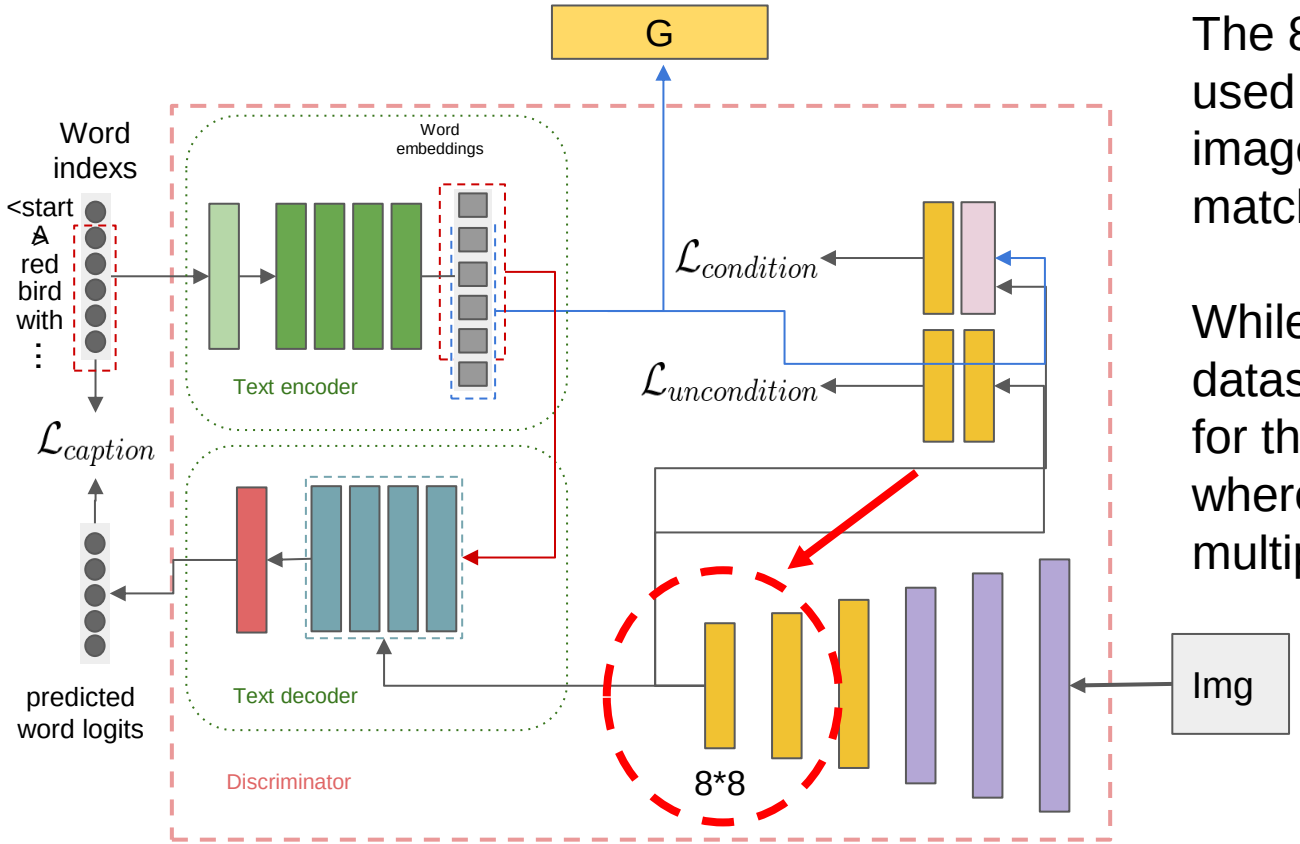
Prior works

1. LSTM model, suffer from gradient vanishing.
2. In prior text-to-image model: the word embeddings that input to the Generator are discrete vectors.
3. The text modules are trained with multiple signals , from both D and G.

Ours

1. Transformer is more robust and easier to train.
2. By using transformer, word embeddings for the same word are different if put in different sentence.
3. We train the text encoder and decoder only on image captioning task. The word embeddings are not trained on making G generate better image.

Methods: Performance on MS-COCO



The 8x8 image feature-map is used for image captioning and image-text consistency matching on Discriminator.

While it works well on CUB bird dataset, it is not a good choice for the MS-COCO dataset, where images containing multiple small objects.

