Releasing the Capacity of GANs in Non-Autoregressive Image Captioning

Da Ren, Qing Li*
Department of Computing, The Hong Kong Polytechnic University
Hong Kong, China
{csdren, csqli}@comp.polyu.edu.hk

Abstract

Building Non-autoregressive (NAR) models in image captioning can fundamentally tackle the high inference latency of autoregressive models. However, existing NAR image captioning models are trained on maximum likelihood estimation, and suffer from their inherent multi-modality problem. Although constructing NAR models based on GANs can theoretically tackle this problem, existing GAN-based NAR models obtain poor performance when transferred to image captioning due to their incapacity of modeling complicated relations between images and text. To tackle this problem, we propose an Adversarial Non-autoregressive Transformer for Image Captioning (CaptionANT) by improving performance from two aspects: 1) modifying the model structure so as to be compatible with contrastive learning to effectively make use of unpaired samples; 2) integrating a reconstruction process to better utilize paired samples. By further combining with other effective techniques and our proposed lightweight structure, CaptionANT can better align input images and output text, and thus achieves new state-of-the-art performance for fully NAR models on the challenging MSCOCO dataset. More importantly, CaptionANT achieves a 26.72× speedup compared to the autoregressive baseline with only 36.3% the number of parameters of the existing best fully NAR model for image captioning.

Keywords: Image Captioning, Non-Autoregressive Models, GANs

1. Introduction

Different with autoregressive (AR) models which generate tokens one-by-one and thus have high decoding latency, non-autoregressive (NAR) models obtain all tokens in parallel and provide a more efficient method to obtain the results (Qian et al., 2021; Ghazvininejad et al., 2019; Huang et al., 2022b). This feature makes NAR models ideal for use in scenarios demanding low latency. However, comparing with the rapid development of NAR models in machine translation (Xiao et al., 2023), its progress in image captioning is relatively slow. Recent study directly enhances performance by significantly sacrificing decoding efficiency (Fei, 2021; Yan et al., 2021).

Exiting work constructs NAR image captioning models based on Maximum Likelihood Estimation (MLE) (Gao et al., 2019; Fei, 2020; Guo et al., 2020), which meets obvious obstacles in their developments. First, MLE-based NAR models can learn the marginal distributions of different candidates, but lose word dependencies and remain non-negative lower bounds in the KL divergence between the learned distributions and real distributions (Huang et al., 2022a). Thus, these models tend to generate ungrammatical sentences by mixing words in different candidates, which is known as the multi-modality problem (Gu et al., 2018). Secondly, the difficulties in the alignment between images and text will cause greater errors in the learned marginal distributions, thereby exacerbating the multi-modality problem by mixing irrelevant candidates.

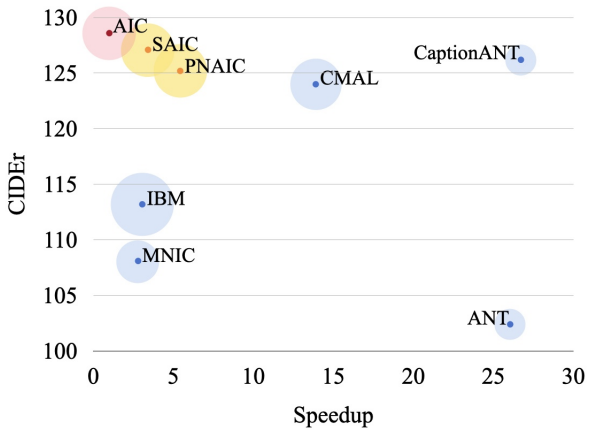

Figure 1: The Performance of Compared Models. The red, yellow and blue points indicate AR, SAR and NAR models, respectively. The area indicates the number of parameters.

Different with MLE, which is inherently incompatible with NAR models, Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) denote a more promising method. Their learned distributions can theoretically converge to the real distributions with one single forward pass (Goodfellow et al., 2014). It exactly fits the needs of NAR models. In text-to-image generation, GANs have been

*Corresponding author: Qing Li

LREC-COLING 2024, pages 13906–13918

demonstrated to be an effective method. They can generate high quality images with much lower latency (Sauer et al., 2023; Kang et al., 2023). However, their potentials in image-to-text generation have not been explored yet.

The main obstacle of adopting GANs in text generation comes from the non-differentiable sampling operation in the generator, which prevents the gradient of the discriminator from being passed to the generator. Recently, Ren and Li (2022) introduce a representation modeling method to tackle this problem by removing the sampling operation during training. It is later extended to NAR models for incomplete information scenarios (Ren and Li, 2023). This model, which is denoted as Adversarial Non-autoregressive Transformer (ANT), obtains poor performance when transferred to image captioning (as shown in Figure 1). ANT is designed for the scenarios with relatively simple input conditions (e.g., class labels). In image captioning, however, the input images are highly diverse, and ANT becomes incapable of building complicated relations between images and text.

In this paper, we release the capacity of GANs in image captioning by proposing an Adversarial Non-autoregressive Transformer for Image Captioning (CaptionANT). To enable the model to build more complicated relations, the discriminator structure in the previous work (Ren and Li, 2023) is modified to be compatible with contrastive learning, so CaptionANT can better align images and text by effectively making use of unpaired samples. In addition, we integrate a reconstruction process to further boost model performance by better making use of paired samples. During the reconstruction process, the key challenge comes from the ambiguous reconstruction target led by the one-to-many mapping relations in image captioning. We tackle this problem by integrating part of target sentences into the input so as to have clearer reconstruction targets. By further combining with other effective techniques (like feature ensemble and the truncation trick) and our proposed lightweight structure, CaptionANT achieves new state-of-the-art performance for fully NAR models on the challenging MSCOCO dataset with much higher speedup and lower parameter number (as shown in Figure 1). The contributions of this paper can be summarized as follows:

- Considering the limitations of MLE-based NAR image captioning models, we propose a GAN-based NAR model—CaptionANT. We redesign the model structure and incorporate contrastive learning in CaptionANT. It can effectively make use of unpaired samples to model complicated relations between images and text. To the best of our knowledge, CaptionANT is the first GAN-based NAR model in image captioning.
- We further propose to incorporate a reconstruction process into the training stage of language GANs based on representation modeling methods. It can further improve model performance by better utilizing paired samples. For the ambiguous reconstruction targets led by the one-to-many mapping relations, we propose to integrate part of target information into the input so to have clear reconstruction targets.
- By further combining with other effective techniques (like feature ensemble and the truncation trick) and our proposed lightweight structure, CaptionANT achieves new state-of-the-art performance for fully NAR models with lower parameter number and faster speed.

