

# Best of both worlds: Transferring knowledge from Discriminative Learning to a Generative Visual Dialog Model

Jiasen Lu, Anitha Kannan, Jianwei Yang, Devi Parikh,  
Dhruv Batra

## Overview

- Problem: MLE trained generative neural dialog models (G) produce ‘safe’, generic responses (‘I don’t know’, ‘I can’t tell’)
- Discriminative dialog models (D) trained to rank a list of candidate human responses outperform their generative counterparts; in terms of automatic metrics, diversity, and informativeness of the responses.
- However, D not useful in practice
- Their approach: best of both worlds – the practical usefulness of G and the strong performance of D – via knowledge transfer from D to G
- End-to-end trainable generative visual dialog model, where G receives gradients from D as a perceptual (not adversarial) loss of the sequence sampled from G.

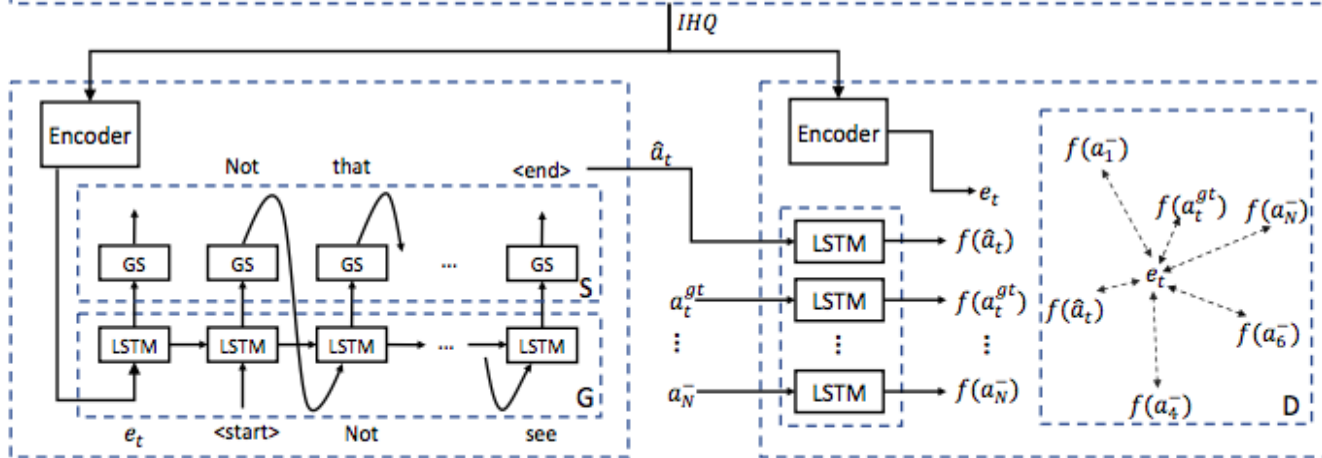
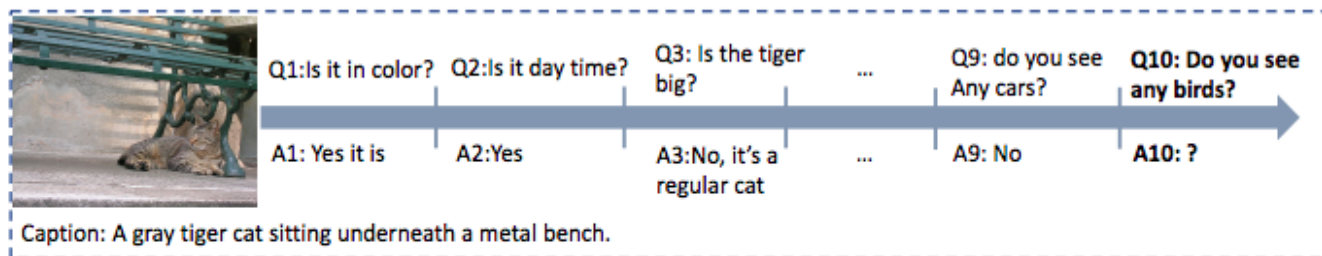
## Introduction

- Discriminative dialog model (D) receives as input a candidate list of possible responses and learns to sort this list from the training dataset
- G aims to produce a sequence that D will rank the highest in the list
- Unlike traditional GANs, discriminator receives a list of candidate responses, explicitly learns to reason about similarities and differences across candidates.
- D learns a task-dependent perceptual similarity and learns to recognize multiple correct responses in the feature space
- Employ metric-learning loss function and a self-attention answer encoding mechanism for D

