

---

# Best of Both Worlds: Transferring Knowledge from Discriminative Learning to a Generative Visual Dialog Model

---

Jiasen Lu<sup>1\*</sup>, Anitha Kannan<sup>2\*</sup>, Jianwei Yang<sup>1</sup>, Devi Parikh<sup>3,1</sup>, Dhruv Batra<sup>3,1</sup>

<sup>1</sup> Georgia Institute of Technology, <sup>2</sup> Curai, <sup>3</sup> Facebook AI Research  
{jiasenlu, jw2yang, parikh, dbatra}@gatech.edu

## Abstract

We present a novel training framework for neural sequence models, particularly for grounded dialog generation. The standard training paradigm for these models is maximum likelihood estimation (MLE), or minimizing the cross-entropy of the human responses. Across a variety of domains, a recurring problem with MLE trained generative neural dialog models ( $G$ ) is that they tend to produce ‘safe’ and generic responses (*‘I don’t know’*, *‘I can’t tell’*). In contrast, discriminative dialog models ( $D$ ) that are 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,  $D$  is not useful in practice since it can not be deployed to have real conversations with users.

Our work aims to achieve the best of both worlds – the practical usefulness of  $G$  and the strong performance of  $D$  – via knowledge transfer from  $D$  to  $G$ . Our primary contribution is an 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  $G$ . We leverage the recently proposed Gumbel-Softmax (GS) approximation to the discrete distribution – specifically, a RNN augmented with a sequence of GS samplers, coupled with the straight-through gradient estimator to enable end-to-end differentiability. We also introduce a stronger encoder for visual dialog, and employ a self-attention mechanism for answer encoding along with a metric learning loss to aid  $D$  in better capturing semantic similarities in answer responses. Overall, our proposed model outperforms state-of-the-art on the VisDial dataset by a significant margin (2.67% on recall@10). The source code can be downloaded from <https://github.com/jiasenlu/visDial.pytorch>

## 1 Introduction

One fundamental goal of artificial intelligence (AI) is the development of perceptually-grounded dialog agents – specifically, agents that can perceive or understand their environment (through vision, audio, or other sensors), and communicate their understanding with humans or other agents in natural language. Over the last few years, neural sequence models (e.g. [47, 44, 46]) have emerged as the dominant paradigm across a variety of setting and datasets – from text-only dialog [44, 40, 23, 3] to more recently, visual dialog [7, 9, 8, 33, 45], where an agent must answer a sequence of questions grounded in an image, requiring it to reason about both visual content and the dialog history.

The standard training paradigm for neural dialog models is maximum likelihood estimation (MLE) or equivalently, minimizing the cross-entropy (under the model) of a ‘ground-truth’ human response. Across a variety of domains, a recurring problem with MLE trained neural dialog models is that they tend to produce ‘safe’ generic responses, such as *‘Not sure’* or *‘I don’t know’* in text-only dialog [23], and *‘I can’t see’* or *‘I can’t tell’* in visual dialog [7, 8]. One reason for this emergent behavior is that

---

\*Work was done while at Facebook AI Research.

the space of possible next utterances in a dialog is *highly* multi-modal (there are many possible paths a dialog may take in the future). In the face of such highly multi-modal output distributions, models ‘game’ MLE by latching on to the head of the distribution or the frequent responses, which by nature tend to be generic and widely applicable. Such safe generic responses break the flow of a dialog and tend to disengage the human conversing with the agent, ultimately rendering the agent useless. It is clear that novel training paradigms are needed; that is the focus of this paper.

One promising alternative to MLE training proposed by recent work [36, 27] is *sequence-level training* of neural sequence models, specifically, using reinforcement learning to optimize task-specific sequence metrics such as BLEU [34], ROUGE [24], CIDEr [48]. Unfortunately, in the case of dialog, *all existing* automatic metrics correlate poorly with human judgment [26], which renders this alternative infeasible for dialog models.

In this paper, inspired by the success of adversarial training [16], we propose to train a *generative* visual dialog model ( $G$ ) to produce sequences that score highly under a *discriminative* visual dialog model ( $D$ ). A discriminative dialog model receives as input a candidate list of possible responses and learns to sort this list from the training dataset. The generative dialog model ( $G$ ) aims to produce a sequence that  $D$  will rank the highest in the list, as shown in Fig. 1.