This paper is structured as follows. In Section 2, we give an overview to non-autoregressive image captioning models and generative adversarial networks. In Section 3, we elaborate the details about our proposed model, CaptionANT. We introduce our experiments and analyze experiment results in Section 4. Finally, we draw a conclusion in Section 5.

2. Related work

2.1. Non-Autoregressive Image Captioning Models

Non-Autoregressive (NAR) models are first proposed in machine translation (Gu et al., 2018) and later extended to other areas (Ren et al., 2019). Although these MLE-based models can learn the marginal distributions of candidates, they tend to mix words in different candidates due to the loss of word dependencies (Huang et al., 2022a). Furthermore, the challenge of aligning image input and text output brings inaccurate marginal distributions, so these NAR models have additional difficulties in image captioning.

Early study of NAR models in image captioning adopts iterative-based methods to accelerate inference (Gao et al., 2019; Fei, 2020). To maintain sentence-level consistent, Guo et al. (2020) integrate the counterfactuals-critical multi-agent learning into the training objective. Their model is still the best fully NAR model since recently proposed models are either based on semi-autoregressive (SAR) structures (Fei, 2021; Yan et al., 2021) or iterative-based methods (Luo et al., 2023). These models enhance performance by significantly sacrificing the inference speed. The study about building fully NAR models, which can generate captions in higher quality and lower latency, has recently come to a halt.

2.2. Generative Adversarial Networks

Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) are widely adopted in image generations (Karras et al., 2019, 2020; Sauer et al., 2022; Zhu et al., 2017). Even competing with the powerful diffusion models (Nichol et al., 2022; Saharia et al., 2022) and autoregressive models (Zhang et al., 2021; Yu et al., 2022), GANs get comparable results with much faster speed in text-to-image generation (Sauer et al., 2023; Kang et al., 2023). However, their potentials in image-to-text generation have not been explored yet.

The main difficulty of adopting GANs in text generation is from the non-differentiable sampling operation when obtaining specific words. This sampling operation stops the gradients from the discriminator passing to the generator. Early study adopts *REINFORCE* (Williams, 1992) or *continuous relaxations* (Jang et al., 2017) to tackle this problem (de Masson d'Autume et al., 2019; Nie et al., 2019). These methods are either high variance or biased (Lin et al., 2020; de Masson d'Autume et al., 2019). Thus, Ren and Li (2022) introduce a representation modeling method to tackle this problem. This method first converts words into representations, and then prompts the generator to recover these representations. The representations are then fed into the discriminator directly. This method avoids the sampling operation during training, so the gradients from the discriminator can be passed through to the generator directly. Recently, this method is further extended to building NAR models for incomplete information scenarios (Ren and Li, 2023). However, it fails to generate high quality results when transferred to image captioning due to its limited capacity in aligning diverse images and text.

3. Model

To allow the gradients from the discriminator to be passed to the generator directly, we adopt the representation modeling method (Ren and Li, 2022), which can avoid the non-differentiable sampling operation during training. More specifically, we first adopt a model, which is denoted as Mapper in this work, to map words into representations. Then, the generator is trained to recover these representations under the guidance of the discriminator. Both the representations from the mapper and the generator are fed into the discriminator as input, and the discriminator needs to identify whether the input is from the mapper or not.

The general structure of CaptionANT is shown in Figure 2. As described above, there are three different models in CaptionANT: Mapper, Discriminator and Generator. All these three models adopt Transformer (Vaswani et al., 2017) as backbones to support highly parallel computation.

3.1. Mapper

The mapper needs to map words into representations. It is trained before the generator and the discriminator. The training process of the mapper is described in the blue dashed box of Figure 2. A certain number of words in a sentence are randomly masked or replaced, and the mapper is trained to reconstruct the original input. We follow the settings in the previous work (Ren and Li, 2023), and incorporate the idea of variational autoencoder (VAE) (Kingma and Welling, 2014) into the training objective. More specifically, after obtaining the mean μ_{x_i} and standard deviation σ_{x_i} for each word x_i, the mapper first adopts reparameterization trick to obtain hidden representations $\mathbf{z}'_i = \mu_{x_i} + \sigma_{x_i} \cdot \mathcal{N}(0, 1)$, and then uses the following objective to train the model:

$$\begin{aligned} L_A = -\mathbb{E}_{\mathbf{z}'_i \sim q(\mathbf{z}'_i|x_i)}(logp(x_i|\mathbf{z}'_i)) \\ +KL(q(\mathbf{z}'_i|x_i)||p(\mathbf{z}'_i)) \end{aligned} \quad (1)$$

where $\mathbf{z}'_i$ is transformed into words by a linear layer F_{LT}. The vector μ_{x_i} will be regarded as the representation of x_i and fed into the discriminator.

This training objective provides a dense and continuous representation space, so representations slightly away from the central point can still be mapped into correct words (Ren and Li, 2023).

3.2. Discriminator

3.2.1. Structure

The discriminator is consisted of a stack of Transformer blocks. Different with the discriminator in the work of Ren and Li (2023), we do not observe improvements from the look-ahead mask, so we remove it and the input in different positions can consider each other directly.