## Visual Dialog

- A visual dialog model is given as input an image  $I$ , caption  $c$  describing the image, a dialog history till round  $t - 1$ , and the followup question  $q_t$  at round  $t$ . The visual dialog history is represented as  $H = (\underbrace{c}_{H_0}, \underbrace{(q_1, a_1)}_{H_1}, \dots, \underbrace{(q_{t-1}, a_{t-1})}_{H_{t-1}})$ .
- Generative models for visual dialog are trained by sequence
- Discriminative models receive both an encoding of the input, as additional input a list of 100 candidate answers  $A_t = \{a^{(1)}_t, \dots, a^{(100)}_t\}$ . Effectively learn to sort the list, hence they cannot be used at test time without a list of candidates available

# Approach

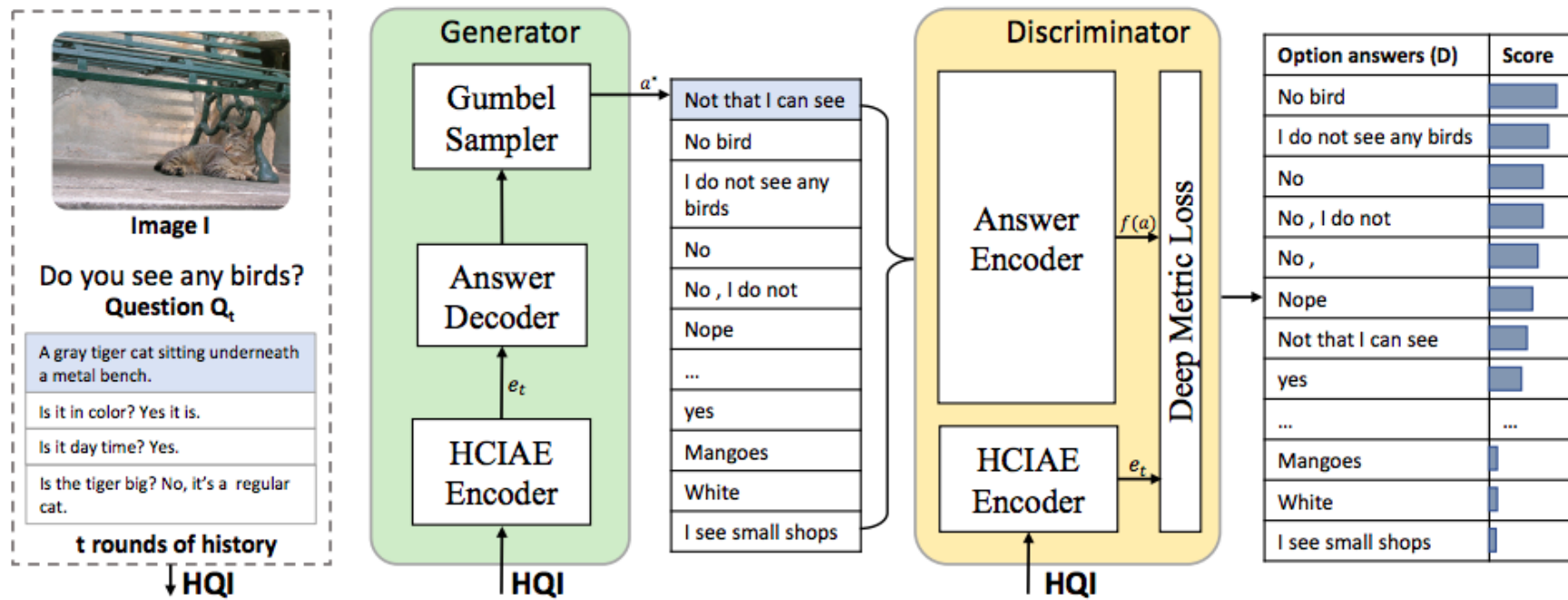


(a)

Option answers (D)	Score
No bird	■
I do not see any birds	■
No	■
<b>No , I do not</b>	■
No ,	■
Nope	■
Not at all	■
<b>Not that I can see</b>	■
yes	■
...	...
Somewhere in his 30 's	■
Mangoes	■
White	■
I see small shops	■

(b)