Note that while our proposed approach is inspired by adversarial training, there are a number of subtle but crucial differences over generative adversarial networks (GANs). Unlike traditional GANs, one novelty in our setup is that our discriminator receives a list of candidate responses and explicitly learns to reason about similarities and differences across candidates. In this process,  $D$  learns a task-dependent perceptual similarity [12, 19, 15] and learns to recognize multiple correct responses in the feature space. For example, as shown in Fig. 1 right, given the image, dialog history, and question ‘Do you see any bird?’, besides the ground-truth answer ‘No, I do not’,  $D$  can also assign high scores to other options that are valid responses to the question, including the one generated by  $G$ : ‘Not that I can see’. The interaction between responses is captured via the similarity between the learned embeddings. This similarity gives an additional signal that  $G$  can leverage in addition to the MLE loss. In that sense, our proposed approach may be viewed as an instance of ‘knowledge transfer’ [17, 5] from  $D$  to  $G$ . We employ a metric-learning loss function and a self-attention answer encoding mechanism for  $D$  that makes it particularly conducive to this knowledge transfer by encouraging perceptually meaningful similarities to emerge. This is especially fruitful since prior work has demonstrated that discriminative dialog models significantly outperform their generative counterparts, but are not as useful since they necessarily need a list of candidate responses to rank, which is only available in a dialog dataset, not in real conversations with a user. In that context, our work aims to achieve the best of both worlds – the practical usefulness of  $G$  and the strong performance of  $D$  – via this knowledge transfer.

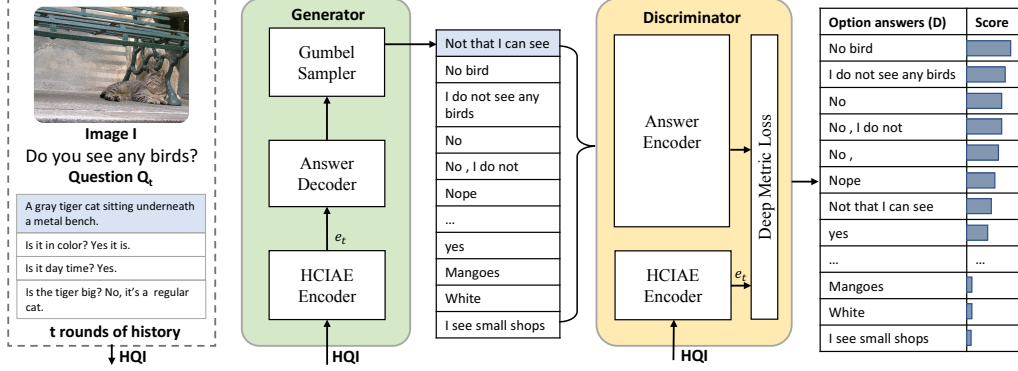
Our primary technical contribution is an end-to-end trainable generative visual dialog model, where the generator receives gradients from the discriminator loss of the sequence sampled from  $G$ . Note that this is challenging because the output of  $G$  is a sequence of discrete symbols, which naïvely is not amenable to gradient-based training. We propose to leverage the recently proposed Gumbel-Softmax (GS) approximation to the discrete distribution [18, 30] – specifically, a Recurrent Neural Network (RNN) augmented with a sequence of GS samplers, which when coupled with the straight-through gradient estimator [2, 18] enables end-to-end differentiability.

Our results show that our ‘knowledge transfer’ approach is indeed successful. Specifically, our discriminator-trained  $G$  outperforms the MLE-trained  $G$  by 1.7% on recall@5 on the VisDial dataset, essentially improving over state-of-the-art [7] by 2.43% recall@5 and 2.67% recall@10. Moreover, our generative model produces more diverse and informative responses (see Table 3).

As a side contribution specific to this application, we introduce a novel encoder for neural visual dialog models, which maintains two separate memory banks – one for visual memory (where do we look in the image?) and another for textual memory (what facts do we know from the dialog history?), and outperforms the encoders used in prior work.

## 2 Related Work

**GANs for sequence generation.** Generative Adversarial Networks (GANs) [16] have shown to be effective models for a wide range of applications involving continuous variables (*e.g.* images) *c.f.* [10, 35, 22, 55]. More recently, they have also been used for discrete output spaces such as language generation – *e.g.* image captioning [6, 41], dialog generation [23], or text generation [53] – by either viewing the generative model as a stochastic parametrized policy that is updated using REINFORCE



**Figure 1:** Model architecture of the proposed model. Given the image, history, and question, the discriminator receives as additional input a candidate list of possible responses and learns to sort this list. The generator aims to produce a sequence that discriminator will rank the highest in the list. The right most block is  $D$ 's score for different candidate answers. Note that the multiple plausible responses all score high. Image from the COCO dataset [25].

with the discriminator providing the reward [53, 6, 41, 23], or (closer to our approach) through continuous relaxation of discrete variables through Gumbel-Softmax to enable backpropagating the response from the discriminator [21, 41].