One key challenge in building the discriminator is how to incorporate conditions into the model. The previous work (Ren and Li, 2023) feeds the condition representation as input. This structure is also adopted in other GAN-based text-to-image generation models (Reed et al., 2016; Zhang et al., 2017; Li et al., 2019). However, it only considers one pair of mismatched samples at a time and can not effectively make use of them to build more accurate alignments between images and text.

To tackle this problem, we separate the condition representation from the input. Instead, we map input text to the same space as the condition representation, so it can measure the correlation between the two by calculating dot product. This modeling method can efficiently utilize unpaired

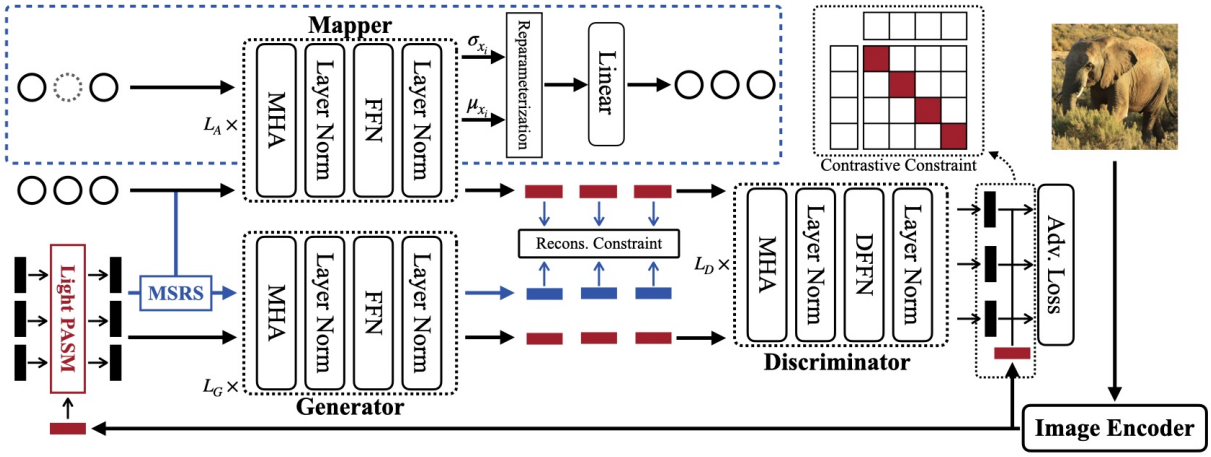

Figure 2: General Structure of CaptionANT.

samples through contrastive learning (which will be introduced in Section 3.2.2). The detailed calculation of this modeling methods is described as follows:

$$\begin{aligned}\hat{\mathbf{h}}_i^{(l)} &= LN(MHA(\mathbf{h}_i^{(l-1)}) + \mathbf{h}_i^{(l-1)}) \\ \mathbf{h}_i^{(l)} &= LN(DFFN(\hat{\mathbf{h}}_i^{(l)}) + \hat{\mathbf{h}}_i^{(l)}) \\ \tilde{\mathbf{h}}_\mathbf{i} &= W_h \mathbf{h}_i^{(L_D)} + b_h \\ \mathbf{y}_i &= \tilde{\mathbf{h}}_\mathbf{i} \cdot \hat{\mathbf{c}}^\intercal\end{aligned} \tag{2}$$

where $MHA(\cdot)$ is the multi-head attention mechanism (here is the self-attention, where query, key and value are the same), $LN(\cdot)$ is layer normalization, $\hat{\mathbf{c}}$ is the normalized image representation ($\hat{\mathbf{c}} = \mathbf{c}/||\mathbf{c}||$, and $\mathbf{c}$ is the image representation provided by the image encoder) and L_D is the layer number. $DFFN(\cdot)$ is the dependency feed forward network (Ren and Li, 2023), which is calculated as follows:

$$\begin{aligned}\hat{\mathbf{g}}_i &= GELU(\hat{\mathbf{h}}_i W_g + b_g) \\ \mathbf{g}_i &= \hat{\mathbf{g}}_{i-1} W_l + \hat{\mathbf{g}}_i W_k + b_o\end{aligned} \tag{3}$$

where $\hat{\mathbf{h}}_i$ is the input of DFFN, and $\mathbf{g}_i$ is the output, which will be incorporated into the calculation of Eq 2. DFFN directly models the relations between $\hat{\mathbf{g}}_{i-1}$ and $\hat{\mathbf{g}}_i$, so it can strengthen the dependency modeling capacity of the discriminator in the unstable training of GANs (Ren and Li, 2023).

3.2.2. Training Objective

We adopt Wasserstein distance (Arjovsky et al., 2017) as the training objective:

$$\begin{aligned}L_{AdvD} = &- \mathbb{E}_{x\sim P_x}[D(M(x), \mathbf{c})] \\ &+ \mathbb{E}_{\mathbf{z}\sim P_\mathbf{z}}[D(G(\mathbf{z}), \mathbf{c})]\end{aligned} \tag{4}$$

where $M(\cdot)$ is the mapper, $\mathbf{c}$ is the condition representation obtained by the image encoder, $D(\cdot)$ and $G(\cdot)$ are the discriminator and the generator, respectively. We adopt Lipschitz penalty (Petzka et al., 2018) to stabilize the training process.

Contrastive Constraint To fully make use of the advantages of our discriminator structure, we further integrate a contrastive constraint into the training objective to regularize the model by considering unpaired samples effectively. We first obtain the representation of the k-th sentence $\mathbf{H}_k$ by calculating the mean of $\tilde{\mathbf{h}}_\mathbf{i}$ in different timesteps. Then, the contrastive constraint is calculated as follows:

$$C_d = -\tau \frac{exp(\mathbf{H}_k \cdot \hat{\mathbf{c}}^\intercal/\tau)}{\sum_{j=1} exp(\mathbf{H}_j \cdot \hat{\mathbf{c}}^\intercal/\tau)} \tag{5}$$

where $\hat{\mathbf{c}}$ is the normalized condition representation. We obtain the negative samples from two different sources: 1) the real but mismatched sentences in the same batch; 2) the synthetic sentences given by the generator with the same batch of condition representations. The real but mismatched sentences can help the model quickly regularize its representations in the early training, while the synthetic sentences can further boost model performance when the generator begins to generate real-like sentences.