## Approach



## History-Conditioned Image Attentive Encoder (HCIAE)

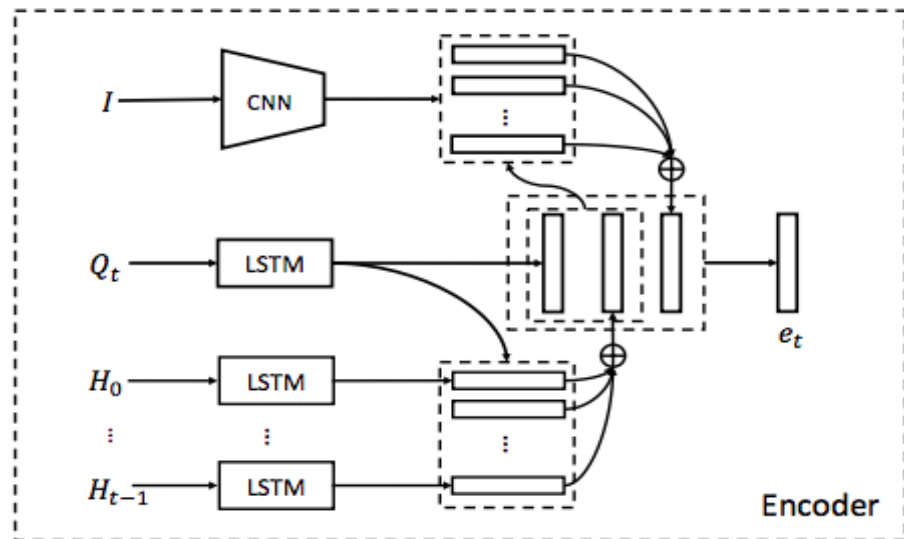
- Use of co-reference to avoid repeating entities that can be contextually resolved, nearly all (98%) dialogs involve at least one pronoun
- Uses current question to attend to exchanges in history, and then uses the question and attended history to attend to the image
- Use the spatial image features  $V \in \mathcal{R}^{d \times k}$  from a convolution layer of a CNN
- $V$  encoded with an LSTM to get a vector
- $V$  encoded separately with another LSTM as  $m_i^q \in \mathcal{R}^d$
- Conditioned on the question embedding, the model attends to the history
- $q_t$  attended representation of the history and the question  $m_i^q \in \mathcal{R}^d$  are concatenated, and used as input to attend to the image  $M_t^h \in \mathcal{R}^{d \times t}$   
 $(H_0, \dots, H_{t-1})$

## History-Conditioned Image Attentive Encoder

$$z_t^h = w_a^T \tanh(W_h M_t^h + (W_q m_t^q) \mathbb{1}^T)$$

$$\alpha_t^h = \text{softmax}(z_t^h)$$

- $\mathbb{1} \in \mathcal{R}^t$  vector with all elements 1
- $W_h, W_q \in \mathcal{R}^{t \times d}$  learned
- $w_a \in \mathcal{R}^k$  history
- $\alpha \in \mathcal{R}^k$  history feature
- concatenate  $\alpha$  and  $\hat{m}_t^h$  to get attended image feature
- Final embedding  $e_t = \tanh(W_e [m_t^q, \hat{m}_t^h, \hat{v}_t])$



$$e_t = \tanh(W_e [m_t^q, \hat{m}_t^h, \hat{v}_t]) \quad W_e \in \mathcal{R}^{d \times 3d}$$



## Discriminator Loss

- Discriminator D produces distribution over candidate answer list  $\mathcal{A}_t$
- Maximize the log-likelihood of
- Loss conducive to knowledge transfer,  $\mathbf{a}_t^{gt}$  encourages perceptually meaningful similarities
- Metric-learning multi-class N-pair loss:

$$\mathcal{L}_D = \mathcal{L}_{n-pair} \left( \{ \mathbf{e}_t, \mathbf{a}_t^{gt}, \{ \mathbf{a}_{t,i}^- \}_{i=1}^{N-1} \}, f \right) = \overbrace{\log \left( 1 + \sum_{i=1}^N \exp \left( \underbrace{\mathbf{e}_t^\top f(\mathbf{a}_{t,i}^-) - \mathbf{e}_t^\top f(\mathbf{a}_t^{gt})}_{\text{score margin}} \right) \right)}^{\text{logistic loss}}$$

- $f$  attention based LSTM encoder, helps deal with paraphrases in answer
- Attention weight is learnt through a 1-layer MLP over LSTM output at each time step

## Knowledge Transfer from D to G

- Transferring knowledge from D to G: G repeatedly queries D with answers generated for input embedding  $\mathbf{e}$  to get feedback and update itself
- G's goal : update parameters to have score higher than ground truth

$$\hat{\mathbf{a}}_t$$

$$\mathcal{L}_G = \mathcal{L}_{1-pair} \left( \{ \mathbf{e}_t, \hat{\mathbf{a}}_t, \mathbf{a}_t^{gt} \}, f \right) = \log \left( 1 + \exp \left( \mathbf{e}_t^\top f(\mathbf{a}_t^{gt}) - \mathbf{e}_t^\top f(\hat{\mathbf{a}}_t) \right) \right)$$

