FLIPDIAL: A Generative Model for Two-Way Visual Dialogue

Daniela Massiceti
University of Oxford, UK
daniela@robots.ox.ac.uk

N. Siddharth
University of Oxford, UK
nsid@robots.ox.ac.uk

Puneet K. Dokania
University of Oxford, UK
puneet@robots.ox.ac.uk

Philip H.S. Torr
University of Oxford, UK
phst@robots.ox.ac.uk

Question	Ground-truth answer	z_1	z_2	z_3	z_4	z_5	z_6
How old is the man?	I can't see his face, but maybe mid or late twenties	He looks thirty	Twenties	I can't tell	Thirtys	I say teenager	Twentys
What race is the man?	White	White	White	In five-thirty away	I can't see	Caucasian	He looks white
Is he wearing a hat?	I don't see his head	No	Yes	Yes	Yes	No	No
Is he wearing a shirt?	Yes	Yes	Yes	No	No	Yes	It
What color?	Dark grey	Grey	Grey	Black	White	Black	Black
What gender is the baby?	I can't really tell, maybe a girl	I	Boy	I	Male	I	I
What is the baby wearing?	A bib	Shirt	T-shirt has pants	Shirt and shirt	Shirt and shirt	Looks like white	Shirt and shirt
What color is the remote?	White and black	White	Silver	Silver	It is white and black	White	Black
Is the chair wood?	No, it's leather	No	No	Yes	No it's a chair	No	Yes
What color is the chair?	Like a light burgundy	Brown	White has white checkered	A light brown	Gray	Brown with white texture	Gray

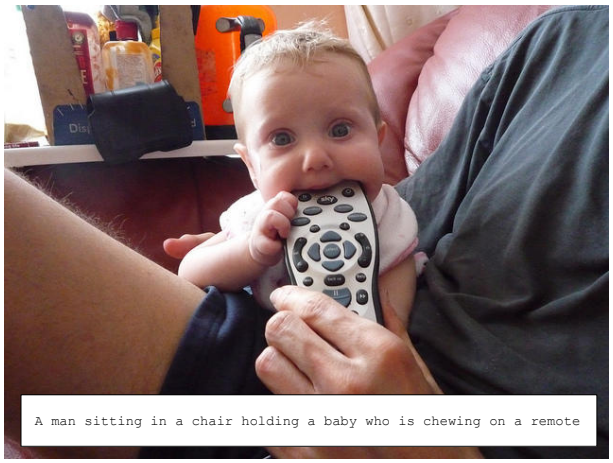

Figure 1: Diverse answers generated by FLIPDIAL in the one-way visual dialogue (1VD) task. For a given time step (row), each column shows a *generated* answer to the current question. Answers are obtained by decoding a latent $\boldsymbol{z}_i$ sampled from the conditional prior – with conditions being the image, caption and dialogue history up until that time step.

Abstract

We present FLIPDIAL*, a generative model for Visual Dialogue that simultaneously plays the role of both participants in a visually-grounded dialogue. Given context in the form of an image and an associated caption summarising the contents of the image,* FLIPDIAL *learns* both *to answer questions and put forward questions, capable of generating entire sequences of dialogue (question-answer pairs) which are diverse and relevant to the image. To do this,* FLIPDIAL *relies on a simple but surprisingly powerful idea: it uses convolutional neural networks (CNNs) to encode entire dialogues directly, implicitly capturing dialogue context, and conditional VAEs to learn the generative model.* FLIPDIAL *outperforms the state-of-the-art model in the sequential answering task (*1VD*) on the VisDial dataset by 5 points in Mean Rank using the generated answers. We are the first to extend this paradigm to full two-way visual dialogue (*2VD*), where our model is capable of generating both questions* and *answers in sequence based on a visual input, for which we propose a set of novel evaluation measures and metrics.*

1. Introduction

A fundamental characteristic of a good human-computer interaction (HCI) system is its ability to effectively acquire *and* disseminate knowledge about the tasks and environments in which it is involved. A particular subclass of such systems, natural-language-driven conversational agents such as *Alexa* and *Siri*, have seen great success in a number of well-defined language-driven tasks. Even such widely adopted systems suffer, however, when exposed to less circumscribed, more free-form situations. Ultimately, an implicit requirement for the wide-scale success of such systems is the effective understanding of the environments and goals of the user – an exceedingly difficult problem in the general case as it involves getting to grips with a variety of sub-problems (semantics, grounding, long-range dependencies) each of which are extremely difficult problems in themselves. One avenue to ameliorate such issues is the incorporation of *visual* context to help explicitly ground the language used – providing a domain in which knowledge can be anchored and extracted from. Conversely, this also provides a way in which language can be used to characterise visual information in richer terms,

for example with sentences describing salient features in the image (referred to as "captioning") [13, 15].