Incorporating the contrastive constraint, the complete training objective of the discriminator is:

$$L_D = L_{AdvD} + \lambda_d \cdot C_d \tag{6}$$

where λ_d is a hyper-parameter which can adjust the importance of the contrastive constraint.

3.3. Generator

3.3.1. Structure

The generator is constructed based on Transformer (Vaswani et al., 2017). The input is a trainable matrix. The vectors obtained by the final Transformer block will be the word representations after

a linear transformation. During training, these representations (denoted as r_i) will be fed into the discriminator, and the discriminator will guide the generator to obtain the representations following same distributions with μ_{x_i} from the mapper. During inference, r_i will be transformed back into words with the same linear layer F_{LT} of the mapper.

Feature Ensemble An effective method to incorporate latent vectors plays a key role in the performance of the generator. Previous work (Ren and Li, 2023; Lee et al., 2022) calculates shift and scale vectors for the normalized input based on latent vectors. We further enhance model performance by adopting feature ensemble which can provide images features from two representation spaces. It is described as follows:

$$\begin{pmatrix} \mathbf{s}'_1 \\ \mathbf{s}'_2 \\ \vdots \\ \mathbf{s}'_N \end{pmatrix} = \mathbf{F}^{\mathbf{1}}_{\mathbf{M}}(\mathbf{z}_1) + \mathbf{F}^{\mathbf{2}}_{\mathbf{M}}(\mathbf{z}_2) \tag{7}$$

$$\mathbf{s}_i = \gamma(\mathbf{s}'_i) \circ LN(X^g_i) + \beta(\mathbf{s}'_i)$$

where X^g_i is the trainable input matrix, $\gamma(\cdot)$ and $\beta(\cdot)$ are linear layers after GELU (Hendrycks and Gimpel, 2016), and $\mathbf{s}_i$ will be fed into a set of Transformer blocks as input. $\mathbf{z}_1$ and $\mathbf{z}_2$ are the concatenations of random noises and image features extracted by two different models. For the random noise, we adopt the **truncation trick** (Brock et al., 2019) which samples the noise in a truncated distribution during inference.

The transformation modules $\mathbf{F}^{\mathbf{1}}_{\mathbf{M}}(\cdot)$ and $\mathbf{F}^{\mathbf{2}}_{\mathbf{M}}(\cdot)$ are in a same structure but with independent parameters. The design of the transformation modules will directly influence the performance, and a detailed discussion is conducted in the following.

Light Position-Aware Self-Modulation Different methods adopt different transformation modules. Self-modulation (Chen et al., 2019) uses same layers at different positions, and the obtained representations at each position are thus too similar to recover the diverse word representations at different positions (Ren and Li, 2023). Ren and Li (2023) tackle this problem by proposing a Position-Aware Self-Modulation (PASM) which adopts unique layers at different positions to obtain diverse representations.

This method, however, has independent layers for each position. It causes a dramatic increase in the number of model parameters, which we find is not necessary. Instead, we propose and adopt a Light Position-Aware Self-Modulation (Light PASM). The transformation module ($\mathbf{F}^{\mathbf{1}}_{\mathbf{M}}$ and $\mathbf{F}^{\mathbf{2}}_{\mathbf{M}}$ in Eq. 7)

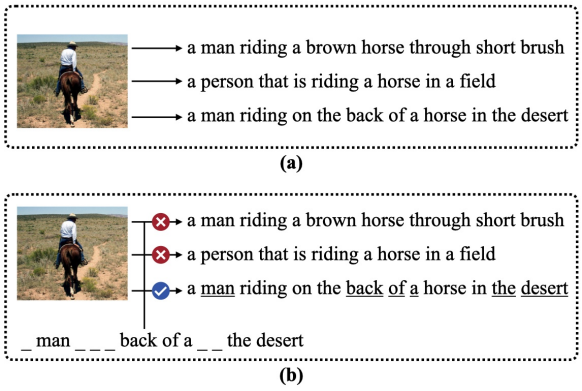

Figure 3: Effectiveness of Masked Sentence Representation Shift (MSRS).

in our proposed model is:

$$\begin{pmatrix} \hat{\mathbf{s}}_1 \\ \hat{\mathbf{s}}_2 \\ \vdots \\ \hat{\mathbf{s}}_{\frac{N}{2}} \end{pmatrix} = \begin{pmatrix} W_1 \\ W_2 \\ \vdots \\ W_{\frac{N}{2}} \end{pmatrix} \cdot \hat{\mathbf{z}} + \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_{\frac{N}{2}} \end{pmatrix} \tag{8}$$

$$\begin{pmatrix} \hat{\mathbf{s}}_{\frac{N}{2}+1} \\ \hat{\mathbf{s}}_{\frac{N}{2}+2} \\ \vdots \\ \hat{\mathbf{s}}_N \end{pmatrix} = W' \cdot \begin{pmatrix} \hat{\mathbf{s}}_1 \\ \hat{\mathbf{s}}_2 \\ \vdots \\ \hat{\mathbf{s}}_{\frac{N}{2}} \end{pmatrix} + b' \tag{9}$$

where $\hat{\mathbf{z}}$ is the $\mathbf{z}_1$ or $\mathbf{z}_2$ in Eq. 7. Light PASM first obtains the hidden representations of the previous half position with unique linear layers. Then, another linear layer is adopted to get the remaining half of the representations. This method can maintain the diversity of representations between different positions while significantly reducing the parameter number.

Different with existing NAR image captioning models (Guo et al., 2020; Fei, 2021) which first use an encoder to process image features and then generate sentences with a decoder, the generator in CaptionANT directly transforms image features into sentences, so it has a lighter and more efficient structure.

