

# Paper presentation

---

Towards diverse and natural image descriptions via a conditional GAN, B. Dai, et. al.

# Paper presentation

---


Towards diverse and natural image descriptions via a conditional GAN, B. Dai, et. al.

Has the problem of  
generating image descriptions  
been solved?

—

# Problem Statement

- Image captioning with a focus on fidelity, naturalness, and diversity

			BLEU	E-GAN
	G-MLE	A cow standing in a field next to houses	<div><div></div></div>	<div><div></div></div>
		A cow standing in a field with houses	<div><div></div></div>	<div><div></div></div>
		A cow standing in a field of grass	<div><div></div></div>	<div><div></div></div>
	G-GAN	Many cows grazing in the grass field in front of houses	<div><div></div></div>	<div><div></div></div>
		Several cows grazing on grassy area in a pasture	<div><div></div></div>	<div><div></div></div>
		A heard of cattle grazing on a lush green field	<div><div></div></div>	<div><div></div></div>
	human	Grey cow walking in a large green field in front of house	<div><div></div></div>	<div><div></div></div>
		A cow in a large open field with a house in the background	<div><div></div></div>	<div><div></div></div>
		A cow standing in a large open grass field	<div><div></div></div>	<div><div></div></div>

Captions generated by 3 different methods for the given image on the left side. [1]

# Motivation

- Less variable and rigid captions produced by the state of the art methods
  - High resemblance captions to the ground truth
- Unavailable metrics to capture variability and naturalness
  - Some of the current state of the art metrics:
    - BLEU
    - METEOR

# Related Work

- Generation
  - Detection-based approaches
    - CRF, SVM, CNN
    - Retrieving sentences from existing data or using sentence templates
  - Encoder-decoder paradigm
    - Maximum likelihood

$$\sum_{(I_i, S_i) \sim D} \sum_{t=0}^{T_i} \log p(w_i^{(t)} | f(I), w_i^{(t-1)}, \dots, w_i^{(t-n)})$$

# Related Work

- Evaluation
  - Classical metrics
    - BLEU: Precision
    - ROUGE: Recall of n-grams
  - METEOR
    - Combination of precision and recall
  - CIDEr
    - Weighted statistics
  - SPICE
    - Focus on linguistic entities reflecting visual concepts

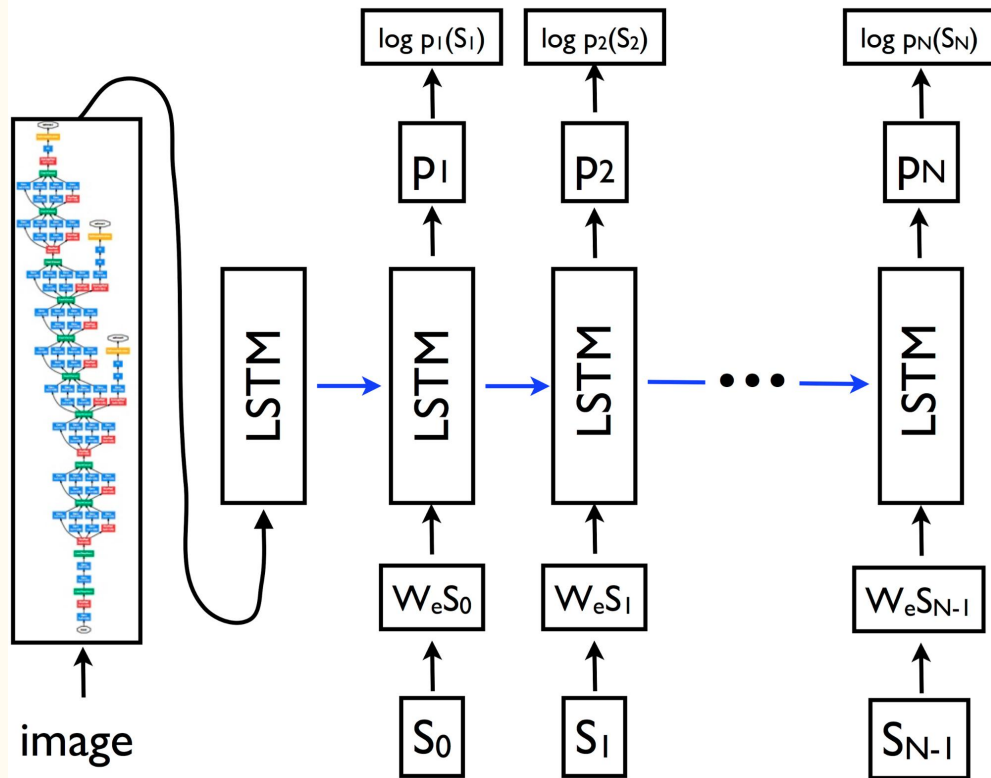
What is the state of the art  
model for image captioning?

—



# State of the art

- Combination of LSTM and CNN
- Metrics
  - CIDEr
  - METEOR
  - BLEU
  - ROUGE



LSTM based model for image captioning. [2]

