THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Releasing the Capacity of GANs in Non-Autoregressive Image Captioning

Da Ren, Qing Li

Department of Computing, The Hong Kong Polytechnic University

{csdren, csqli}@comp.polyu.edu.hk

LREC-COLING 2024

Outline

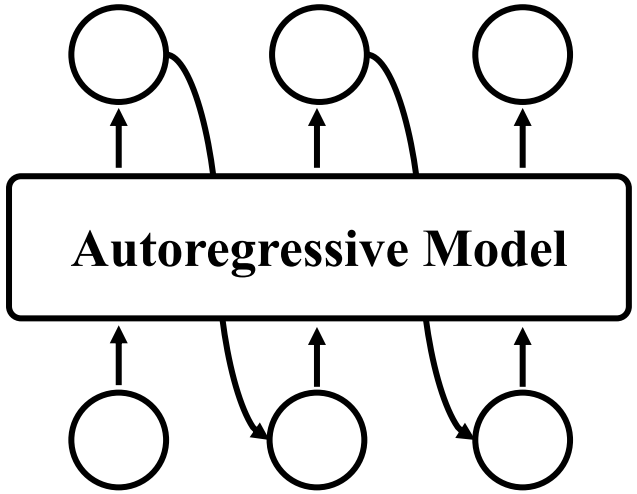
THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

- Introduction
- Model
- Experiment
- Conclusion

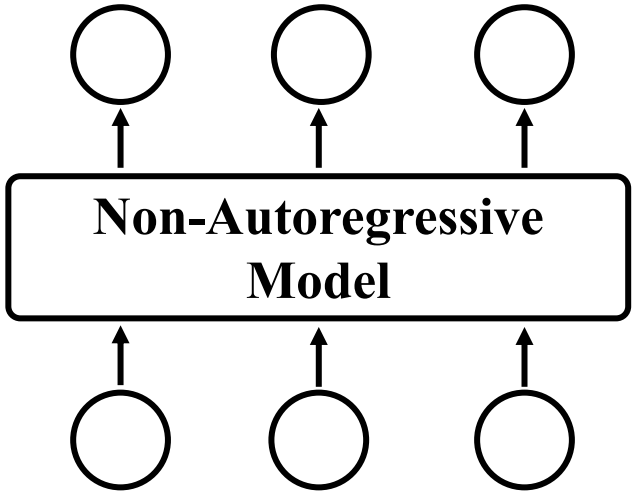
Introduction

Introduction

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學



- Generate tokens **one-by-one**
- **High latency**



- Generate tokens **in parallel**
- **Low latency**

Introduction

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Existing Non-autoregressive (NAR) Models in image captioning:

- Based on Maximum Likelihood Estimation (MLE)

Learn the **marginal distributions**, but **lose word dependencies** → **Multi-modality problem**: Mixing words in different candidates.

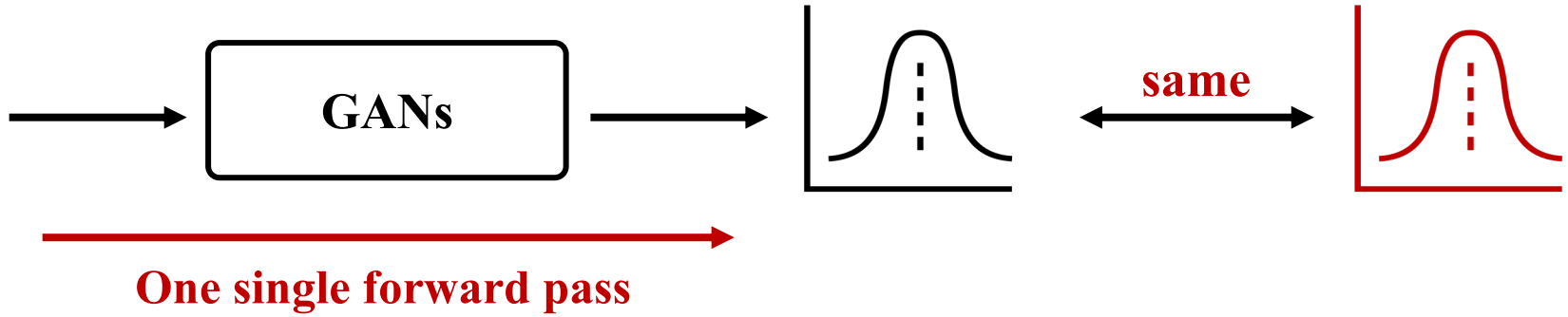
Difficulties in the **alignment** between images and text → **Greater errors** in the learned marginal distributions.