3.3.2. Training Objective

Corresponding to the discriminator, the adversarial training objective of the generator is:

$$L_{AdvG} = -\mathbb{E}_{\mathbf{z}\sim P_{\mathbf{z}}}[D(G(\mathbf{z}), \mathbf{c})] \tag{10}$$

In addition, we also adopt the following constraints to boost its performance.

Contrastive Constraint Similar to the discriminator, we adopt a contrastive constraint to better

Model	BLEU-1	BLEU-4	METEOR	ROUGE	SPICE	CIDEr	#Param.	Speedup
Autoregressive Models								
Up-Down (Anderson et al., 2018)	79.8	36.3	27.7	56.9	21.4	120.1	-	-
M2-T (Cornia et al., 2020)	80.8	39.1	29.2	58.6	22.6	131.2	-	-
$\mathcal{A}^2$-Transformer (Fei, 2022)	**81.5**	**39.8**	**29.6**	**59.1**	**23.0**	**133.9**	-	-
AIC (bw=1)	80.3	38.9	28.7	58.5	22.4	127.1	54.9M	1.22×
AIC (bw=3)	80.4	39.2	28.8	58.6	22.5	128.6		1.00×
Semi-Autoregressive Models								
PNAIC (Fei, 2021)	79.9	37.5	28.2	58.0	21.8	125.2	54.9M	5.43×
SAIC (Yan et al., 2021)	**80.3**	**38.4**	**29.0**	**58.1**	**21.9**	**127.1**		3.42×
Non-Autoregressive Models								
MNIC (Gao et al., 2019)	75.4	30.9	27.5	55.6	21.0	108.1	36.0M	2.80×
IBM (Fei, 2020)	77.2	36.6	27.8	56.2	20.9	113.2	77.0M	3.06×
CMAL (Guo et al., 2020)	80.3	37.3	28.1	58.0	21.8	124.0	50.1M	13.90×
CaptionANT	**80.8**	**38.0**	**28.7**	**58.7**	**22.5**	**126.2**	18.2M	26.72×

Table 1: Evaluation Results on the "Karpathy" Split of MSCOCO Dataset.

align input images and output text.

$$C_g = -\tau \frac{exp(\mathbf{H}'_k \cdot \hat{\mathbf{c}}^\intercal/\tau)}{\sum_{j=1} exp(\mathbf{H}'_j \cdot \hat{\mathbf{c}}^\intercal/\tau)} \tag{11}$$

where $\mathbf{H}'_k$ is the mean of $\tilde{\mathbf{h}}_\mathbf{i}$ from the discriminator in different timesteps, and the negative samples are the captions generated based on the unpaired conditions in the same batch.

Reconstruction Constraint Reconstruction constraint has been adopted to stabilize the training and enhance model performance in image GANs (Zhu et al., 2017). It provides a more effective way to utilize paired samples. However, how to incorporate reconstruction constraint in language GANs, which are based on the representation modeling method, has not been explored yet. The key challenge is from the diverse words in the same positions among different candidates. We give an example in Figure 3 (a). If the model is trained to fit all candidates together, it will try being close to the diverse word representations in different candidates and finally degenerate to learn mean values instead of specific representations.

We tackle this problem by proposing a **Masked Sentence Representation Shift (MSRS)**. When calculating the reconstruction constraint term, the input representations $\mathbf{s}_i$ obtained by Eq. 7 are added with shift vectors as follows:

$$\begin{aligned} \mathbf{e}_i &= Emb(x_i) + pos_i \\ \hat{\mathbf{e}}_i &= Mask(\mathbf{e}_i, \rho) \\ \dot{\mathbf{s}}_i &= \omega \circ MHA(\mathbf{s}_i, \hat{\mathbf{e}}_i, \hat{\mathbf{e}}_i) \\ \hat{\mathbf{s}}_i &= \mathbf{s}_i + \dot{\mathbf{s}}_i \end{aligned} \tag{12}$$

where x_i is the i-th word of the sentence, $Emb(\cdot)$ is an embedding layer, pos_i is the positional encoding for the i-th position, ρ is the mask rate, $MHA(query, key, value)$ is the multi-head attention, ω is a trainable vector which can directly control the scale of $\dot{\mathbf{s}}_i$. This process is shown in the blue path of Figure 2. It should be noted that $\hat{\mathbf{s}}_i$ is only used when calculating the reconstruction constraint, and the generator still adopts $\mathbf{s}_i$ as input when calculating the adversarial loss and generating captions in inference stage.

The effectiveness of the MSRS is described in Figure 3 (b). By providing shift vectors $\dot{\mathbf{s}}_i$, MSRS incorporates unmasked words into the input representations. This approach reduces the number of possible candidates and transforms the mapping relations from input to output to a roughly one-to-one relation. Thus, the model can learn to reconstruct specific word representations instead of ambiguous ones. The reconstruction constraint is:

$$C_r = ||\mu_{x_i} - r'_i||^2 + \lambda_s ||\dot{\mathbf{s}}_i||^2 \tag{13}$$

where μ_{x_i} is the representation of the word x_i obtained by the mapper, r'_i is the i-th word representation given by the generator, and λ_s is a hyper-parameter. The norm of $\dot{\mathbf{s}}_i$ is also minimized, so the shifted representation $\hat{\mathbf{s}}_i$ can be as close to the original input representation $\mathbf{s}_i$ as possible.

With the constraints above, the complete training objective of the generator is:

$$L_G = L_{AdvG} + \lambda_g \cdot C_g + \lambda_r \cdot C_r \tag{14}$$

where λ_g and λ_r are both hyper-parameters which can control the effects from the constraints.