# State of the art

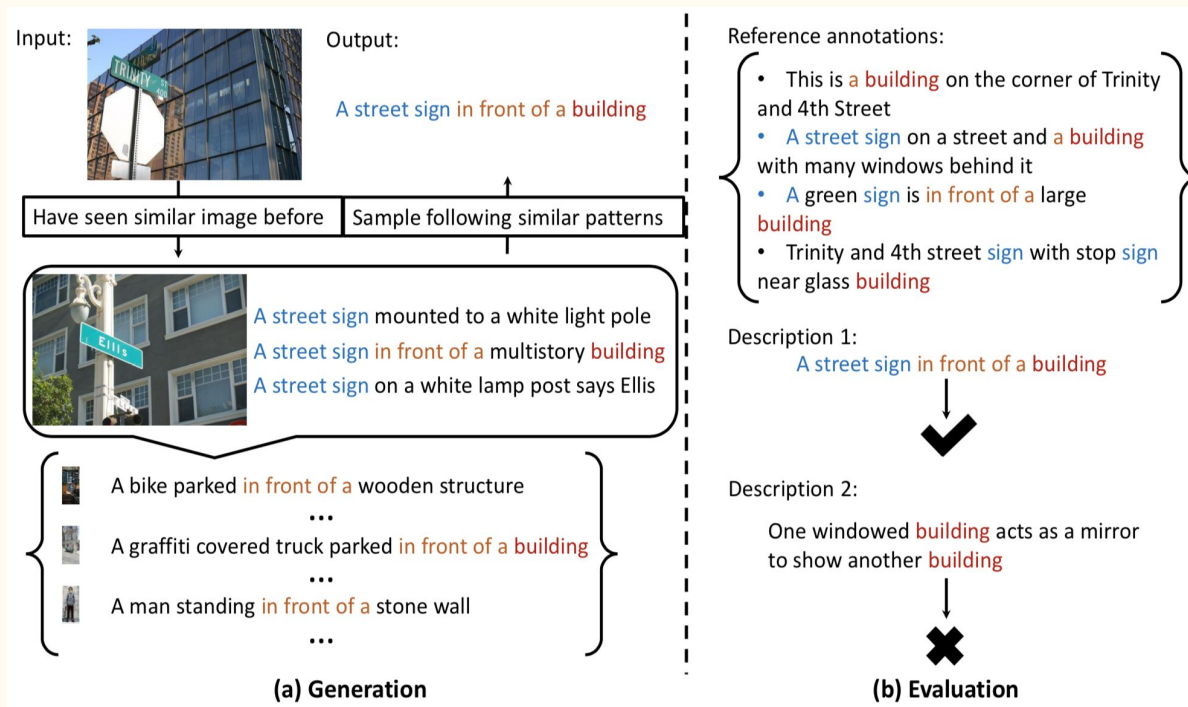



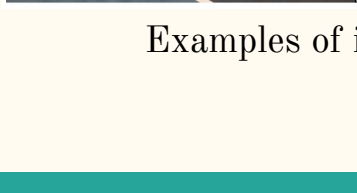


Image description generation and evaluation for the state of the art approaches. [1]

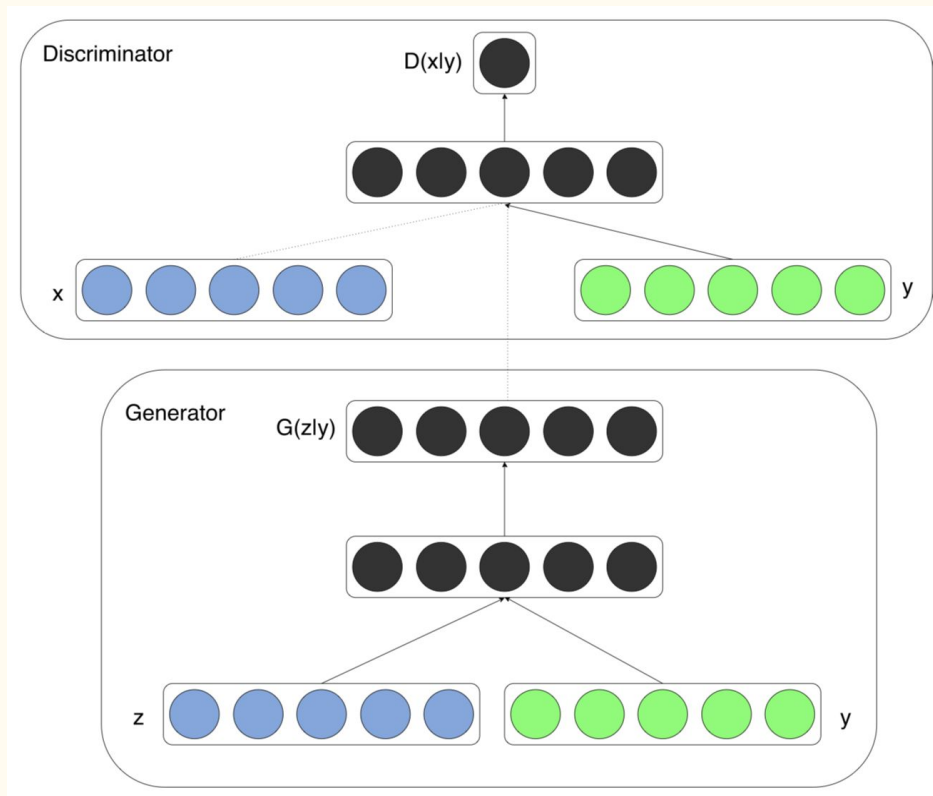
# State of the art

	a woman holding a skateboard on a street						
	0.71	0.61	0.75	0.36	1.49	0.28	0.05
	B3	B4	ROUGE	METEOR	CIDEr	SPICE	E-GAN
	0.25	0.01	0.48	0.19	0.36	0.14	0.37
	three women one with a skateboard outside a store						
	a baseball player swinging a bat at a ball						
	0.71	0.65	0.78	0.39	2.21	0.28	0.48
	B3	B4	ROUGE	METEOR	CIDEr	SPICE	E-GAN
	0.01	0.01	0.31	0.23	0.82	0.25	0.82
	the umpire stands over a catcher as the batter swings						
	a man holding a tennis racquet on a tennis court						
	0.99	0.99	1.0	1.0	3.53	0.58	0.69
	B3	B4	ROUGE	METEOR	CIDEr	SPICE	E-GAN
	0.01	0.01	0.48	0.28	1.03	0.2	0.67
	a man getting ready to serve a tennis ball						

Examples of images with two semantically similar descriptions. [1]

# Background

- Conditional GAN
  - Conditioned on some extra information  $y$ .