In recent years, there has been considerable interest in visually-guided language generation in the form of visual question-answering (VQA) [1] and subsequently visual dialogue [6], both involving the task of *answering* questions in the context of an image. In the particular case of visual dialogue, along with the image, previously seen questions and answers (i.e. the dialogue history) are also accepted, and a relevant answer at the current time produced. We refer to this one-sided or answer-only form of visual dialogue as one-way visual dialogue (1VD). Inspired by these models and aiming to extend their capabilities, we establish the task of two-way visual dialogue (2VD) whereby an agent must be capable of acting as both the questioner and the answerer.

Our motivation for this is simple – AI agents need to be able to both ask questions *and* answer them, often interchangeably, rather do either one exclusively. For example, a vision-based home-assistant (e.g. Amazon's *Alexa*) may need to ask questions based on her visual input ("There is no toilet paper left. Would you like me to order more?") but may also need to answer questions asked by humans ("Did you order the two-ply toilet paper?"). The same question-answer capability is true for other applications. For example, with aids for the visually-impaired, a user may need the answer to "Where is the tea and kettle?", but the system may equally need to query "Are you looking for an Earl Grey or Rooibos teabag?" to resolve potential ambiguities.

We take one step toward this broad research goal with FLIPDIAL, a generative model capable of both 1VD and 2VD. The generative aspect of our model is served by using the conditional variational auto-encoder (CVAE), a framework for learning deep conditional generative models while simultaneously amortising the cost of inference in such models over the dataset [17, 24]. Furthermore, inspired by the recent success of convolutional neural networks (CNNs) in language generation and prediction tasks [11, 14, 21], we explore the use of CNNs on sequences of sequences (i.e. a dialogue) to *implicitly* capture all sequential dependences through the model. Demonstrating the surprising effectiveness of this approach, we show sets of sensible and diverse answer generations for the 1VD task in Fig. 1.

We here provide a brief treatment of works related to visual dialogue. We reserve a thorough comparison to Das et.al. [6] for §4.3, noting here that our fully-generative convolutional extension of their model outperforms their state-of-the-art results on the answering of sequential visual-based questions (1VD). In another work, Das et.al. [7] present a Reinforcement Learning based model to do 1VD, where they instantiate two separate agents, one each for questioning and answering. Crucially, the two agents are given *different* information – with one (QBot) given the caption, and the other (ABot) given the image. While this sets up the interesting task of performing image retrieval from natural-language descriptions, it is also fundamentally different from having a single agent perform both roles. Jain et.al. [12] explore a complementary task to VQA [1] where the goal is instead to generate a (diverse) set of relevant *questions* given an image. In their case, however, there is no dependence on a history of questions and answers. Finally, we note that Zhao et.al. [27] employ a similar model structure to ours, using a CVAE to model dialogue, but condition their model on discourse-based constraints for a purely linguistic (rather than visuo-linguistic) dataset. The tasks we target, our architectural differences (CNNs), and the dataset and metrics we employ are distinct.

Our primary contributions in this work are therefore:

- A fully-generative, convolutional framework for visual dialogue that outperforms state-of-the-art models on sequential question answering (1VD) using the generated answers, and establishes a baseline in the challenging two-way visual dialogue task (2VD).
- Evaluation using the *predicted* (not ground-truth) dialogue – essential for real-world conversational agents.
- Novel evaluation metrics for generative models of two-way visual dialogue to quantify answer-generation quality, question relevance, and the models's generative capacity.

2. Preliminaries

Here we present a brief treatment of the preliminaries for deep generative models – a conglomerate of deep neural networks and generative models. In particular, we discuss the variational auto-encoder (VAE) [17] which given a dataset $\mathcal{X}$ with elements $\boldsymbol{x} \in \mathcal{X}$, simultaneously learns i) a variational approximation $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$[1] to the unknown posterior distribution $p_\theta(\boldsymbol{z} \mid \boldsymbol{x})$ for latent variable $\boldsymbol{z}$, and ii) a generative model $p_\theta(\boldsymbol{x}, \boldsymbol{z})$ over data and latent variables. These are both highly attractive prospects as the ability to approximate the posterior distribution helps *amortise* inference for any given data point $\boldsymbol{x}$ over the entire dataset $\mathcal{X}$, and learning a generative model helps effectively capture the underlying abstractions in the data. Learning in this model is achieved through a unified objective, involving the marginal likelihood (or *evidence*) of the data, namely:

$$\begin{aligned}\log p_\theta(\boldsymbol{x}) &= \mathbb{D}_{\text{KL}}(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \parallel p_\theta(\boldsymbol{z} \mid \boldsymbol{x})) \\ &\quad + \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}, \boldsymbol{z}) - \log q_\phi(\boldsymbol{z} \mid \boldsymbol{x})] \\ &\geq \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - \mathbb{D}_{\text{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \| p_\theta(\boldsymbol{z}))\end{aligned} \tag{1}$$