4. Experiment

4.1. Experiment Setup

The MSCOCO dataset (Chen et al., 2015) is one of most popular dataset in image captioning. We adopt the widely used "Karpathy" splits (Karpathy and Fei-Fei, 2015) to conduct experiments. It contains 113,287 images for the training set, 5,000

Model	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		ROUGE-L		CIDEr	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
Up-Down (Anderson et al., 2018)	80.2	95.2	64.1	88.8	49.1	79.4	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
M2-T (Cornia et al., 2020)	81.6	96.0	66.4	90.8	51.8	82.7	39.7	72.8	29.4	39.0	59.2	74.8	129.3	132.1
$\mathcal{A}^2$-Transformer (Fei, 2022)	**82.2**	**96.4**	**67.0**	**91.5**	**52.4**	**83.6**	**40.2**	**73.8**	**29.7**	**39.3**	**59.5**	**75.0**	**132.4**	**134.7**
CMAL (Guo et al., 2020)	79.8	94.3	63.8	87.2	48.8	77.2	36.8	66.1	27.9	36.4	57.6	72.0	119.3	121.2
CaptionANT	**80.3**	**94.7**	**64.5**	**88.2**	**49.4**	**78.5**	**37.1**	**67.3**	**28.4**	**37.3**	**58.2**	**73.0**	**120.9**	**124.7**

Table 2: Evaluation Results on the Online MSCOCO Test Server.

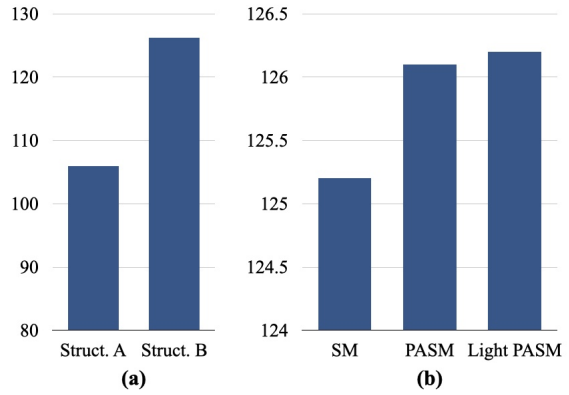

Figure 4: CIDEr Scores of Different Structures.

images for the validation set and the test set, respectively.

4.2. Evaluation Metric

We adopt standard evaluation metrics to compare the performance of different models comprehensively: BLEU (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007), ROUGE-L (Lin, 2004), SPICE (Anderson et al., 2016), CIDEr (Vedantam et al., 2015). Besides, we also show the parameter numbers of different models and the speedup value. The speedup value of CaptionANT is calculated based on the average latency of generating 10,000 sentences.

4.3. Implementation Details

The input size of the mapper and the generator is set to be 384, and the hidden size of the FFN is set to be 1,536, while the input size of the discriminator is 768 and the hidden size of DFFN is 3072. The head numbers are all set to be 8. They are all stacked with 4 blocks. We adopt AdamW as the optimizer of the mapper ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $weight\_decay = 1e{-}5$) and the discriminator ($\beta_1 = 0.5$, $\beta_2 = 0.9$, $weight\_decay = 1e-4$), and Adam ($\beta_1 = 0.5$, $\beta_2 = 0.9$) as the optimizer of the generator. The λ_d is Eq. 6, λ_g and λ_r in Eq. 14 are all set to be 1. The λ_s in Eq 13 is set to be 5.

For the discriminator, we use OpenCLIP ViT-G/14 (Ilharco et al., 2021) as the image encoder. For the generator, we additionally use the features from OpenCLIP ConvNext-XXLarge for the feature ensemble module. All the parameters of the image encoders are fixed during the training process. Knowledge distillation (Kim and Rush, 2016) is adopted as in previous work (Guo et al., 2020; Fei, 2021). The mapper is first trained and its parameters are fixed during the training of the discriminator and generator. Different with the previous work (Guo et al., 2020) which needs a careful adjustment of learning rates, our model can obtain remarkable performance with fixed ones. The learning rates of the mapper, generator and discriminator are set to be 1e-4, 1e-4 and 2e-4, respectively.

Our model is implemented based on Tensorflow[1] and trained on NVIDIA GeForce RTX 3090. Our source code will be released to the public in the near future[2].

	B1	B4	M	R	S	C
CaptionANT	**80.8**	**38.0**	**28.7**	**58.7**	**22.5**	**126.2**
- w/o T.	80.0	37.1	28.3	58.3	22.0	123.5
- w/o F.	79.9	36.4	28.1	57.9	22.0	121.4
- w/o R.	78.5	35.1	27.3	56.8	20.7	116.0
- w/o P. (ANT)	74.9	31.1	25.7	54.3	19.0	102.4

Table 3: Ablation Study of CaptionANT.

4.4. Experimental Result

Overall Performance We compare the performance of CaptionANT with both AR models (Anderson et al., 2018; Cornia et al., 2020; Fei, 2022), SAR models (Fei, 2021; Yan et al., 2021) and NAR models (Gao et al., 2019; Fei, 2020; Guo et al., 2020). Following previous work (Guo et al., 2020), we choose AIC as our AR baseline. AIC is a Transformer based AR model which is first trained with cross entropy and then fine-tuned with SCST (Rennie et al., 2017).