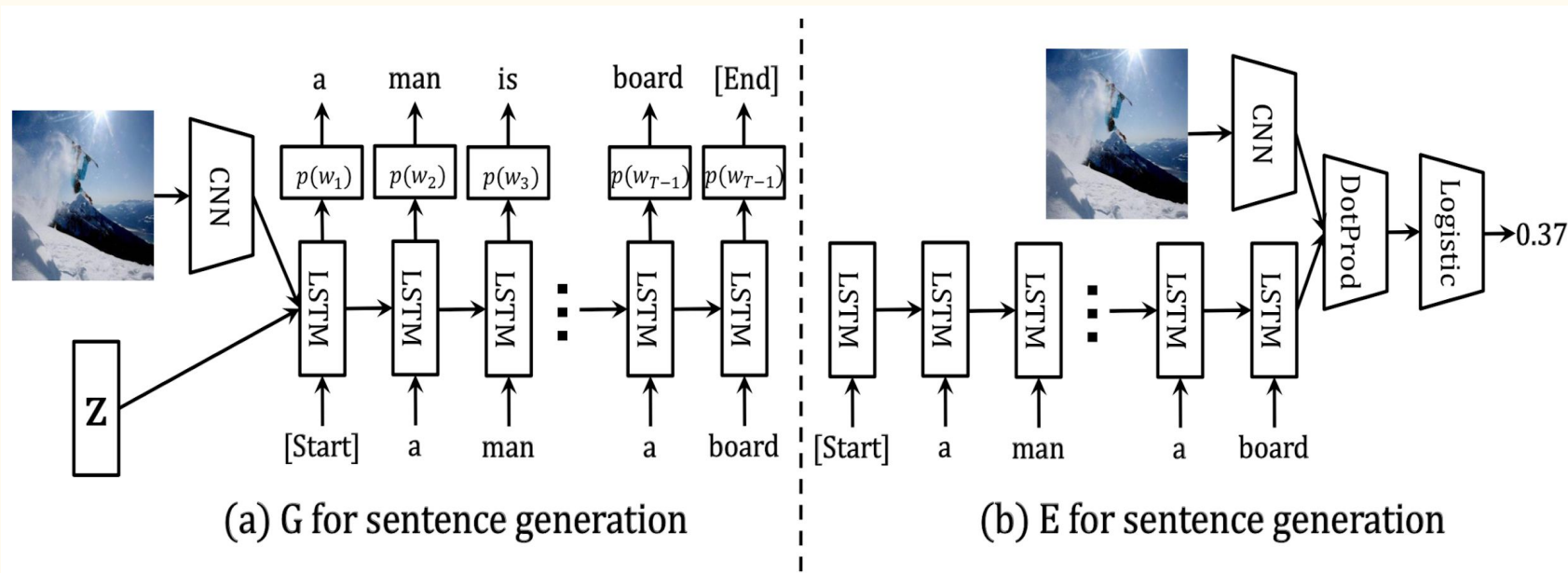


Conditional adversarial net. [3]

# Background

- Sequential Sampling
  - Non-probability sampling technique
    - Not equal chances of being selected for all members of the population
  - Picking a single or a group of objects in every time interval
  - Analyze the result, pick another sample
- Monte Carlo tree search
  - A heuristic search algorithm for some kinds of decision processes esp. Game plays
  - Steps
    - Selection
    - Expansion
    - Simulation
    - Backpropagation

# Proposed Approach



Overall structure of both generator and evaluator of a CGAN. [1]

# Overall formulation

- Evaluator
  - Quality of descriptions

$$r_{\boldsymbol{\eta}}(I, S) = \sigma(\langle \mathbf{f}(I, \boldsymbol{\eta}_I), \mathbf{h}(S, \boldsymbol{\eta}_S) \rangle)$$

- Learning objective

$$\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\eta}} \mathcal{L}(G_{\boldsymbol{\theta}}, E_{\boldsymbol{\eta}})$$

$$\mathbb{E}_{S \sim \mathcal{P}_I} [\log r_{\boldsymbol{\eta}}(I, S)] + \mathbb{E}_{\mathbf{z} \sim \mathcal{N}_0} [\log(1 - r_{\boldsymbol{\eta}}(I, G_{\boldsymbol{\theta}}(I, \mathbf{z})))]$$

The production of sentences  
is non-differentiable. How can  
we backpropagate?

—



# Challenges

- Using policy gradient for generating linguistic description
  - Sequential sampling procedure
    - Sampling a discrete token at each time step
  - Non-differentiable
- Expected future reward for early feedback using Monte Carlo rollouts
  - Vanishing gradients
  - Error propagation

# Training G

- Policy gradient
  - Action space: words
  - Conditional policy

$$\pi_{\theta}(w_t | \mathbf{f}(I), \mathbf{z}, S_{1:t-1})$$

- Reward given by the evaluator for a sequence of actions S

$$r_{\eta}(I, S)$$

# Training G

- Early feedback
  - Expected future reward

$$V_{\boldsymbol{\theta}, \boldsymbol{\eta}}(I, \mathbf{z}, S_{1:t}) = \mathbb{E}_{S_{t+1:T} \sim G_{\boldsymbol{\theta}}(I, \mathbf{z})} [r_{\boldsymbol{\eta}}(I, S_{1:t} \oplus S_{t+1:T})]$$

- Gradient of the objective w.r.t.  $\theta$

$$\tilde{\mathbb{E}} \left[ \sum_{t=1}^{T_{\max}} \sum_{w_t \in \mathcal{V}} \nabla_{\boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(w_t | I, \mathbf{z}, S_{1:t-1}) \cdot V_{\boldsymbol{\theta}', \boldsymbol{\psi}}(I, \mathbf{z}, S_{1:t} \oplus w_t) \right]$$

# Training E

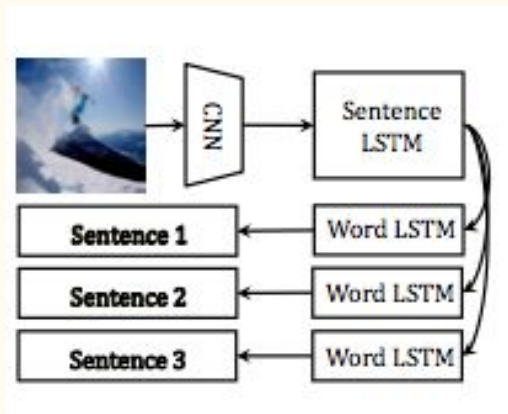
- Enforcing naturalness and semantic relevance
  - Set of descriptions provided by human
  - Descriptions from the generator
  - Human descriptions uniformly sampled from other descriptions not associated with the given image

$$\max_{\boldsymbol{\eta}} \mathcal{L}_E(\boldsymbol{\eta}) = \frac{1}{N} \sum_{i=1}^N \mathcal{L}_E(I_i; \boldsymbol{\eta})$$

$$\begin{aligned} \mathcal{L}_E(I; \boldsymbol{\eta}) = & \mathbb{E}_{S \in \mathcal{S}_I} \log r_{\boldsymbol{\eta}}(I, S) \\ & + \alpha \cdot \mathbb{E}_{S \in \mathcal{S}_G} \log(1 - r_{\boldsymbol{\eta}}(I, S)) \\ & + \beta \cdot \mathbb{E}_{S \in \mathcal{S}_{\setminus I}} \log(1 - r_{\boldsymbol{\eta}}(I, S)) \end{aligned}$$

# Generating paragraphs

- Hierarchical LSTM
  - Sentence level
  - Word level



Adding extension for paragraph generation. [1]

# Experiment

- Datasets
  - MSCOCO
  - Flickr30k
- Settings
  - Removing non-alphabet characters
  - Converting to lowercase
  - Replacing less frequent words, less than 5 times, with UNK
  - Max length of 16
  - Pretrain G for 20 epochs based on MLE
  - Pretrain E with supervised training for 5 epochs
  - Batch = 64, lr = 0.0001, n = 16 (Monte Carlo)

# Experiment

- Performance

		BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE	E-NGAN	E-GAN
COCO	human	0.290	0.192	0.240	0.465	0.849	<b>0.211</b>	0.527	<b>0.626</b>
	G-MLE	<b>0.393</b>	<b>0.299</b>	<b>0.248</b>	<b>0.527</b>	<b>1.020</b>	0.199	0.464	0.427
	G-GAN	0.305	0.207	0.224	0.475	0.795	0.182	<b>0.528</b>	0.602
Flickr	human	0.269	0.185	0.194	0.423	0.627	0.159	0.482	<b>0.464</b>
	G-MLE	<b>0.372</b>	<b>0.305</b>	<b>0.215</b>	<b>0.479</b>	<b>0.767</b>	<b>0.168</b>	0.465	0.439
	G-GAN	0.153	0.088	0.132	0.330	0.202	0.087	<b>0.582</b>	0.456