There are a few subtle but significant differences w.r.t. to our application, motivation, and approach. In these prior works, both the discriminator and the generator are trained in tandem, and from scratch. The goal of the discriminator in those settings has primarily been to discriminate ‘fake’ samples (*i.e.* generator’s outputs) from ‘real’ samples (*i.e.* from training data). In contrast, we would like to transfer knowledge from the discriminator to the generator. We start with pre-trained  $D$  and  $G$  models suited for the task, and then transfer knowledge from  $D$  to  $G$  to further improve  $G$ , while keeping  $D$  fixed. As we show in our experiments, this procedure results in  $G$  producing diverse samples that are close in the embedding space to the ground truth, due to perceptual similarity learned in  $D$ . One can also draw connections between our work and Energy Based GAN (EBGAN) [54] – without the adversarial training aspect. The “energy” in our case is a deep metric-learning based scoring mechanism, instantiated in the visual dialog application.

**Modeling image and text attention.** Models for tasks at the intersection of vision and language – *e.g.*, image captioning [11, 13, 20, 49], visual question answering [1, 14, 31, 37], visual dialog [7, 9, 8, 45, 33] – typically involve attention mechanisms. For image captioning, this may be attending to relevant regions in the image [49, 51, 28]. For VQA, this may be attending to relevant image regions alone [4, 50, 52] or co-attending to image regions and question words/phrases [29].

In the context of visual dialog, [7] uses attention to identify utterances in the dialog history that may be useful for answering the current question. However, when modeling the image, the entire image embedding is used to obtain the answer. In contrast, our proposed encoder HCIAE (Section 4.1) localizes the region in the image that can help reliably answer the question. In particular, in addition to the history and the question guiding the image attention, our visual dialog encoder also reasons about the history when identifying relevant regions of the image. This allows the model to implicitly resolve co-references in the text and ground them back in the image.

### 3 Preliminaries: Visual Dialog

We begin by formally describing the visual dialog task setup as introduced by Das *et al.* [7]. The machine learning task is as follows. A visual dialog model is given as input an image  $I$ , caption  $c$  describing the image, a dialog history till round  $t - 1$ ,  $\mathbf{H} = (\underbrace{c}_{H_0}, \underbrace{(q_1, a_1)}_{H_1}, \dots, \underbrace{(q_{t-1}, a_{t-1})}_{H_{t-1}})$ , and

the followup question  $q_t$  at round  $t$ . The visual dialog agent needs to return a valid response to the question.

Given the problem setup, there are two broad classes of methods – generative and discriminative models. Generative models for visual dialog are trained by maximizing the log-likelihood of the ground truth answer sequence  $a_t^{gt} \in \mathcal{A}_t$  given the encoded representation of the input  $(I, \mathbf{H}, q_t)$ .

On the other hand, discriminative models receive both an encoding of the input  $(I, H, q_t)$  and as additional input a list of 100 candidate answers  $\mathcal{A}_t = \{a_t^{(1)}, \dots, a_t^{(100)}\}$ . These models effectively learn to sort the list. Thus, by design, they cannot be used at test time without a list of candidates available.

#### 4 Approach: Backprop Through Discriminative Losses for Generative Training

In this section, we describe our approach to transfer knowledge from a discriminative visual dialog model ( $D$ ) to generative visual dialog model ( $G$ ). Fig. 1 (a) shows the overview of our approach. Given the input image  $I$ , dialog history  $H$ , and question  $q_t$ , the encoder converts the inputs into a joint representation  $e_t$ . The generator  $G$  takes  $e_t$  as input, and produces a distribution over answer sequences via a recurrent neural network (specifically an LSTM). At each word in the answer sequence, we use a Gumbel-Softmax sampler  $S$  to sample the answer token from that distribution. The discriminator  $D$  in it's standard form takes  $e_t$ , ground-truth answer  $a_t^{gt}$  and  $N - 1$  "negative" answers  $\{a_{t,i}^-\}_{i=1}^{N-1}$  as input, and learns an embedding space such that  $\text{similarity}(e_t, f(a_t^{gt})) > \text{similarity}(e_t, f(a_{t,i}^-))$ , where  $f(\cdot)$  is the embedding function. When we enable the communication between  $D$  and  $G$ , we feed the sampled answer  $\hat{a}_t$  into discriminator, and optimize the generator  $G$  to produce samples that get higher scores in  $D$ 's metric space.

We now describe each component of our approach in detail.

##### 4.1 History-Conditioned Image Attentive Encoder (HCIAE)

An important characteristic in dialogs is the use of co-reference to avoid repeating entities that can be contextually resolved. In fact, in the VisDial dataset [7] nearly all (98%) dialogs involve at least one pronoun. This means that for a model to correctly answer a question, it would require a reliable mechanism for co-reference resolution.

A common approach is to use an encoder architecture with an attention mechanism that implicitly performs co-reference resolution by identifying the portion of the dialog history that can help in answering the current question [7, 38, 39, 32]. while using a holistic representation for the image. Intuitively, one would also expect that the answer is also localized to regions in the image, and be consistent with the attended history.