Exacerbating

Introduction

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Generative Adversarial Networks (GANs):



Exactly meets the needs of NAR models

Introduction

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Generative Adversarial Networks (GANs):

- **Incapacity** in building complicated mapping relations between **images** and **text**.

↓ Modify the discriminator structure

↓ Better make use of **unpaired samples**

↓ Integrate the reconstruction process

↓ Better utilize **paired samples**

Introduction

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Contributions

- 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.

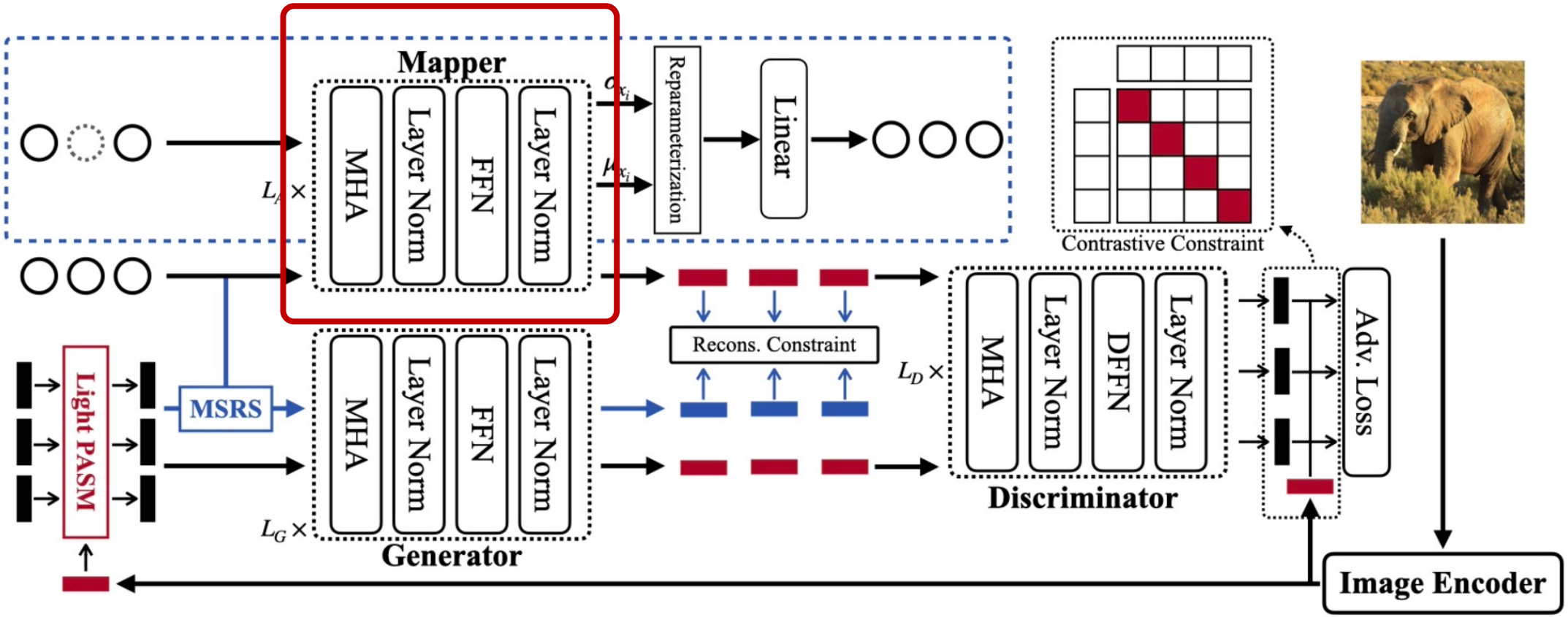
- We further propose to incorporate a **reconstruction process** into the training stage of language GANs. It can further improve model performance by better utilizing **paired samples**. For the ambiguous reconstruction targets led by the one-to-many mapping relations, we propose to integrate part of 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**.