The unknown true posterior $p_\theta(\boldsymbol{z} \mid \boldsymbol{x})$ in the first Kullback-Leibler (KL) divergence is intractable to compute making the objective difficult to optimise directly. Rather a lower-bound

[1] Following the literature, the terms recognition model or inference network may also be used to refer to the posterior variational approximation.

of the marginal log-likelihood $\log p_\theta(\boldsymbol{x})$, referred to as the evidence lower bound (ELBO), is maximised instead.

By introducing a condition variable $\boldsymbol{y}$, we capture a *conditional* posterior approximation $q_\phi(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{y})$ and a *conditional* generative model $p_\theta(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{y})$, thus deriving the CVAE [24]. Similar to Eq. (1), the conditional ELBO is:

$$\begin{aligned}\log p_\theta(\boldsymbol{x} \mid \boldsymbol{y}) \geq \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})}[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{y})] \\ - \mathbb{D}_{\text{KL}}(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{y}) \,\|\, p_\theta(\boldsymbol{z} \mid \boldsymbol{y}))\end{aligned} \tag{2}$$

where the first term is referred to as the reconstruction or negative cross entropy (CE) term, and the second, the regularisation or KL divergence term. Here too, similar to the VAE, $q_\phi(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{y})$ and $p_\theta(\boldsymbol{z} \mid \boldsymbol{y})$ are typically taken to be isotropic multivariate Gaussian distributions, whose parameters $(\boldsymbol{\mu}_q, \boldsymbol{\sigma}_q^2)$ and $(\boldsymbol{\mu}_p, \boldsymbol{\sigma}_p^2)$ are provided by deep neural networks (DNNs) with parameters ϕ and θ, respectively. The generative model likelihood $p_\theta(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{y})$, whose form varies depending on the data type – Gaussian or Laplace for images and Categorical for language models – is also parametrised similarly. In this work, we employ the CVAE model for the task of eliciting dialogue *given* contextual information from vision (images) and language (captions).

3. Generative Models for Visual Dialogue

In applying deep generative models to visual dialogue, we begin by characterising a preliminary step toward it, VQA. In VQA, the goal is to answer a single question in the context of a visual cue, typically an image. The primary goal for such a model is to ensure that the elicited answer conforms to a stronger notion of relevance than simply answering the given question – it must also relate to the visual cue provided. This notion can be extended to one-way visual dialogue (1VD) which we define as the task of answering a *sequence* of questions contextualised by an image (and a short caption describing its contents), similar to [6]. Being able to exclusively answer questions, however, is not fully encompassing of true conversational agents. We therefore extend 1VD to the more general and realistic task of two-way visual dialogue (2VD). Here the model must elicit not just answers given questions, but questions given answers as well – generating *both* components of a dialogue, contextualised by the given image and caption. Generative 1VD and 2VD models introduce stochasticity in the latent representations.

As such, we begin by characterising our generative approach to 2VD using a CVAE. For a given image $\boldsymbol{i}$ and associated caption $\boldsymbol{c}$, we define a dialogue as a sequence of question-answer pairs $\boldsymbol{d}_{1:T} = \langle(\boldsymbol{q}_t, \boldsymbol{a}_t)\rangle_{t=1}^{T}$, simply denoted $\boldsymbol{d}$ when sequence indexing is unnecessary. Additionally, we denote a dialogue context $\boldsymbol{h}$. When indexed by step as $\boldsymbol{h}_t$, it captures the dialogue subsequence $\boldsymbol{d}_{1:t}$.

With this formalisation, we characterise a generative model for 2VD under latent variable $\boldsymbol{z}$ as $p_\theta(\boldsymbol{d}, \boldsymbol{z} \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}) = p_\theta(\boldsymbol{d} \mid \boldsymbol{z}, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h})\ p_\theta(\boldsymbol{z} \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h})$, with the corresponding recognition model defined as $q_\phi(\boldsymbol{z} \mid \boldsymbol{d}, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h})$. Note that with relation to Eq. (2), data $\boldsymbol{x}$ is dialogue $\boldsymbol{d}$ and the condition variable is $\boldsymbol{y} = \{\boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}\}$, giving:

$$\begin{aligned}&\log p_\theta(\boldsymbol{d} \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}) \\ &\geq \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{d},\boldsymbol{i},\boldsymbol{c},\boldsymbol{h})}[\log p_\theta(\boldsymbol{d} \mid \boldsymbol{z}, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h})] \\ &\quad - \mathbb{D}_{\text{KL}}(q_\phi(\boldsymbol{z} \mid \boldsymbol{d}, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}) \,\|\, p_\theta(\boldsymbol{z} \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h})),\end{aligned} \tag{3}$$

with the graphical model structures shown in Fig. 2.