With this motivation, we propose a novel encoder architecture (called HCIAE) shown in Fig. 2. Our encoder first uses the current question to attend to the exchanges in the history, and then use the question and attended history to attend to the image, so as to obtain the final encoding.

Specifically, we use the spatial image features  $V \in \mathcal{R}^{d \times k}$  from a convolution layer of a CNN.  $q_t$  is encoded with an LSTM to get a vector  $m_t^q \in \mathcal{R}^d$ . Simultaneously, each previous round of history  $(H_0, \dots, H_{t-1})$  is encoded separately with another LSTM as  $M_t^h \in \mathcal{R}^{d \times t}$ . Conditioned on the question embedding, the model attends to the history. The attended representation of the history and the question embedding are concatenated, and used as input to attend to the image:

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

$$\alpha_t^h = \text{softmax}(z_t^h) \quad (2)$$

where  $\mathbb{1} \in \mathcal{R}^t$  is a vector with all elements set to 1.

$W_h, W_q \in \mathcal{R}^{t \times d}$  and  $w_a \in \mathcal{R}^k$  are parameters to be learned.  $\alpha \in \mathcal{R}^k$  is the attention weight over history. The attended history feature  $\hat{m}_t^h$  is a convex combination of columns of  $M_t$ , weighted appropriately by the elements of  $\alpha_t^h$ . We further concatenate  $m_t^q$  and  $\hat{m}_t^h$  as the query vector and get the attended image feature  $\hat{v}_t$  in the similar manner. Subsequently, all three components are used to obtain the final embedding  $e_t$ :

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

where  $W_e \in \mathcal{R}^{d \times 3d}$  is weight parameters and  $[\cdot]$  is the concatenation operation.

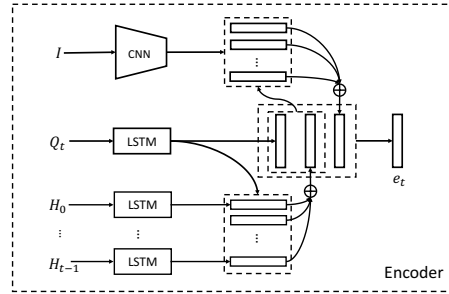


Figure 2: Structure of the proposed encoder.

## 4.2 Discriminator Loss

Discriminative visual dialog models produce a distribution over the candidate answer list  $\mathcal{A}_t$  and maximize the log-likelihood of the correct option  $\mathbf{a}_t^{gt}$ . The loss function for  $D$  needs to be conducive for knowledge transfer. In particular, it needs to encourage perceptually meaningful similarities. Therefore, we use a metric-learning multi-class N-pair loss [43] defined as:

$$\mathcal{L}_D = \mathcal{L}_{n-pair}(\{\mathbf{e}_t, \mathbf{a}_t^{gt}, \{\mathbf{a}_{t,i}^-\}_{i=1}^{N-1}\}, f) = \log \left( 1 + \overbrace{\sum_{i=1}^N \exp \left( \underbrace{\mathbf{e}_t^\top f(\mathbf{a}_{t,i}^-) - \mathbf{e}_t^\top f(\mathbf{a}_t^{gt})}_{\text{score margin}} \right)}^{\text{logistic loss}} \right) \quad (4)$$

where  $f$  is an attention based LSTM encoder for the answer. This attention can help the discriminator better deal with paraphrases across answers. The attention weight is learnt through a 1-layer MLP over LSTM output at each time step. The N-pair loss objective encourages learning a space in which the ground truth answer is scored higher than other options, and at the same time, encourages options similar to ground truth answers to score better than dissimilar ones. This means that, unlike the multiclass logistic loss, the options that are correct but different from the correct option may not be overly penalized, and thus can be useful in providing a reliable signal to the generator. See Fig. 1 for an example. Following [43], we regularize the L2 norm of the embedding vectors to be small.

## 4.3 Discriminant Perceptual Loss and Knowledge Transfer from $D$ to $G$

At a high-level, our approach for transferring knowledge from  $D$  to  $G$  is as follows:  $G$  repeatedly queries  $D$  with answers  $\hat{\mathbf{a}}_t$  that it generates for an input embedding  $\mathbf{e}_t$  to get feedback and update itself. In each such update,  $G$ 's goal is to update its parameters to try and have  $\hat{\mathbf{a}}_t$  score higher than the correct answer,  $\mathbf{a}_t^{gt}$ , under  $D$ 's learned embedding and scoring function. Formally, the perceptual loss that  $G$  aims to optimize is given by:

$$\mathcal{L}_G = \mathcal{L}_{1-pair}(\{\mathbf{e}_t, \hat{\mathbf{a}}_t, \mathbf{a}_t^{gt}\}, f) = \log \left( 1 + \exp \left( \mathbf{e}_t^\top f(\mathbf{a}_t^{gt}) - \mathbf{e}_t^\top f(\hat{\mathbf{a}}_t) \right) \right) \quad (5)$$

where  $f$  is the embedding function learned by the discriminator as in (4). Intuitively, updating generator parameters to minimize  $\mathcal{L}_G$  can be interpreted as learning to produce an answer sequence  $\hat{\mathbf{a}}_t$  that 'fools' the discriminator into believing that this answer should score higher than the human response  $\mathbf{a}_t^{gt}$  under the discriminator's learned embedding  $f(\cdot)$  and scoring function.