The evaluation results of the "Karpathy" split and the online server can be found in Table 1 and Table 2, respectively. CaptionANT obtains new state-of-the-art performance for fully NAR models. For the "Karpathy" split, it achieves 126.2 for CIDEr, which is 2.2 higher than the existing best fully NAR model, CMAL (Guo et al., 2020). Besides, it is also

[1] https://www.tensorflow.org

[2] https://github.com/compdren/CaptionANT/

	B1	B4	M	R	S	C
Only C_d	80.5	37.5	28.4	58.4	22.1	124.3
Only C_g	79.3	36.8	28.0	57.8	21.5	121.7
Both	**80.8**	**38.0**	**28.7**	**58.7**	**22.5**	**126.2**

Table 4: Effectiveness of the Contrastive Constraints.

the only fully NAR model which can outperform the reported results of PNAIC[3]. It obtains extremely close performance compared with AIC (bw=1), and even outperforms it on some metrics. It is the first time a fully NAR model can achieve such remarkable performance. More importantly, existing SAR and NAR models need more than 50M parameters to obtain close performance as the AR baseline, while CaptionANT obtains the remarkable performance with only 18.2M parameters. It is only 33.1% parameters of the models like AIC and PNAIC, and 36.3% parameters of CMAL. Different with SAR models, which improve model performance by sacrificing speedup, CaptionANT is 26.72× faster than AIC (bw=3). This speedup is much higher than other NAR models. In addition, Guo et al. (2020) also introduce a variant of CMAL which makes use of additional unlabeled data to get further improvement (its CIDEr is 125.5 on the "Karpathy" split). Even comparing with it, our model still obtains better performance without making use of any additional data. More details about this variant can be found in the paper of CMAL (Guo et al., 2020). These experimental results demonstrate that CaptionANT can achieve better performance with much fewer parameters and faster speed.

Performance in Different Strucures We also explore the differences brought by different discriminator structures. We compare the performance between Struct. A: the structure which uses image representations as additional input of the discriminator as in the previous work (Ren and Li, 2023), and Struct. B: the structure adopted in CaptionANT. The results can be found in Figure 4 (a). Compared to Struct. A, Struct. B can effectively make use of unpaired samples to regularize hidden representations. The discriminator thus can better align images and texts, and finally obtains better performance.

In addition, we replace the Light PASM in CaptionANT with Self-modulation (SM) and PASM. Their performance is shown in Figure 4 (b). Both PASM and Light PASM outperform Self-Modulation. It is consistent with the results in the previous work (Ren and Li, 2023). PASM and Light PASM provide diverse input signals which can help models recover different word representations more effectively. The performance between PASM and Light PASM is extremely close, but the parameter number is significantly reduced after adopting Light PASM (the parameter number of the model with PASM is 27.0M while the number of adopting Light PASM is 18.2M). It demonstrates that Light PASM can make the model lighter while maintaining the original performance.

[3]SAR models can further improve performance by sacrificing speedup. More experimental results on these models can be found in their original papers.

AIC: an old truck sitting in a field of flowers
Only C_d: an old truck parked in the grass in a field
Only C_g: an old truck parked in a field with a field
Both: an old truck is parked in a field of flowers
Ground Truth: a rusted out truck parked next to some yellow flowers

AIC: two bikes parked next to a blue bus
Only C_d: two blue bus with a bike on top of it
Only C_g: two blue buses with a bike on the top of it
Both: a bus with two bikes on the front of it
Ground Truth: a bus that has bikes mounted on the front of it

Figure 5: Examples of Generated Captions.

CaptionANT: a woman wearing sunglasses and a white tie
Ground Truth: an image of a very cute girl with face piercings

CaptionANT: a group of zebras are running in a field
Ground Truth: two zebras in a jungle area fighting in a dirt area

CaptionANT: a blue and white airplane is in the grass
Ground Truth: model airplane with an american insignia and stripes on wings

Figure 6: Failure Cases.

Ablation Study Furthermore, we explore the effectiveness of the adopted techniques and show the results in Table 3. The "T.", "F." and "R." indicate the truncation trick, feature ensemble and the reconstruction constraint, respectively. The "P." indicates the projection structure in the discriminator of CaptionANT. After further removing it, the settings will be similar to ANT (Ren and Li, 2023). The performance continuously decreases after removing these techniques, which demonstrates their effectiveness.

The contrastive constraints are adopted in the training objectives of the discriminator and the generator. We also conduct experiments to explore its effectiveness and demonstrate the results in Table 4. Both C_d and C_g contribute to the improve-

ment of model performance, while the contribution from C_d is more important. C_d can help the discriminator obtain more reasonable hidden representations, and identify irrelevant captions more accurately.

Case Study The effectiveness of the contrastive constraints can also be illustrated with the samples in Figure 5. In the first case, the model fails to capture the detail "flower" if one of the constraints is disabled, while the detail is captured accurately when using the two constraints together. In the second case, the models confuse the numbers of "bicycles" and the "bus" if the constraints are lost. With the two constraints, the model describes the numbers correctly.

To perform a complete analysis of CaptionANT, we also show failure cases in Figure 6. In the first case, CaptionANT meets an image in a less common style and uses unrelated words (like sunglasses, and white tie) to describe it. For the second case, although the style is a common one, the content that describes two fighting zebras is not frequent, and CaptionANT fails to describe this image accurately. For the third case, CaptionANT gives a general description, but misunderstands the relatively complicated details (black and white stripes). And it also fails to recognize that this plane is a model airplane. These cases demonstrate that the capacity of CaptionANT in processing less common image styles or content and identifying complicated details requires further enhancement.

5. Conclusion

In the paper, we first analyze the limitations of existing MLE-based NAR models, whose inherent multi-modality problem will be exacerbated in image captioning. Although GANs have potential to tackle this problem, the existing GAN-based NAR model fails to learn complicated relations between images and text, and thus obtains poor performance when transferred to image captioning.

To tackle this problem, we propose CaptionANT. CaptionANT is constructed based on GANs, so it is naturally free from the multi-modality problem. To model the complicated relations between various images and text, we first modify the discriminator structure to enable the use of contrastive learning. The model thus can effectively make use of unpaired samples. Then, we integrate a reconstruction process into the training to better utilize paired samples. By further combining with other effective techniques (like feature ensemble and the truncation trick) and our proposed lightweight structure, CaptionANT achieves new state-of-the-art performance for fully NAR models on the MSCOCO dataset with 36.3% parameters of the existing best fully NAR model and 26.72$\times$ speedup compared with the AR baseline.

6. Ethics Statement

As a generative model, CaptionANT may generate biased or offensive sentences (especially when training data include these kinds of sentences). To avoid this problem, additional filters can be adopted before returning generated results to users.

7. Acknowledgements

This work was supported by the Hong Kong Research Grants Council through the General Research Fund (Project No. 15200023).