Experiments: Performance on MS-COCO

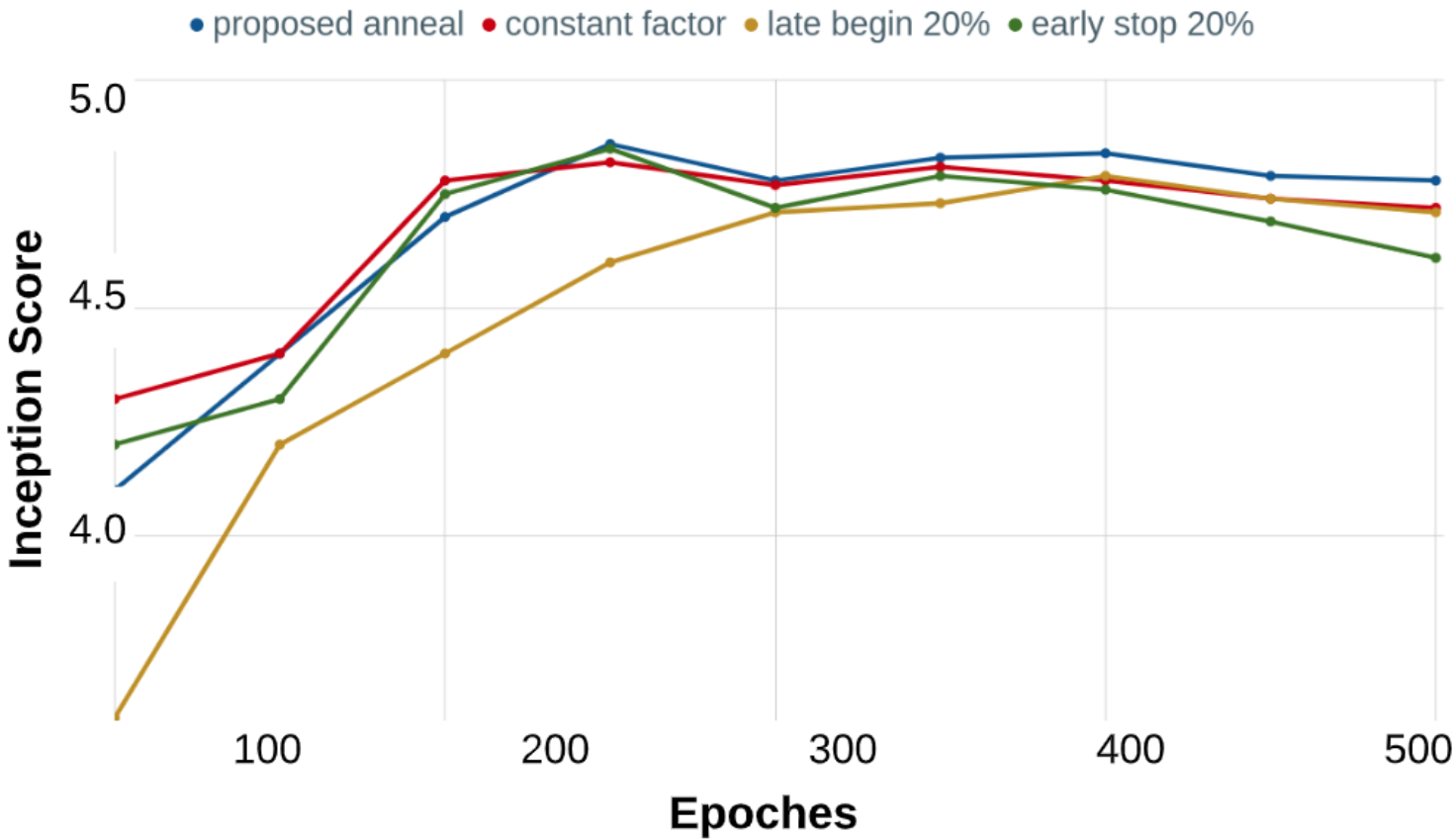
	Inception Score ↑	R-precision ↑
AttnGAN	4.36 ± 0.03	67.82 ± 4.43
Tf-h1-l1	4.38 ± 0.06	66.96 ± 5.21
Tf-h4-l1	4.42 ± 0.06	68.58 ± 4.39
Tf-h4-l2	$\mathbf{4.48} \pm 0.03$	$\mathbf{69.72} \pm 4.23$
Tf-h4-l4	4.33 ± 0.02	67.42 ± 4.31
Tf-h8-l4	4.28 ± 0.03	62.32 ± 4.25