While it is straightforward to sample an answer  $\hat{\mathbf{a}}_t$  from the generator and perform a forward pass through the discriminator, naïvely, it is not possible to backpropagate the gradients to the generator parameters since sampling discrete symbols results in zero gradients w.r.t. the generator parameters. To overcome this, we leverage the recently introduced continuous relaxation of the categorical distribution – the Gumbel-softmax distribution or the Concrete distribution [18, 30].

At an intuitive level, the Gumbel-Softmax (GS) approximation uses the so called 'Gumbel-Max trick' to reparametrize sampling from a categorical distribution and replaces argmax with softmax to obtain a continuous relaxation of the discrete random variable. Formally, let  $\mathbf{x}$  denote a  $K$ -ary categorical random variable with parameters denoted by  $(p_1, \dots, p_K)$ , or  $\mathbf{x} \sim \text{Cat}(\mathbf{p})$ . Let  $(g_i)_{i=1}^K$  denote  $K$  IID samples from the standard Gumbel distribution,  $g_i \sim F(g) = e^{-e^{-g}}$ . Now, a sample from the Concrete distribution can be produced via the following transformation:

$$y_i = \frac{e^{(\log p_i + g_i)/\tau}}{\sum_{j=1}^K e^{(\log p_j + g_j)/\tau}} \quad \forall i \in \{1, \dots, K\} \quad (6)$$

where  $\tau$  is a temperature parameter that control how close samples  $\mathbf{y}$  from this Concrete distribution approximate the one-hot encoding of the categorical variable  $\mathbf{x}$ .

As illustrated in Fig. 1, we augment the LSTM in  $G$  with a sequence of GS samplers. Specifically, at each position in the answer sequence, we use a GS sampler to sample an answer token from that conditional distribution. When coupled with the straight-through gradient estimator [2, 18] this enables end-to-end differentiability. Specifically, during the forward pass we discretize the GS samples into discrete samples, and in the backward pass use the continuous relaxation to compute gradients. In our experiments, we held the temperature parameter fixed at 0.5.

## 5 Experiments

**Dataset and Setup.** We evaluate our proposed approach on the VisDial dataset [7], which was collected by Das *et al.* by pairing two subjects on Amazon Mechanical Turk to chat about an image. One person was assigned the role of a ‘questioner’ and the other of ‘answerer’. One worker (the questioner) sees only a single line of text describing an image (caption from COCO [25]); the image remains hidden to the questioner. Their task is to ask questions about this hidden image to “imagine the scene better”. The second worker (the answerer) sees the image and caption and answers the questions. The two workers take turns asking and answering questions for 10 rounds. We perform experiments on VisDial v0.9 (the latest available release) containing 83k dialogs on COCO-train and 40k on COCO-val images, for a total of 1.2M dialog question-answer pairs. We split the 83k into 82k for `train`, 1k for `val`, and use the 40k as `test`, in a manner consistent with [7]. The caption is considered to be the first round in the dialog history.

**Evaluation Protocol.** Following the evaluation protocol established in [7], we use a retrieval setting to evaluate the responses at each round in the dialog. Specifically, every question in VisDial is coupled with a list of 100 candidate answer options, which the models are asked to sort for evaluation purposes.  $D$  uses its score to rank these answer options, and  $G$  uses the log-likelihood of these options for ranking. Models are evaluated on standard retrieval metrics – (1) mean rank, (2) recall @ $k$ , and (3) mean reciprocal rank (MRR) – of the human response in the returned sorted list.

**Pre-processing.** We truncate captions/questions/answers longer than 24/16/8 words respectively. We then build a vocabulary of words that occur at least 5 times in `train`, resulting in 8964 words.

**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.

### 5.1 Results and Analysis

**Baselines.** We compare our proposed techniques to the current state-of-art generative and discriminative models developed in [7]. Specifically, [7] introduced 3 encoding architectures – Late Fusion (**LF**), Hierarchical Recurrent Encoder (**HRE**), Memory Network (**MN**) – each trained with a generative (**-G**) and discriminative (**-D**) decoder. We compare to all 6 models.

**Our approaches.** We present a few variants of our approach to systematically study the individual contributions of our training procedure, novel encoder (HCIAE), self-attentive answer encoding (ATT), and metric-loss (NP).