- Gumbel-Softmax (GS) approximation to sample answer from generator, coupled with the straight-through gradient estimator (discretize GS samples through forward pass)

## Training details

**Training Details** In our experiments, all 3 LSTMs are single layer with  $512d$  hidden state. We use VGG-19 [42] to get the representation of image. We first rescale the images to be  $224 \times 224$  pixels, and take the output of last pooling layer ( $512 \times 7 \times 7$ ) as image feature. We use the Adam optimizer with a base learning rate of  $4e-4$ . We pre-train  $G$  using standard MLE for 20 epochs, and  $D$  with supervised training based on Eq (4) for 30 epochs. Following [43], we regularize the  $L^2$  norm of the embedding vectors to be small. Subsequently, we train  $G$  with  $\mathcal{L}_G + \alpha\mathcal{L}_{MLE}$ , which is a combination of discriminative perceptual loss and MLE loss. We set  $\alpha$  to be 0.5. We found that including  $\mathcal{L}_{MLE}$  (with teacher-forcing) is important for encouraging  $G$  to generate grammatically correct responses.

## Results



**Table 1:** Results (generative) on VisDial dataset. “MRR” is mean reciprocal rank and “Mean” is mean rank.

Model	MRR	R@1	R@5	R@10	Mean
LF-G [7]	0.5199	41.83	61.78	67.59	17.07
HREA-G [7]	0.5242	42.28	62.33	68.17	16.79
MN-G [7]	0.5259	42.29	62.85	68.88	17.06
HCIAE-G-MLE	0.5386	44.06	63.55	69.24	16.01
HCIAE-G-DIS	<b>0.5467</b>	<b>44.35</b>	<b>65.28</b>	<b>71.55</b>	<b>14.23</b>

**Table 2:** Results (discriminative) on VisDial dataset.

Model	MRR	R@1	R@5	R@10	Mean
LF-D [7]	0.5807	43.82	74.68	84.07	5.78
HREA-D [7]	0.5868	44.82	74.81	84.36	5.66
MN-D [7]	0.5965	45.55	76.22	85.37	5.46
HCIAE-D-MLE	0.6140	47.73	77.50	86.35	5.15
HCIAE-D-NP	0.6182	47.98	78.35	87.16	4.92
HCIAE-D-NP-ATT	<b>0.6222</b>	<b>48.48</b>	<b>78.75</b>	<b>87.59</b>	<b>4.81</b>

**Table 3: Qualitative comparison.**

Image+Caption	Question	GT Answer	G-MLE Answer	G-DIS Answer
 <p>A couple of people in the snow on skis.</p>	Is the person male or female? About how old is the person? What color is his hair? What is he wearing? What kind of bag does he have? Is he on a mountain? Is the mountain also snow covered? Can you see clouds? Is the sun shining? Does he look happy?	Male Male Black UNK and pant with shoes Luggage bag Yes Yes, it is snow mountain Yes No Not sure	I can't tell I can't tell Brown Black jacket A small 1 Yes Yes Yes No yes	Male Can't tell Brown Snowsuit Backpack Yes Yes Yes Yes Can't tell
 <p>Two zebra standing next to each other in a dry grass field.</p>	Are zebras in zoo? Are they adults or babies? Are they standing next to each other? Is there water source in image? Is it sunny out? Any trees? Are zebras eating? Any other animals in picture? Is there any type of enclosure? Is grass green?	No zebras not in zoo I would say they are adults 1 lead in front , while other 1 No It does look like sunny There are few trees but is not in No , they are walking No No No	No They are adult Yes they are No water Yes Yes, many trees No, they are just standing around No No Yes	No They appear to be adult Yes they are I do not see any water Yes it is Yes, lots of trees in background No they are not eating No just zebras No Yes

**Table 4: Adversarial training results on VisDial dataset.**

Model	Discriminative					Generative				
	MRR	R@1	R@5	R@10	Mean	MRR	R@1	R@5	R@10	Mean
HCIAE-D-NP-ATT	0.6222	48.48	78.75	87.59	4.81	-	-	-	-	-
HCIAE-G-DIS	-	-	-	-	-	0.5467	44.35	65.28	71.55	14.23
HCIAE-GAN1	0.2177	8.82	32.97	52.14	18.53	0.5298	43.12	62.74	68.58	16.25
HCIAE-GAN2	0.6050	46.20	77.92	87.20	4.97	0.5459	44.33	65.05	71.40	14.34