Performances of different generators on MSCOCO and Flickr30k. [1]

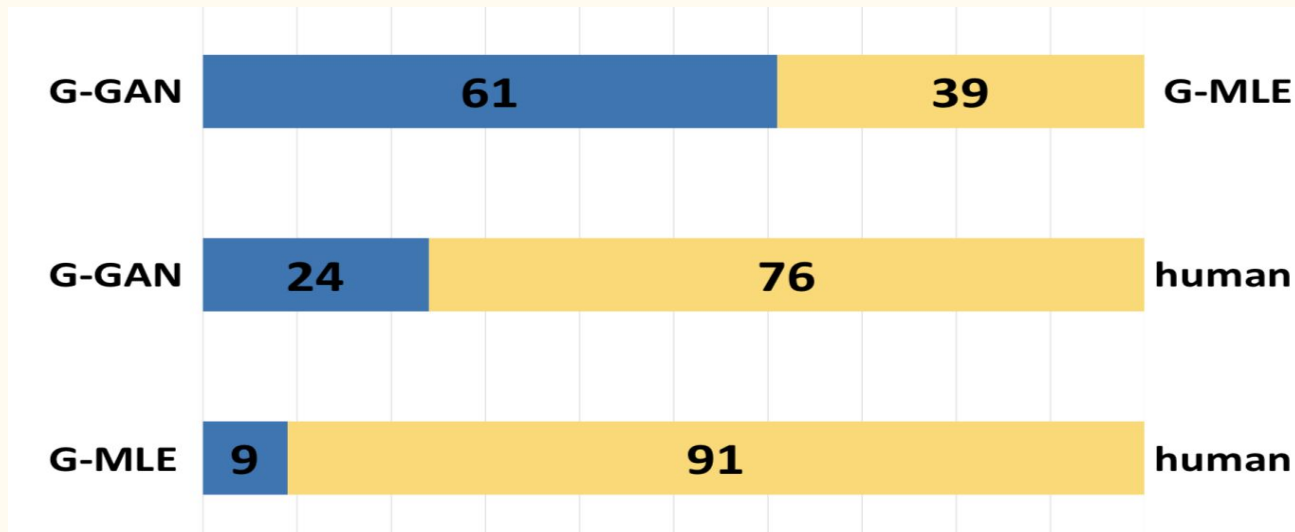
Has the proposed approach  
been able to achieve more  
natural descriptions?

—



# Experiment









- User Study



Human comparison results between each pair of generators. [1]

# Experiment

- User Study

				
G-MLE	a man flying through the air while riding a snowboard	a man flying through the air while riding a snowboard	a man flying through the air while riding a snowboard	a man flying through the air while riding a skateboard
G-GAN	a man on a skateboard in a snowy park	a man skiing down the slope near a mountain	a man performing a grind trick on a skateboard ramp	a man with stunts on his skis in the snow
				
G-MLE	a group of people standing around a boat	a group of people sitting around a table	a group of people sitting at a table	a group of people sitting around a living room
G-GAN	the bench is sitting on the ground by the water	a group of people watching each other	a table with a lot of stuff on it	furnished living room with furniture and built area

Corresponding G-GAN captions for images with similar descriptions in G-MLE. [1]

# Experiment

- Diverse descriptions

				
$\mathbf{z}_1$	a baseball player holds a bat up to hit the ball	a man riding a snowboard down a slope	a group of people sitting around a table having a meal in a restaurant	a group of men dressed in suits posing for a photo
$\mathbf{z}_2$	a baseball player holding white bat and wear blue baseball uniform	a person standing on a snowboard sliding down a hill	a young man sitting at a table with coffee and a lot of food	a couple of men standing next to each other wearing glasses
$\mathbf{z}_3$	a professional baseball player holds up his bat as he watches	a man is jumping over a snow covered hill	a pretty young man sitting next to two men in lots of people	some people dressed in costume and cups

Generated descriptions with different  $\mathbf{z}$ . [1]

# Experiment

- Evaluating semantic relevance by retrieval

		R@1	R@3	R@5	R@10
S	G-MLE	5.06	12.28	18.24	29.30
	G-GAN	<b>14.30</b>	<b>30.88</b>	<b>40.06</b>	<b>55.82</b>
P	G-MLE	9.88	20.12	27.30	39.94
	G-GAN	<b>12.04</b>	<b>23.88</b>	<b>30.70</b>	<b>41.78</b>



Image rankings for different generators. [1]

S  $\rightarrow$  E-GAN

P  $\rightarrow$  Log-likelihood

# Experiment

- Paragraph generation with different  $z$  values

	human	G-GAN, $z_1$	G-GAN, $z_2$	G-MLE
	people are on motorcycles. there are green cars behind them. the signs are all brown with chinese written on it.	men are riding on a motorcycle. the man is wearing tan boots, and a white and blue jacket with beige stripes on. the street is made of cobblestone. there are tall bright green trees on the sidewalk.	two people are riding motorcycles. there are many trees on the sidewalk. there is a red and white painted letter on the side of the ledge. tall buildings are on the background.	a man is riding a bike. there are trees on the sidewalk. there are people walking on the sidewalk. there is a tall building in the background.
	A baseball player is swinging a bat. He is wearing a black helmet and a black and white uniform. A catcher is behind him wearing a gray uniform. The catcher has a brown glove on his hand. Two men can be seen standing behind a green fence.	a baseball player in a white and blue uniform is holding a white bat. there is a umpire behind the batter in the blue and white uniform. he is getting ready to catch the ball. there is a crowd of people behind him watching him.	men are on a baseball field on a sunny day. the player is wearing a black and white uniform. there is a catcher behind him. the field is green with brown dirt and white shiny lines.	a baseball player is standing on a baseball field. he is wearing a blue helmet on his head. the catcher is wearing a black and gray uniform. the court is green with white lines.

Different paragraph descriptions generated by human, G-GAN, and G-MLE with different  $z$  values. [1]

# Failure Analysis

- Incorrect details
  - Colors
  - Counts
    - Few samples for each special detail
    - Increased risk of putting more incorrect details due to focus on diversity

Potential extensions?

—

# Future works

- Use VAE instead of GAN
- Use other similarity metrics instead of dot product in evaluator



# Presenters

- Kevin Dsouza
- Ainaz Hajimoradlou

# References

1. B. Dai, S. Fidler, R. Urtasun, and D. Lin, Towards Diverse and Natural Image Descriptions via a Conditional GAN, 2017, *arXiv preprint arXiv: 1703.06029*
2. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge, *arXiv preprint arXiv: 1609.06647*
3. M. Mirza, S. Osindero, Conditional Generative Adversarial Nets, *arXiv preprint arXiv: 1411.1784*

Thank you.