- **HCIAE-G-MLE** is a generative model with our proposed encoder trained under the MLE objective. Comparing this variant to the generative baselines from [7] establishes the improvement due to our encoder (HCIAE).
- **HCIAE-G-DIS** is a generative model with our proposed encoder trained under the mixed MLE and discriminator loss (knowledge transfer). This forms our best generative model. Comparing this model to **HCIAE-G-MLE** establishes the improvement due to our discriminative training.
- **HCIAE-D-MLE** is a discriminative model with our proposed encoder, trained under the standard discriminative cross-entropy loss. The answer candidates are encoded using an LSTM (no attention). Comparing this variant to the discriminative baselines from [7] establishes the improvement due to our encoder (HCIAE) in the discriminative setting.
- **HCIAE-D-NP** is a discriminative model with our proposed encoder, trained under the n-pair discriminative loss (as described in Section 4.2). The answer candidates are encoded using an LSTM (no attention). Comparing this variant to **HCIAE-D-MLE** establishes the improvement due to the n-pair loss.
- **HCIAE-D-NP-ATT** is a discriminative model with our proposed encoder, trained under the n-pair discriminative loss (as described in Section 4.2), and using the self-attentive answer encoding. Comparing this variant to **HCIAE-D-NP** establishes the improvement due to the self-attention mechanism while encoding the answers.

**Table 1:** Results (generative) on VisDial dataset. “MRR” is mean reciprocal rank and “Mean” is mean rank. **Table 2:** Results (discriminative) on VisDial dataset.

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>

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>

**Results.** Tables 1, 2 present results for all our models and baselines in generative and discriminative settings. The key observations are:

1. **Main Results for HCIAE-G-DIS:** Our final generative model with all ‘bells and whistles’, **HCIAE-G-DIS**, uniformly performs the best under all the metrics, outperforming the previous state-of-art model **MN-G** by 2.43% on R@5. This shows the importance of the knowledge transfer from the discriminator and the benefit from our encoder architecture.
2. **Knowledge transfer vs. encoder for  $G$ :** To understand the relative importance of the proposed history conditioned image attentive encoder (HCIAE) and the knowledge transfer, we compared the performance of **HCIAE-G-DIS** with **HCIAE-G-MLE**, which uses our proposed encoder but without any feedback from the discriminator. This comparison highlights two points: first, **HCIAE-G-MLE** improves R@5 by 0.7% over the current state-of-art method (**MN-D**) confirming the benefits of our encoder. Secondly, and importantly, its performance is lower than **HCIAE-G-DIS** by 1.7% on R@5, confirming that the modifications to encoder alone will not be sufficient to gain improvements in answer generation; knowledge transfer from  $D$  greatly improves  $G$ .
3. **Metric loss vs. self-attentive answer encoding:** In the purely discriminative setting, our final discriminative model (**HCIAE-D-NP-ATT**) also beats the performance of the corresponding state-of-art models [7] by 2.53% on R@5. The n-pair loss used in the discriminator is not only helpful for knowledge transfer but it also improves the performance of the discriminator by 0.85% on R@5 (compare **HCIAE-D-NP** to **HCIAE-D-MLE**). The improvements obtained by using the answer attention mechanism leads to an additional, albeit small, gains of 0.4% on R@5 to the discriminator performance (compare **HCIAE-D-NP** to **HCIAE-D-NP-ATT**).





## 5.2 Does updating discriminator help?

Recall that our model training happens as follows: we independently train the generative model **HCIAE-G-MLE** and the discriminative model **HCIAE-D-NP-ATT**. With **HCIAE-G-MLE** as the initialization, the generative model is updated based on the feedback from **HCIAE-D-NP-ATT** and this results in our final **HCIAE-G-DIS**.

We performed two further experiments to answer the following questions:

- What happens if we continue training **HCIAE-D-NP-ATT** in an adversarial setting? In particular, we continue training by maximizing the score of the ground truth answer  $\alpha_i^{gt}$  and minimizing the score of the generated answer  $\hat{\alpha}_t$ , effectively setting up an adversarial training regime  $\mathcal{L}_D = -\mathcal{L}_G$ . The resulting discriminator **HCIAE-GAN1** has significant drop in performance, as can be seen in Table. 4 (32.97% R@5). This is perhaps expected because **HCIAE-GAN1** updates its parameters based on only two answers, the ground truth and the generated sample (which is likely to be similar to ground truth). This wrecks the structure that **HCIAE-D-NP-ATT** had previously learned by leveraging additional incorrect options.
- What happens if we continue structure-preserving training of **HCIAE-D-NP-ATT**? In addition to providing **HCIAE-D-NP-ATT** samples from  $G$  as fake answers, we also include incorrect options as negative answers so that the structure learned by the discriminator is preserved. **HCIAE-D-NP-ATT** continues to train under loss  $\mathcal{L}_D$ . In this case (**HCIAE-GAN2** in Table. 4), we find that there is a small improvement in the performance of  $G$ . The additional computational overhead to training the discriminator supersedes the performance improvement. Also note that **HCIAE-D-NP-ATT** itself gets worse at the dialog task.