8. Bibliographical References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: semantic propositional image caption evaluation. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*, volume 9909 of *Lecture Notes in Computer Science*, pages 382–398. Springer.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6077–6086. Computer Vision Foundation / IEEE Computer Society.

Martín Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR.

Andrew Brock, Jeff Donahue, and Karen Simonyan. 2019. Large scale GAN training for high fidelity natural image synthesis. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Ting Chen, Mario Lucic, Neil Houlsby, and Sylvain Gelly. 2019. On self modulation for generative adversarial networks. In *7th International Conference on Learning Representations, ICLR 2019,*

New Orleans, LA, USA, May 6-9, 2019. OpenReview.net.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10575–10584.

Cyprien de Masson d'Autume, Shakir Mohamed, Mihaela Rosca, and Jack W. Rae. 2019. Training language gans from scratch. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 4302–4313.

Zhengcong Fei. 2020. Iterative back modification for faster image captioning. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 3182–3190. ACM.

Zhengcong Fei. 2021. Partially non-autoregressive image captioning. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 1309–1316. AAAI Press.

Zhengcong Fei. 2022. Attention-aligned transformer for image captioning. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 607–615. AAAI Press.

Junlong Gao, Xi Meng, Shiqi Wang, Xia Li, Shanshe Wang, Siwei Ma, and Wen Gao. 2019. Masked non-autoregressive image captioning. *CoRR*, abs/1906.00717.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-predict: Parallel decoding of conditional masked language models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 6111–6120. Association for Computational Linguistics.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor O. K. Li, and Richard Socher. 2018. Non-autoregressive neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Longteng Guo, Jing Liu, Xinxin Zhu, Xingjian He, Jie Jiang, and Hanqing Lu. 2020. Non-autoregressive image captioning with counterfactuals-critical multi-agent learning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 767–773. ijcai.org.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *CoRR*, abs/1606.08415.

Fei Huang, Tianhua Tao, Hao Zhou, Lei Li, and Minlie Huang. 2022a. On the learning of non-autoregressive transformers. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 9356–9376. PMLR.

Xiao Shi Huang, Felipe Pérez, and Maksims Volkovs. 2022b. Improving non-autoregressive translation models without distillation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. Openclip.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April*

24-26, 2017, Conference Track Proceedings. OpenReview.net.

Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. 2023. Scaling up gans for text-to-image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 10124–10134. IEEE.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3128–3137.

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4401–4410. Computer Vision Foundation / IEEE.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8107–8116. Computer Vision Foundation / IEEE.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1317–1327. The Association for Computational Linguistics.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: an automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation, WMT@ACL 2007, Prague, Czech Republic, June 23, 2007*, pages 228–231. Association for Computational Linguistics.

Kwonjoon Lee, Huiwen Chang, Lu Jiang, Han Zhang, Zhuowen Tu, and Ce Liu. 2022. Vitgan: Training gans with vision transformers. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip H. S. Torr. 2019. Controllable text-to-image generation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 2063–2073.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Chun-Hsing Lin, Siang-Ruei Wu, Hung-yi Lee, and Yun-Nung Chen. 2020. Taylorgan: Neighbor-augmented policy update towards sample-efficient natural language generation. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Jianjie Luo, Yehao Li, Yingwei Pan, Ting Yao, Jianlin Feng, Hongyang Chao, and Tao Mei. 2023. Semantic-conditional diffusion networks for image captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 23359–23368. IEEE.

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 16784–16804. PMLR.

Weili Nie, Nina Narodytska, and Ankit Patel. 2019. Relgan: Relational generative adversarial networks for text generation. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Henning Petzka, Asja Fischer, and Denis Lukovnikov. 2018. On the regularization of

wasserstein gans. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Lihua Qian, Hao Zhou, Yu Bao, Mingxuan Wang, Lin Qiu, Weinan Zhang, Yong Yu, and Lei Li. 2021. Glancing transformer for non-autoregressive neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1993–2003. Association for Computational Linguistics.

Scott E. Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1060–1069. JMLR.org.

Da Ren and Qing Li. 2022. Initialgan: A language gan with completely random initialization. *CoRR*, abs/2208.02531.

Da Ren and Qing Li. 2023. An adversarial non-autoregressive model for text generation with incomplete information. *CoRR*, abs/2305.03977.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: Fast, robust and controllable text to speech. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3165–3174.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1179–1195. IEEE Computer Society.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. 2022. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. 2023. Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 30105–30118. PMLR.

Axel Sauer, Katja Schwarz, and Andreas Geiger. 2022. Stylegan-xl: Scaling stylegan to large diverse datasets. In *SIGGRAPH '22: Special Interest Group on Computer Graphics and Interactive Techniques Conference, Vancouver, BC, Canada, August 7 - 11, 2022*, pages 49:1–49:10. ACM.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575. IEEE Computer Society.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, 8:229–256.

Yisheng Xiao, Lijun Wu, Junliang Guo, Juntao Li, Min Zhang, Tao Qin, and Tie-Yan Liu. 2023. A survey on non-autoregressive generation for neural machine translation and beyond. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(10):11407–11427.

Xu Yan, Zhengcong Fei, Zekang Li, Shuhui Wang, Qingming Huang, and Qi Tian. 2021. Semi-autoregressive image captioning. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 2708–2716. ACM.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin

Li, Han Zhang, Jason Baldridge, and Yonghui Wu. 2022. Scaling autoregressive models for content-rich text-to-image generation. *Trans. Mach. Learn. Res.*

Han Zhang, Tao Xu, and Hongsheng Li. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5908–5916. IEEE Computer Society.

Han Zhang, Weichong Yin, Yewei Fang, Lanxin Li, Boqiang Duan, Zhihua Wu, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-vilg: Unified generative pre-training for bidirectional vision-language generation. *CoRR*, abs/2112.15283.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2242–2251. IEEE Computer Society.