Model

Model

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學



Model

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Mapper

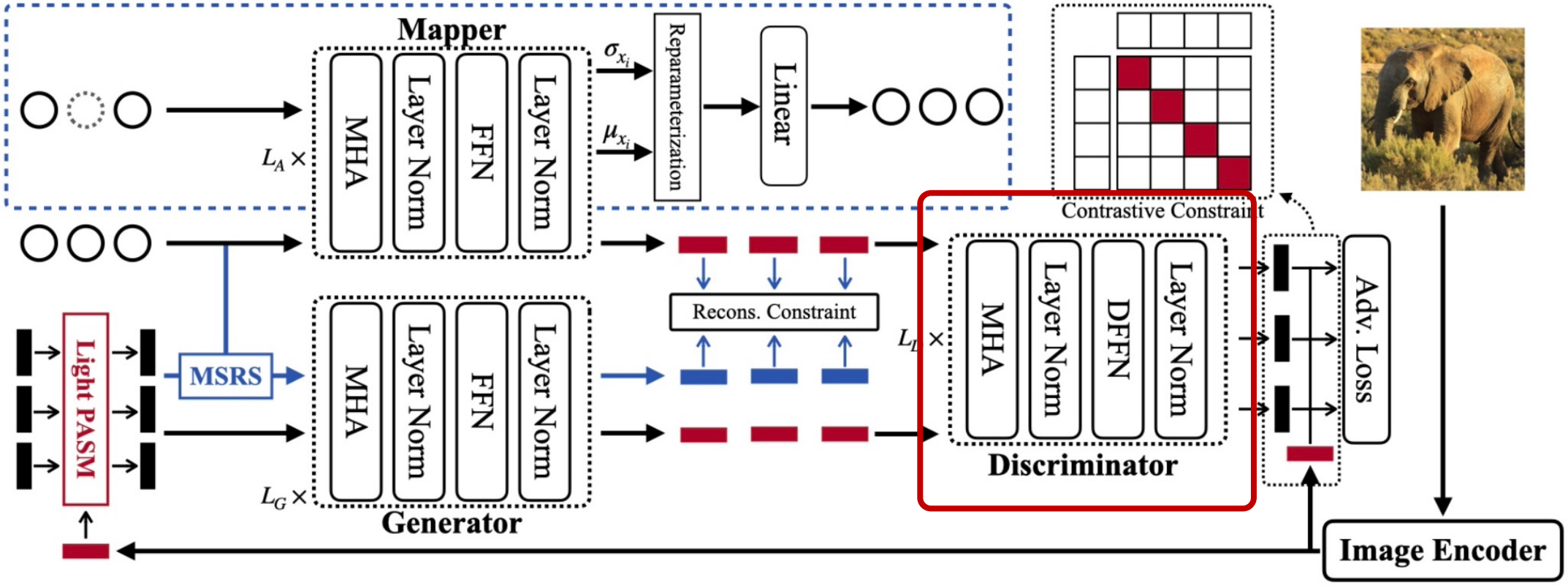
- The mapper needs to map words into representations.
- A certain number of words in a sentence are randomly masked or replaced, and the mapper is trained to reconstruct the original input.
- After obtaining the mean and standard deviation for each word, the mapper first adopts reparameterization trick to obtain hidden representations:

$$\mathbf{z}'_i = \mu_{x_i} + \sigma_{x_i} \cdot \mathcal{N}(0, 1)$$

- Use the following training objective:

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

Model


Mapper
MHA
Layer Norm
FFN
Layer Norm
$L_A \times$
σ_{x_i}
μ_{x_i}
Reparameterization
Linear
Contrastive Constraint
Recons. Constraint
Light PASM
MSRS
MHA
Layer Norm
FFN
Layer Norm
$L_G \times$
Generator
$L_D \times$
MHA
Layer Norm
DFFN
Layer Norm
Discriminator
Adv. Loss
Image Encoder

Model

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Discriminator

- One key challenge in building the discriminator is how to incorporate conditions into the model.
- Previous work feeds the condition representation as **input**. It can only consider **one pair of mismatched samples at a time**.
- Map input text to the **same space** as the condition representation, so it can measure the correlation between the two by calculating dot product.

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

Model

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Discriminator

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

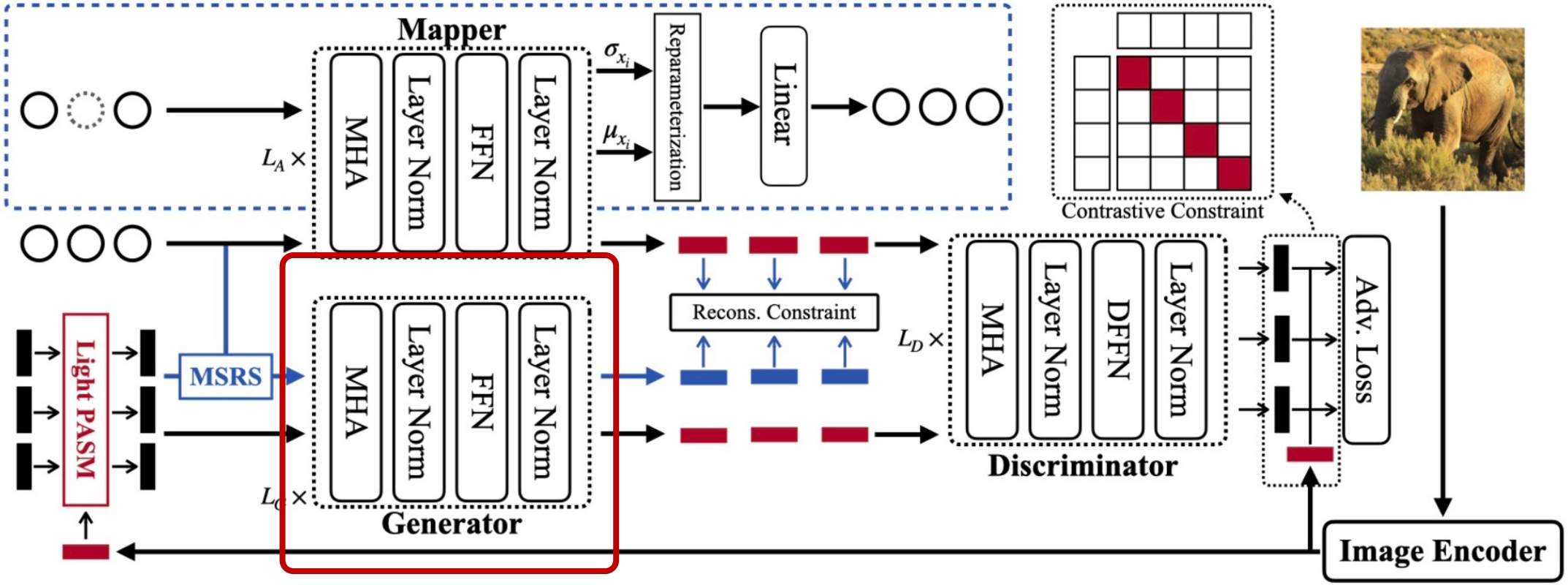
Adversarial loss

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

Contrastive constraint

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

Model


Mapper
MHA
Layer Norm
FFN
Layer Norm
$L_A \times$
σ_{x_i}
μ_{x_i}
Reparameterization
Linear
Contrastive Constraint
Light PASM
MSRS
MHA
Layer Norm
FFN
Layer Norm
$L_G \times$
Generator
Recons. Constraint
$L_D \times$
MHA
Layer Norm
DFFN
Layer Norm
Discriminator
Adv. Loss
Image Encoder