**Table 3:** Qualitative comparison. “Ours” are samples from G-DIS model with different gumbel noise  $z$ . Images from the COCO dataset [25].

			
<p><b>Q:</b> Is it a home or restaurant?</p> <p><b>A:</b> I think restaurant.</p> <p><b>G-MLE:</b> I can't tell.</p> <p><b>Ours</b> (<math>z1</math>): Hard to say.</p> <p><b>Ours</b> (<math>z2</math>): It looks like a restaurant.</p> <p><b>Ours</b> (<math>z3</math>): I can't tell because it is too close.</p>	<p><b>Q:</b> Can you see his face?</p> <p><b>A:</b> I am not sure.</p> <p><b>G-MLE:</b> Yes.</p> <p><b>Ours</b> (<math>z1</math>): I can only see the back of his body.</p> <p><b>Ours</b> (<math>z2</math>): No.</p> <p><b>Ours</b> (<math>z3</math>): No, he's too far away.</p>	<p><b>Q:</b> How old does the man seem to be?</p> <p><b>A:</b> 20's.</p> <p><b>G-MLE:</b> Late teens.</p> <p><b>Ours</b> (<math>z1</math>): On his 20s.</p> <p><b>Ours</b> (<math>z2</math>): In his twenties.</p> <p><b>Ours</b> (<math>z3</math>): Mid 20's.</p>	<p><b>Q:</b> Can you see broccoli?</p> <p><b>A:</b> Yes, 3 larger pieces and some small.</p> <p><b>G-MLE:</b> No, just the broccoli.</p> <p><b>Ours</b> (<math>z1</math>): I can see broccoli, slightly butter.</p> <p><b>Ours</b> (<math>z2</math>): Yes, there is broccoli.</p> <p><b>Ours</b> (<math>z3</math>): Yes, broccoli is green.</p>

**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

One might wonder, why not train a GAN for visual dialog? Formulating the task in a GAN setting would involve  $G$  and  $D$  training in tandem with  $D$  providing feedback as to whether a response that  $G$  generates is real or fake. We found this to be a particularly unstable setting, for two main reasons: First, consider the case when the ground truth answer and the generated answers are the same. This happens for answers that are typically short or ‘cryptic’ (e.g. ‘yes’). In this case,  $D$  can not train itself or provide feedback, as the answer is labeled both positive and negative. Second, in cases where the ground truth answer is descriptive but the generator provides a short answer,  $D$  can quickly become powerful enough to discard generated samples as fake. In this case,  $D$  is not able to provide any information to  $G$  to get better at the task. Our experience suggests that the discriminator, if one were to consider a ‘GANs for visual dialog’ setting, can not merely be focused on differentiating fake from real. It needs to be able to score similarity between the ground truth and other answers. Such a scoring mechanism provides a more reliable feedback to  $G$ . In fact, as we show in the previous two results, a pre-trained  $D$  that captures this structure is the key ingredient in sharing knowledge with  $G$ . The adversarial training of  $D$  is not central.

### 5.3 Qualitative Comparison

In Table 3 we present a couple of qualitative examples that compares the responses generated by G-MLE and G-DIS. G-MLE predominantly produces ‘safe’ and less informative answers, such as ‘Yes’ and or ‘I can’t tell’. In contrast, our proposed model G-DIS does so less frequently, and often generates more diverse yet informative responses.

## 6 Conclusion

Generative models for (visual) dialog are typically trained with an MLE objective. As a result, they tend to latch on to safe and generic responses. Discriminative (or retrieval) models on the other hand have been shown to significantly outperform their generative counterparts. However, discriminative models can not be deployed as dialog agents with a real user where canned candidate responses are not available. In this work, we propose transferring knowledge from a powerful discriminative visual dialog model to a generative model. We leverage the Gumbel-Softmax (GS) approximation to the discrete distribution –specifically, a RNN augmented with a sequence of GS samplers, coupled with a ST gradient estimator for end-to-end differentiability. We also propose a novel visual dialog encoder that reasons about image-attention informed by the history of the dialog; and employ a metric learning loss along with a self-attentive answer encoding to enable the discriminator to learn meaningful structure in dialog responses. The result is a generative visual dialog model that significantly outperforms state-of-the-art.