Has the problem of  
generating image descriptions  
been solved?

—

# Problem Statement

- Image captioning with a focus on fidelity, naturalness, and diversity

			BLEU	E-GAN
	G-MLE	A cow standing in a field next to houses	<div><div></div></div>	<div><div></div></div>
		A cow standing in a field with houses	<div><div></div></div>	<div><div></div></div>
		A cow standing in a field of grass	<div><div></div></div>	<div><div></div></div>
	G-GAN	Many cows grazing in the grass field in front of houses	<div><div></div></div>	<div><div></div></div>
		Several cows grazing on grassy area in a pasture	<div><div></div></div>	<div><div></div></div>
		A heard of cattle grazing on a lush green field	<div><div></div></div>	<div><div></div></div>
	human	Grey cow walking in a large green field in front of house	<div><div></div></div>	<div><div></div></div>
		A cow in a large open field with a house in the background	<div><div></div></div>	<div><div></div></div>
		A cow standing in a large open grass field	<div><div></div></div>	<div><div></div></div>

Captions generated by 3 different methods for the given image on the left side. [1]

# Motivation

- Less variable and rigid captions produced by the state of the art methods
  - High resemblance captions to the ground truth
- Unavailable metrics to capture variability and naturalness
  - Some of the current state of the art metrics:
    - BLEU
    - METEOR

# Related Work

- Generation
  - Detection-based approaches
    - CRF, SVM, CNN
    - Retrieving sentences from existing data or using sentence templates
  - Encoder-decoder paradigm
    - Maximum likelihood

$$\sum_{(I_i, S_i) \sim D} \sum_{t=0}^{T_i} \log p(w_i^{(t)} | f(I), w_i^{(t-1)}, \dots, w_i^{(t-n)})$$

# Related Work

- Evaluation
  - Classical metrics
    - BLEU: Precision
    - ROUGE: Recall of n-grams
  - METEOR
    - Combination of precision and recall
  - CIDEr
    - Weighted statistics
  - SPICE
    - Focus on linguistic entities reflecting visual concepts

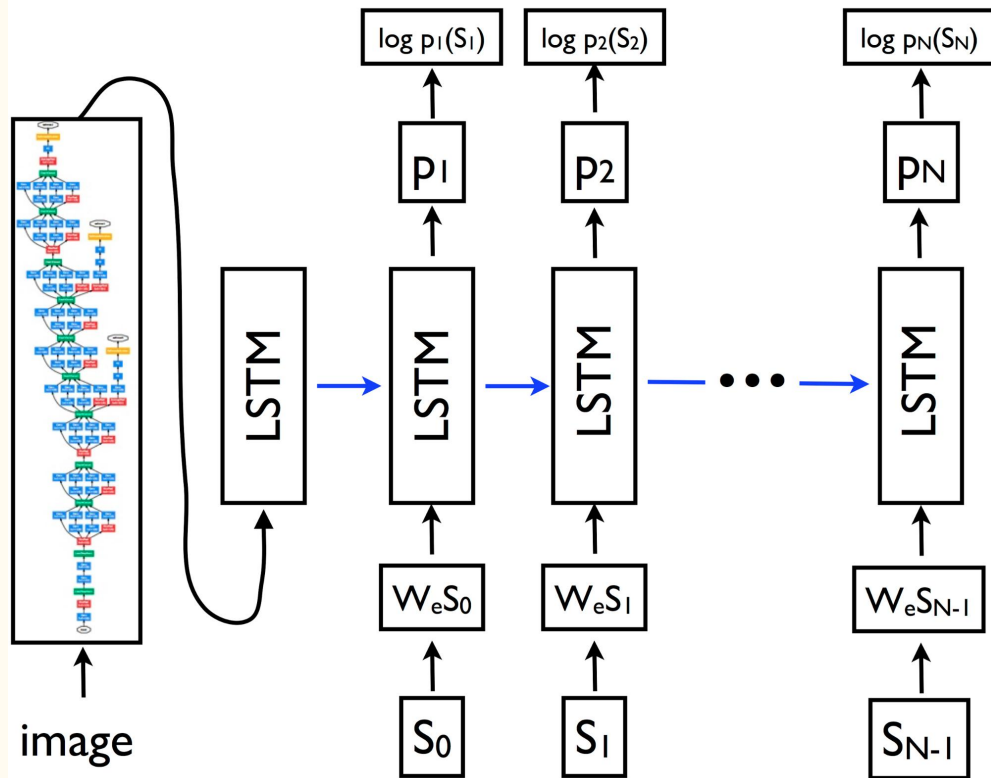


What is the state of the art  
model for image captioning?

—

# State of the art

- Combination of LSTM and CNN
- Metrics
  - CIDEr
  - METEOR
  - BLEU
  - ROUGE



LSTM based model for image captioning. [2]