CUB

	AttnGAN	Tf-h1-l1	Tf-h4-l1	Tf-h4-l2	Tf-h4-l4	Tf-h8-l4
Inception Score ↑	25.89	26.58	26.42	27.48	27.85	27.153
R-precision ↑	83.53	86.46	88.58	89.72	89.57	88.32

MS-COCO

Experiments: The Annealing Schedule of the Loss



Experiments: Text-to-Image Performance

	CUB-bird	MS-COCO
FID	14.3	31.14
Inception Score	4.91	30.85
SOA-C	N/A	32.78
R-precision	71.57	89.57

Experiments: Qualitative Results from our Model


A people are skiing on a snowy ground.


A large amount of vegetables and animals on a table.


A group of black cows standing near a woman.


A plate of hotdog and carrots and fries.

(b) TIME image-captioning results on MS-COCO with errors


A train driving down a empty road.


Profile of the head and neck of a giraffe.


A young boy that is playing at front of a cake.


Baseball player is swinging a baseball to a game.

(c) TIME image-caption results on MS-COCO without errors


A parking lot with many different motor bikes.


This is a picture of a kitchen with cabinets large appliances and dishes.


A few cows grazing grass on a farm .


A large crowd of people riding horses walks along a trail.


Two people standing on a ski slope looking down the hill.


A team of baseball players playing a game of baseball.


A tall clock tower with trees in front of it .


Broccoli zucchini and peppers are mixed together on a plate.

(d) TIME text-to-image results on MS-COCO

Thanks for watching!

Contact me at: bingchen.liu@rutgers.edu

TIME: Text and Image Mutual-Translation Adversarial Networks

Bingchen Liu, Kunpeng Song, Yizhe Zhu, Gerard de Melo, Ahmed elgammal

Rutgers University

Background

Recent years have witnessed substantial progress in the text-to-image task owing largely to the success of deep generative models. Conditional GANs by Reed et al. first synthesized plausible images from text descriptions. Stack-GAN and AttnGAN then took the quality to the next level. MirrorGAN incorporates a pre-trained text re-description RNN to better align the images with the given texts. ControlGAN employs a channel-wise attention in G, and SDGAN includes a contrastive loss to strengthen the image–text correlation.

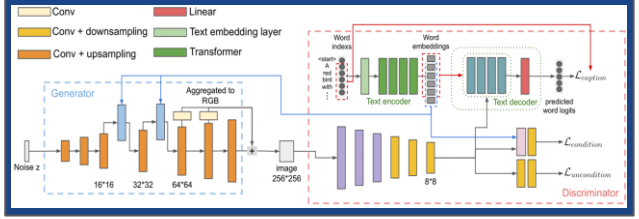
These works all depend on: 1. a pre-trained text encoder for word and sentence embeddings; 2. an additional image encoder to as- certain image–text consistency. Can we replace the extra CNN in the DAMSM module and the extra CNN in the DAMSM module with elegantly trained end-to-end generative model?

We propose Text and Image Mutual-Translation Adversarial Networks (TIME), a lightweight but effective model that jointly learns a T2I generator G and an image captioning discriminator D under the Generative Adversarial Network framework. While previous methods tackle the T2I problem as a uni-directional task and use pre-trained language models to enforce the image–text consistency, TIME requires neither extra modules nor pre-training. We show that the performance of G can be boosted substantially by training it jointly with D as a language model. Specifically, we adopt Transformers to model the cross-modal connections between the image features and word embeddings, and design an annealing conditional hinge loss that dynamically balances the adversarial learning.

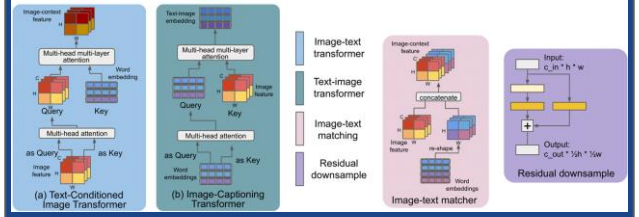
Method

The upper panel in Fig.1 shows the overall structure of TIME, consisting of a Text-to-Image Generator and an Image-Captioning Discriminator D. The main components are:

- Aggregated Generator: To trim the model size of the StackGAN structure, we present the design of an aggregated G. G still yields RGB outputs at multiple resolutions, but these RGB outputs are rescaled and added together as a single aggregated image output. Therefore, only one Discriminator is needed.