Model

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Generator

- Feature ensemble

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

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

- Light Position-Aware Self-Modulation

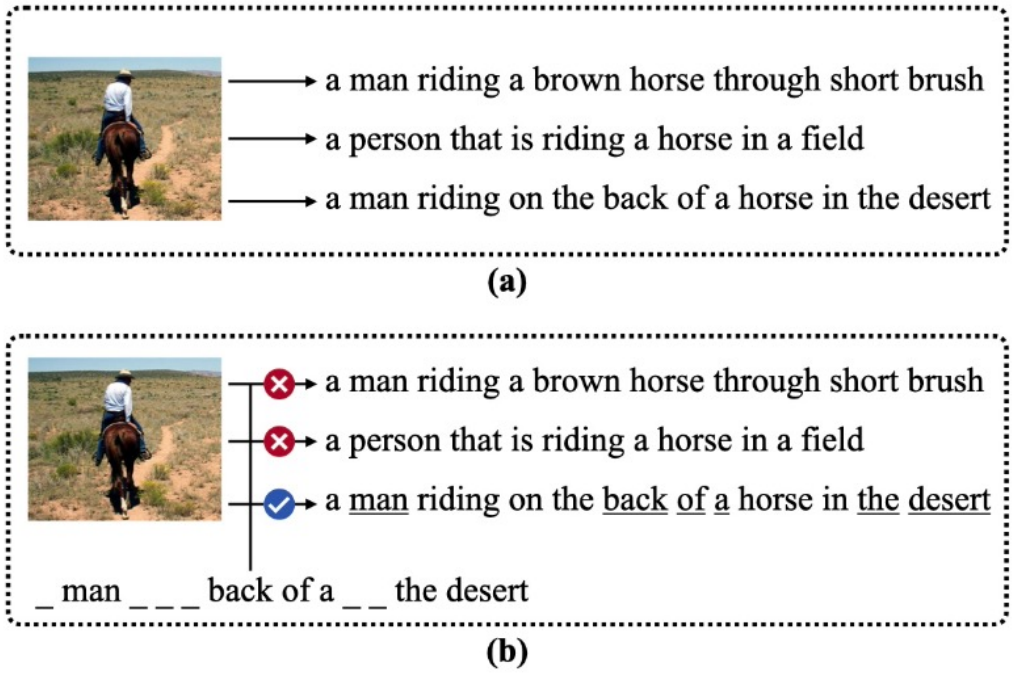
$$\begin{pmatrix} \hat{\mathbf{s}}_1 \\ \hat{\mathbf{s}}_2 \\ \vdots \\ \hat{\mathbf{s}}_{\frac{N}{2}} \end{pmatrix} = \begin{pmatrix} W_1 \\ W_2 \\ \vdots \\ W_{\frac{N}{2}} \end{pmatrix} \cdot \hat{\mathbf{z}} + \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_{\frac{N}{2}} \end{pmatrix} \qquad \begin{pmatrix} \hat{\mathbf{s}}_{\frac{N}{2}+1} \\ \hat{\mathbf{s}}_{\frac{N}{2}+2} \\ \vdots \\ \hat{\mathbf{s}}_N \end{pmatrix} = W' \cdot \begin{pmatrix} \hat{\mathbf{s}}_1 \\ \hat{\mathbf{s}}_2 \\ \vdots \\ \hat{\mathbf{s}}_{\frac{N}{2}} \end{pmatrix} + b'$$