# State of the art

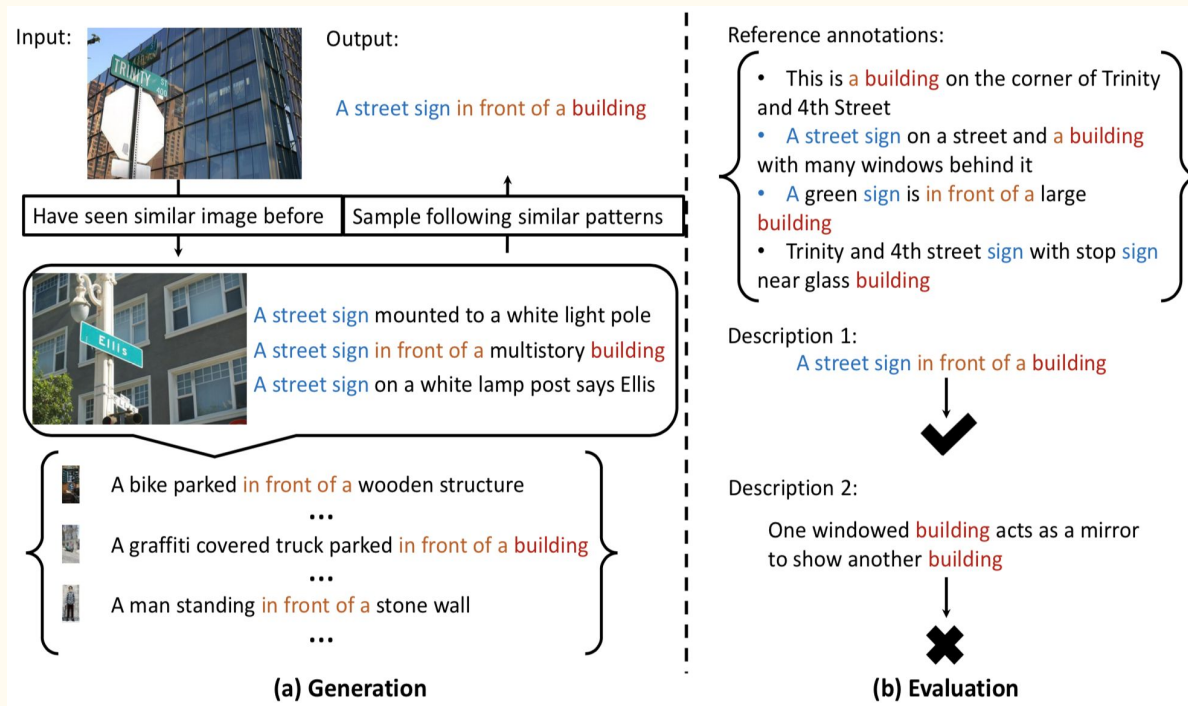





Image description generation and evaluation for the state of the art approaches. [1]

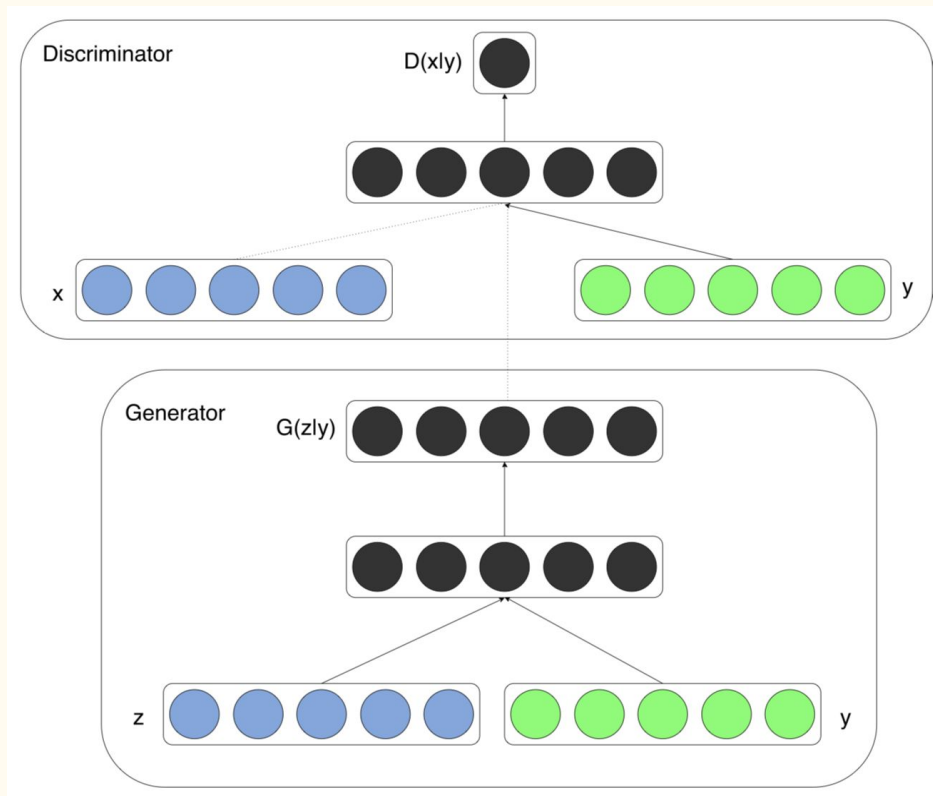
# State of the art

	a woman holding a skateboard on a street						
	0.71	0.61	0.75	0.36	1.49	0.28	0.05
	B3	B4	ROUGE	METEOR	CIDEr	SPICE	E-GAN
	0.25	0.01	0.48	0.19	0.36	0.14	0.37
	three women one with a skateboard outside a store						
	a baseball player swinging a bat at a ball						
	0.71	0.65	0.78	0.39	2.21	0.28	0.48
	B3	B4	ROUGE	METEOR	CIDEr	SPICE	E-GAN
	0.01	0.01	0.31	0.23	0.82	0.25	0.82
	the umpire stands over a catcher as the batter swings						
	a man holding a tennis racquet on a tennis court						
	0.99	0.99	1.0	1.0	3.53	0.58	0.69
	B3	B4	ROUGE	METEOR	CIDEr	SPICE	E-GAN
	0.01	0.01	0.48	0.28	1.03	0.2	0.67
	a man getting ready to serve a tennis ball						

Examples of images with two semantically similar descriptions. [1]

# Background

- Conditional GAN
  - Conditioned on some extra information  $y$ .

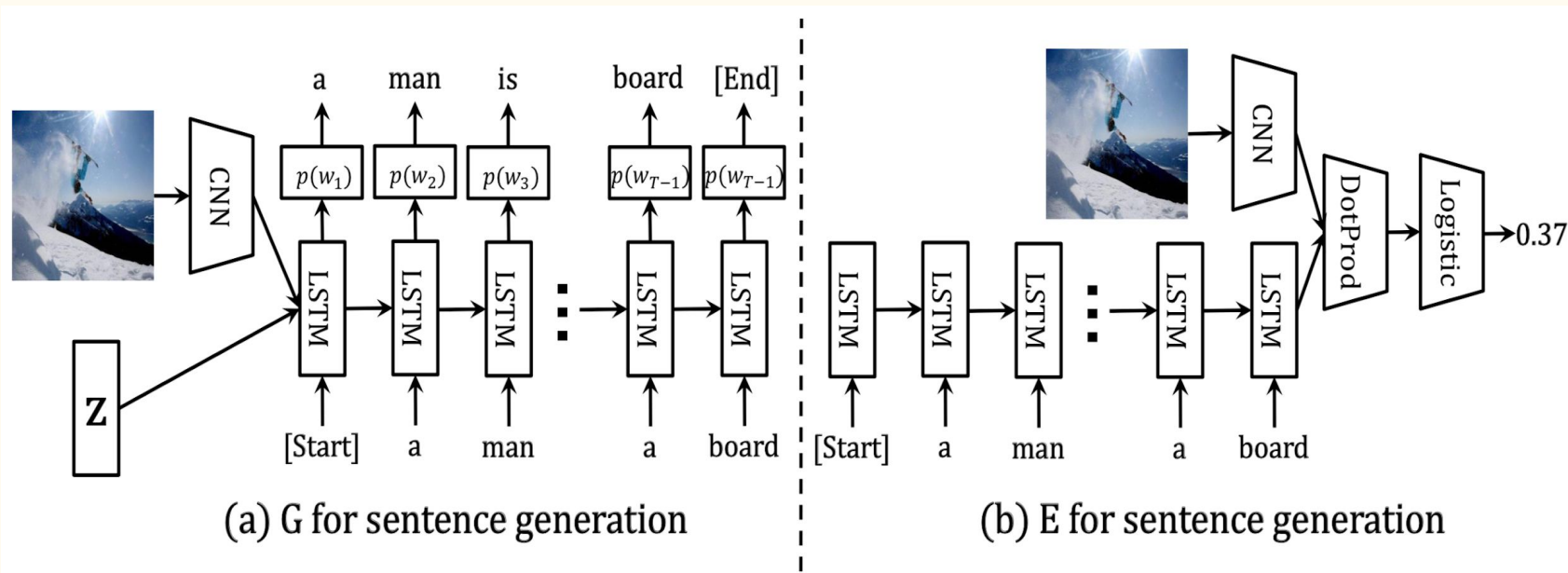


Conditional adversarial net. [3]