- Text-conditioned-Image Transformer: as illustrated in Fig. 2-(a). The Transformer replicates the attention module in a halti-head manner, thus adding more flexibility for eachimage region to account for multiple words. We stacked the attention layers in a residual structure in a multi-layer manner for better performance by provisioning multiple attention layers and recurrently revisingthe learned feature.



- Image-Captioning Discriminator: We treat the text encoder and text decoder as a part of D. Text encoder is a Transformer that maps the word indices into the embeddings while adding contextual information to them. Decoder is a Transformer decoder that performs image captioning by predicting the next word's probability from the masked word embeddings and the image features.
- Image-Captioning Transformer: Symmetric to Text-Conditioned Image Transformer, the inverse operation, which extracts text information from the image and converts it into text embeddings. It is designed to be a 4-layer-4-head transformer.

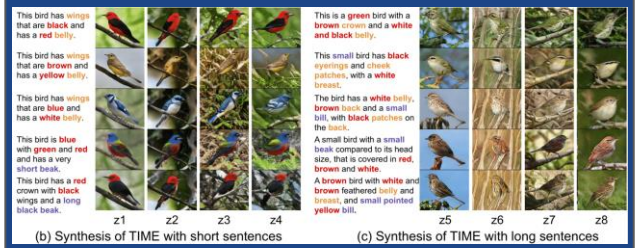
Objective function: The loss function of contains the adversarial loss (conditioned and unconditioned), and an annealing Image-Text matching loss. Please refer to the paper for details.

Results

TIME demonstrates competitive performance on MS-COCO and CUB datasets with the new state-of-the-art IS and FID. Unlike the other models that require a well pre-trained language module and an Inception-v3 image encoder, TIME itself is sufficient to learn the relationships between image and language. The results are among the top performers on both datasets.

		StackGAN	AttnGAN	ControlGAN	MirrorGAN	DMGAN	TIME	Real-Image
CUB	Inception Score ↑	3.82 ± 0.06	4.36 ± 0.03	4.51 ± 0.06	4.56 ± 0.05	4.71 ± 0.02	**4.91** ± 0.03	5.04
	FID ↓	N/A	23.98	N/A	N/A	16.09	**14.3**	0
	R-precision ↑	10.37 ± 5.88	67.82 ± 4.43	69.33 ± 3.21	69.58 ± 4.39	72.31 ± 0.91	71.57 ± 1.2	N/A
COCO	Inception Score ↑	8.45 ± 0.03	25.89 ± 0.47	24.06 ± 0.6	26.47 ± 0.4	30.49 ± 0.5	**30.85** ± 0.7	36.5
	FID ↓	N/A	35.49	N/A	N/A	32.64	**31.14**	0
	R-precision ↑	N/A	83.53 ± 0.43	82.43 ± 2.21	84.21 ± 0.39	91.87 ± 0.28	89.57 ± 0.9	N/A
	SOA-C ↑	N/A	25.88	25.64	27.52	**33.44**	32.78	74.97

We follow the same convention as in previous T2I works to split the training/testing set. Experiment results of CUB dataset are shown in the image below. They are cases where changing just a single word leads to unpredictably large changes in the image. We adopt the Transformer as the text encoder, where the word embeddings already come with contextual information. So our model does not need an aggregated sentence embedding, giving it the capability of accurate local editing.



We apply PCA on the learned word embeddings and visualize them in the image below. Words with similar meanings reside close to each other. "Large" ends up close to "red", as the latter often applies to large birds, while "small" is close to "brown" and "grey", which often apply to small birds. This also proves the model successfully learned meaningful word vectors. This helps the Generator to converge faster. As shown in image below, our generator can learn a good semantic visual translation at very early iterations.

Conclusions

In this paper, we propose the Text-and-Image Mutual-translation adversarial Network (TIME), a unified framework trained with an adversarial schema that accomplishes both the text-to-image and image-captioning tasks. While previous works in the T2I field require pre-training several supportive modules, TIME achieves the new state-of-the-art T2I performance without pre-training. The Joint process of learning in TIME bridges the gap between the visual and language domains, unveiling the potential of mutual translations between the two modalities within a single model.