Model

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Generator

- Reconstruction Constraint



$$\mathbf{e}_i = Emb(x_i) + pos_i$$
$$\hat{\mathbf{e}}_i = Mask(\mathbf{e}_i, \rho)$$
$$\dot{\mathbf{s}}_i = \omega \circ MHA(\mathbf{s}_i, \hat{\mathbf{e}}_i, \hat{\mathbf{e}}_i)$$
$$\hat{\mathbf{s}}_i = \mathbf{s}_i + \dot{\mathbf{s}}_i$$

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

Model

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Generator

- Training Objective

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

Adversarial loss ↑

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

Contrastive constraint ↓

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

Reconstruction constraint ↓

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

Experiment

Experiment

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Experimental results

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.

Experiment

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Experimental results

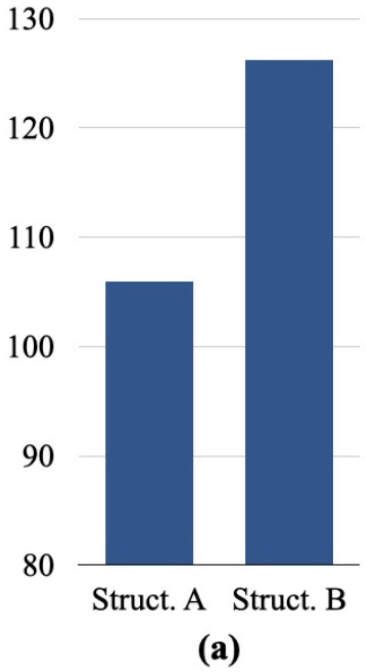
Model	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		ROUGE-L		CIDEr	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
Up-Down (Anderson et al., 2018)	80.2	95.2	64.1	88.8	49.1	79.4	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
M2-T (Cornia et al., 2020)	81.6	96.0	66.4	90.8	51.8	82.7	39.7	72.8	29.4	39.0	59.2	74.8	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.

Experiment

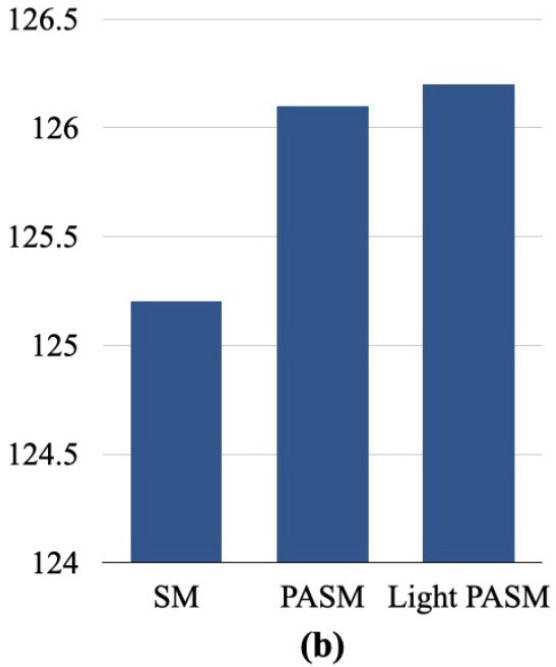
THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Experimental results



(a)

- Struct. A: uses image representations as additional input of the discriminator
- Struct. B: the structure adopted in CaptionANT



(b)

- SM: Self-Modulation
- PASM: Position-Aware Self-Modulation (#Param.: 27M)
- Light PASM: Light Position-Aware Self-Modulation (#Param.: 18.2M)

Experiment

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Experimental results

	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

	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**

Experiment

Case Study

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

Conclusion

Conclusion

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

- To tackle the inherent **multi-modality problem** in MLE-based NAR models, we propose a GAN-based NAR model: **CaptionANT**.
- We first modify the discriminator structure to enable the use of **contrastive learning**, so the model 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 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× speedup compared with the AR baseline**.

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Thanks for your time

Da Ren

csdren@comp.polyu.edu.hk

The Hong Kong Polytechnic University