# Background

- Sequential Sampling
  - Non-probability sampling technique
    - Not equal chances of being selected for all members of the population
  - Picking a single or a group of objects in every time interval
  - Analyze the result, pick another sample
- Monte Carlo tree search
  - A heuristic search algorithm for some kinds of decision processes esp. Game plays
  - Steps
    - Selection
    - Expansion
    - Simulation
    - Backpropagation

# Proposed Approach



Overall structure of both generator and evaluator of a CGAN. [1]

# Overall formulation

- Evaluator
  - Quality of descriptions

$$r_{\boldsymbol{\eta}}(I, S) = \sigma(\langle \mathbf{f}(I, \boldsymbol{\eta}_I), \mathbf{h}(S, \boldsymbol{\eta}_S) \rangle)$$

- Learning objective

$$\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\eta}} \mathcal{L}(G_{\boldsymbol{\theta}}, E_{\boldsymbol{\eta}})$$

$$\mathbb{E}_{S \sim \mathcal{P}_I} [\log r_{\boldsymbol{\eta}}(I, S)] + \mathbb{E}_{\mathbf{z} \sim \mathcal{N}_0} [\log(1 - r_{\boldsymbol{\eta}}(I, G_{\boldsymbol{\theta}}(I, \mathbf{z})))]$$



The production of sentences  
is non-differentiable. How can  
we backpropagate?

—

# Challenges

- Using policy gradient for generating linguistic description
  - Sequential sampling procedure
    - Sampling a discrete token at each time step
  - Non-differentiable
- Expected future reward for early feedback using Monte Carlo rollouts
  - Vanishing gradients
  - Error propagation

# Training G

- Policy gradient
  - Action space: words
  - Conditional policy

$$\pi_{\theta}(w_t | \mathbf{f}(I), \mathbf{z}, S_{1:t-1})$$

- Reward given by the evaluator for a sequence of actions S

$$r_{\eta}(I, S)$$

# Training G

- Early feedback
  - Expected future reward

$$V_{\boldsymbol{\theta}, \boldsymbol{\eta}}(I, \mathbf{z}, S_{1:t}) = \mathbb{E}_{S_{t+1:T} \sim G_{\boldsymbol{\theta}}(I, \mathbf{z})} [r_{\boldsymbol{\eta}}(I, S_{1:t} \oplus S_{t+1:T})]$$

- Gradient of the objective w.r.t.  $\boldsymbol{\theta}$

$$\tilde{\mathbb{E}} \left[ \sum_{t=1}^{T_{\max}} \sum_{w_t \in \mathcal{V}} \nabla_{\boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(w_t | I, \mathbf{z}, S_{1:t-1}) \cdot V_{\boldsymbol{\theta}', \boldsymbol{\psi}}(I, \mathbf{z}, S_{1:t} \oplus w_t) \right]$$

# Training E

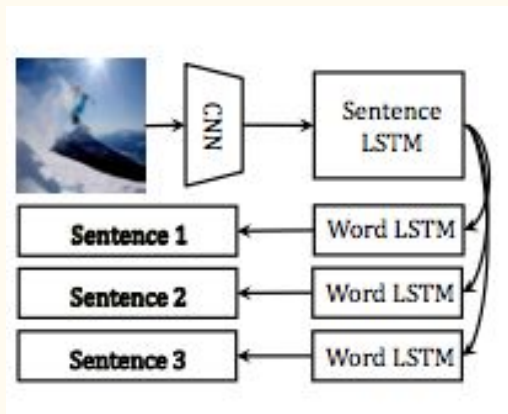
- Enforcing naturalness and semantic relevance
  - Set of descriptions provided by human
  - Descriptions from the generator
  - Human descriptions uniformly sampled from other descriptions not associated with the given image

$$\max_{\boldsymbol{\eta}} \mathcal{L}_E(\boldsymbol{\eta}) = \frac{1}{N} \sum_{i=1}^N \mathcal{L}_E(I_i; \boldsymbol{\eta})$$

$$\begin{aligned} \mathcal{L}_E(I; \boldsymbol{\eta}) = & \mathbb{E}_{S \in \mathcal{S}_I} \log r_{\boldsymbol{\eta}}(I, S) \\ & + \alpha \cdot \mathbb{E}_{S \in \mathcal{S}_G} \log(1 - r_{\boldsymbol{\eta}}(I, S)) \\ & + \beta \cdot \mathbb{E}_{S \in \mathcal{S}_{\setminus I}} \log(1 - r_{\boldsymbol{\eta}}(I, S)) \end{aligned}$$

# Generating paragraphs

- Hierarchical LSTM
  - Sentence level
  - Word level



Adding extension for paragraph generation. [1]

# Experiment

- Datasets
  - MSCOCO
  - Flickr30k
- Settings
  - Removing non-alphabet characters
  - Converting to lowercase
  - Replacing less frequent words, less than 5 times, with UNK
  - Max length of 16
  - Pretrain G for 20 epochs based on MLE
  - Pretrain E with supervised training for 5 epochs
  - Batch = 64, lr = 0.0001, n = 16 (Monte Carlo)

# Experiment

- Performance

		BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE	E-NGAN	E-GAN
COCO	human	0.290	0.192	0.240	0.465	0.849	<b>0.211</b>	0.527	<b>0.626</b>
	G-MLE	<b>0.393</b>	<b>0.299</b>	<b>0.248</b>	<b>0.527</b>	<b>1.020</b>	0.199	0.464	0.427
	G-GAN	0.305	0.207	0.224	0.475	0.795	0.182	<b>0.528</b>	0.602
Flickr	human	0.269	0.185	0.194	0.423	0.627	0.159	0.482	<b>0.464</b>
	G-MLE	<b>0.372</b>	<b>0.305</b>	<b>0.215</b>	<b>0.479</b>	<b>0.767</b>	<b>0.168</b>	0.465	0.439
	G-GAN	0.153	0.088	0.132	0.330	0.202	0.087	<b>0.582</b>	0.456