## References

- [1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015.
- [2] Yoshua Bengio, Nicholas Léonard, and Aaron C. Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *CoRR*, abs/1308.3432, 2013.
- [3] Antoine Bordes and Jason Weston. Learning end-to-end goal-oriented dialog. *arXiv preprint arXiv:1605.07683*, 2016.
- [4] Kan Chen, Jiang Wang, Liang-Chieh Chen, Haoyuan Gao, Wei Xu, and Ram Nevatia. Abc-cnn: An attention based convolutional neural network for visual question answering. *arXiv preprint arXiv:1511.05960*, 2015.
- [5] Tianqi Chen, Ian Goodfellow, and Jonathon Shlens. Net2net: Accelerating learning via knowledge transfer. *arXiv preprint arXiv:1511.05641*, 2015.
- [6] Bo Dai, Dahua Lin, Raquel Urtasun, and Sanja Fidler. Towards diverse and natural image descriptions via a conditional gan. *arXiv preprint arXiv:1703.06029*, 2017.
- [7] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [8] Abhishek Das, Satwik Kottur, José MF Moura, Stefan Lee, and Dhruv Batra. Learning cooperative visual dialog agents with deep reinforcement learning. *arXiv preprint arXiv:1703.06585*, 2017.
- [9] Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. Guesswhat?! visual object discovery through multi-modal dialogue. *arXiv preprint arXiv:1611.08481*, 2016.
- [10] Emily L. Denton, Soumith Chintala, Arthur Szlam, and Robert Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. *Neural Information Processing Systems*, 2015.
- [11] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term Recurrent Convolutional Networks for Visual Recognition and Description. In *CVPR*, 2015.
- [12] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. In *Advances in Neural Information Processing Systems*, pages 658–666, 2016.
- [13] Hao Fang, Saurabh Gupta, Forrest N. Iandola, Rupesh Kumar Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. From Captions to Visual Concepts and Back. In *CVPR*, 2015.
- [14] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. Are you talking to a machine? dataset and methods for multilingual image question answering. In *NIPS*, 2015.
- [15] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.
- [16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- [17] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
- [18] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.
- [19] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016.
- [20] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.
- [21] Matt J. Kusner and José Miguel Hernández-Lobato. Gans for sequences of discrete elements with the gumbel-softmax distribution. *CoRR*, abs/1611.04051, 2016.
- [22] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. *CoRR*, abs/1609.04802, 2016.
- [23] Jiwei Li, Will Monroe, Tianlin Shi, Alan Ritter, and Dan Jurafsky. Adversarial learning for neural dialogue generation. *arXiv preprint arXiv:1701.06547*, 2017.
- [24] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *ACL 2004 Workshop*, 2004.

- [25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.
- [26] Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*, 2016.
- [27] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Optimization of image description metrics using policy gradient methods. *arXiv preprint arXiv:1612.00370*, 2016.
- [28] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. *CoRR*, abs/1612.01887, 2016.
- [29] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297, 2016.
- [30] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.
- [31] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *ICCV*, 2015.
- [32] Hongyuan Mei, Mohit Bansal, and Matthew R Walter. Coherent dialogue with attention-based language models. *arXiv preprint arXiv:1611.06997*, 2016.
- [33] Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios P Spithourakis, and Lucy Vanderwende. Image-grounded conversations: Multimodal context for natural question and response generation. *arXiv preprint arXiv:1701.08251*, 2017.
- [34] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.
- [35] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015.
- [36] Marc’Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*, 2015.
- [37] Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. In *NIPS*, 2015.
- [38] Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. *arXiv preprint arXiv:1507.04808*, 2015.
- [39] Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. A hierarchical latent variable encoder-decoder model for generating dialogues. *arXiv preprint arXiv:1605.06069*, 2016.
- [40] Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
- [41] Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. Speaking the same language: Matching machine to human captions by adversarial training. *CoRR*, abs/1703.10476, 2017.
- [42] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [43] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*, pages 1849–1857, 2016.
- [44] Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. A neural network approach to context-sensitive generation of conversational responses. *arXiv preprint arXiv:1506.06714*, 2015.
- [45] Florian Strub, Harm de Vries, Jeremie Mary, Bilal Piot, Aaron Courville, and Olivier Pietquin. End-to-end optimization of goal-driven and visually grounded dialogue systems. *arXiv preprint arXiv:1703.05423*, 2017.
- [46] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. In *Advances in neural information processing systems*, pages 2440–2448, 2015.
- [47] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.
- [48] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015.

- [49] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2015.
- [50] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European Conference on Computer Vision*, pages 451–466. Springer, 2016.
- [51] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *CoRR*, abs/1502.03044, 2015.
- [52] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 21–29, 2016.
- [53] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: Sequence generative adversarial nets with policy gradient. *AAAI Conference on Artificial Intelligence*, 2017.
- [54] Junbo Jake Zhao, Michaël Mathieu, and Yann LeCun. Energy-based generative adversarial network. *CoRR*, abs/1609.03126, 2016.
- [55] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*, 2017.