Performances of different generators on MSCOCO and Flickr30k. [1]

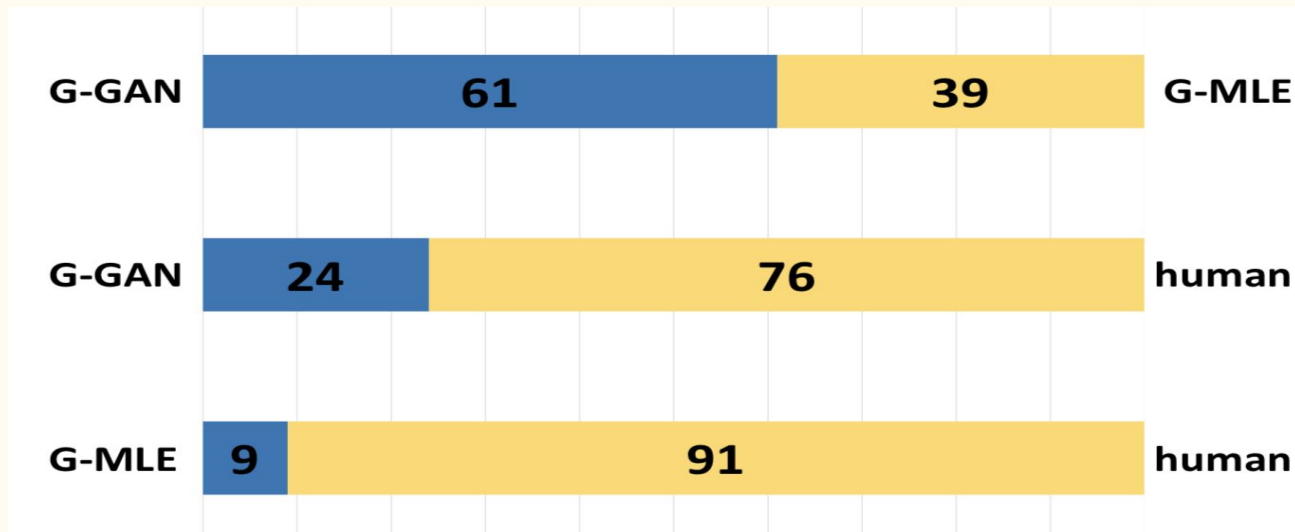


Has the proposed approach  
been able to achieve more  
natural descriptions?

—

# Experiment









- User Study



Human comparison results between each pair of generators. [1]

# Experiment

- User Study

				
G-MLE	a man flying through the air while riding a snowboard	a man flying through the air while riding a snowboard	a man flying through the air while riding a snowboard	a man flying through the air while riding a skateboard
G-GAN	a man on a skateboard in a snowy park	a man skiing down the slope near a mountain	a man performing a grind trick on a skateboard ramp	a man with stunts on his skis in the snow
				
G-MLE	a group of people standing around a boat	a group of people sitting around a table	a group of people sitting at a table	a group of people sitting around a living room
G-GAN	the bench is sitting on the ground by the water	a group of people watching each other	a table with a lot of stuff on it	furnished living room with furniture and built area

Corresponding G-GAN captions for images with similar descriptions in G-MLE. [1]

# Experiment

- Diverse descriptions

				
$\mathbf{z}_1$	a baseball player holds a bat up to hit the ball	a man riding a snowboard down a slope	a group of people sitting around a table having a meal in a restaurant	a group of men dressed in suits posing for a photo
$\mathbf{z}_2$	a baseball player holding white bat and wear blue baseball uniform	a person standing on a snowboard sliding down a hill	a young man sitting at a table with coffee and a lot of food	a couple of men standing next to each other wearing glasses
$\mathbf{z}_3$	a professional baseball player holds up his bat as he watches	a man is jumping over a snow covered hill	a pretty young man sitting next to two men in lots of people	some people dressed in costume and cups

Generated descriptions with different  $\mathbf{z}$ . [1]

# Experiment

- Evaluating semantic relevance by retrieval

		R@1	R@3	R@5	R@10
S	G-MLE	5.06	12.28	18.24	29.30
	G-GAN	<b>14.30</b>	<b>30.88</b>	<b>40.06</b>	<b>55.82</b>
P	G-MLE	9.88	20.12	27.30	39.94
	G-GAN	<b>12.04</b>	<b>23.88</b>	<b>30.70</b>	<b>41.78</b>



Image rankings for different generators. [1]

S  $\rightarrow$  E-GAN

P  $\rightarrow$  Log-likelihood

# Experiment

- Paragraph generation with different  $z$  values

	human	G-GAN, $z_1$	G-GAN, $z_2$	G-MLE
	people are on motorcycles. there are green cars behind them. the signs are all brown with chinese written on it.	men are riding on a motorcycle. the man is wearing tan boots, and a white and blue jacket with beige stripes on. the street is made of cobblestone. there are tall bright green trees on the sidewalk.	two people are riding motorcycles. there are many trees on the sidewalk. there is a red and white painted letter on the side of the ledge. tall buildings are on the background.	a man is riding a bike. there are trees on the sidewalk. there are people walking on the sidewalk. there is a tall building in the background.
	A baseball player is swinging a bat. He is wearing a black helmet and a black and white uniform. A catcher is behind him wearing a gray uniform. The catcher has a brown glove on his hand. Two men can be seen standing behind a green fence.	a baseball player in a white and blue uniform is holding a white bat. there is a umpire behind the batter in the blue and white uniform. he is getting ready to catch the ball. there is a crowd of people behind him watching him.	men are on a baseball field on a sunny day. the player is wearing a black and white uniform. there is a catcher behind him. the field is green with brown dirt and white shiny lines.	a baseball player is standing on a baseball field. he is wearing a blue helmet on his head. the catcher is wearing a black and gray uniform. the court is green with white lines.

Different paragraph descriptions generated by human, G-GAN, and G-MLE with different  $z$  values. [1]

# Failure Analysis

- Incorrect details
  - Colors
  - Counts
    - Few samples for each special detail
    - Increased risk of putting more incorrect details due to focus on diversity

Potential extensions?

—



# Future works

- Use VAE instead of GAN
- Use other similarity metrics instead of dot product in evaluator

# Presenters

- Kevin Dsouza
- Ainaz Hajimoradlou

# References

1. B. Dai, S. Fidler, R. Urtasun, and D. Lin, Towards Diverse and Natural Image Descriptions via a Conditional GAN, 2017, *arXiv preprint arXiv: 1703.06029*
2. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge, *arXiv preprint arXiv: 1609.06647*
3. M. Mirza, S. Osindero, Conditional Generative Adversarial Nets, *arXiv preprint arXiv: 1411.1784*

Thank you.