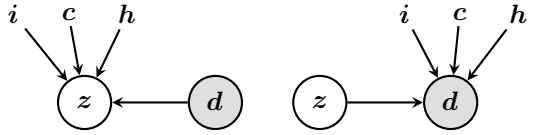

Figure 2: **Left:** Conditional recognition model and **Right:** conditional generative model for 2VD.

The formulation in Eq. (3) is general enough to be applied to single question-answering (VQA) all the way to full two-way dialogue generation (2VD). Taking a step back from generative 2VD, we can re-frame the formulation for generative 1VD (i.e. sequential answer generation) by considering the generated component to be the answer to a particular question at step t, given context from the image, caption and the sequence of previous question-answers. Simply put, this corresponds to the data $\boldsymbol{x}$ being the answer $\boldsymbol{a}_t$, conditioned on the image, its caption, the dialogue history to t-1, and the current question, or $\boldsymbol{y} = \{\boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_{t-1}, \boldsymbol{q}_t\}$. For simplicity, we denote a compound context as $\boldsymbol{h}_t^+ = \langle \boldsymbol{h}_{t-1}, \boldsymbol{q}_t \rangle$ and reformulate Eq. (3) for 1VD as:

$$\begin{aligned}&\log p_\theta(\boldsymbol{d} \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}) = \sum_{t=1}^{T} \log p_\theta(\boldsymbol{a}_t \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+), \\ &\log p_\theta(\boldsymbol{a}_t \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+) \\ &\geq \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{a}_t,\boldsymbol{i},\boldsymbol{c},\boldsymbol{h}_t^+)}[\log p_\theta(\boldsymbol{a}_t \mid \boldsymbol{z}, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+)] \\ &\quad - \mathbb{D}_{\text{KL}}(q_\phi(\boldsymbol{z} \mid \boldsymbol{a}_t, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+) \,\|\, p_\theta(\boldsymbol{z} \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+)),\end{aligned} \tag{4}$$

with the graphical model structures shown in Fig. 3.

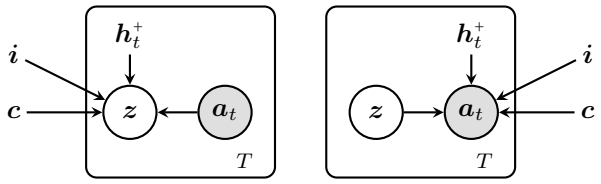

Figure 3: **Left:** Conditional recognition model and **Right:** conditional generative model for 1VD.

Our baseline [6] for the 1VD model can also be represented in our formulation by taking the variational posterior and generative prior to be conditional Dirac-Delta distributions. That is, $q_\phi(\boldsymbol{z} \mid \boldsymbol{a}_t, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+) = p_\theta(\boldsymbol{z} \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+) = \delta(\boldsymbol{z} \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+)$. This transforms the objective from Eq. (4)

by a) replacing the expectation of the log-likelihood over the recognition model by an evaluation of the log-likelihood for a *single* encoding (one that satisfies the Dirac-Delta), and b) ignoring the $\mathbb{D}_{\text{KL}}$ regulariser, which is trivially 0. This computes the marginal likelihood directly as just the model likelihood $\log p_\theta(\boldsymbol{a}_t \mid \boldsymbol{z}, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+)$, where $\boldsymbol{z} \sim \delta(\boldsymbol{z} \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+)$.

Note that while such models can "generate" answers to questions by sampling from the likelihood function, we typically don't call them generative since they effectively make the encoding of the data and conditions fully deterministic. We explore and demonstrate the benefit of a fully generative treatment of 1VD in §4.3. It also follows trivially that the basic VQA model (for single question-answering) itself can be obtained from this 1VD model by simply assuming there is no dialogue history (i.e. step length $T = 1$).

3.1. "Colouring" Visual Dialogue with Convolutions

FLIPDIAL's convolutional formulation allows us to *implicitly* capture the sequential nature of sentences and sequences of sentences. Here we introduce how we encode questions, answers, and whole dialogues with CNNs.

We begin by noting the prevalence of recurrent approaches (e.g. LSTM [10], GRU [5]) in modelling both visual dialogue and general dialogue to date [6, 7, 8, 12, 27]. Typically recurrence is employed at two levels – at the lower level to sequentially generate the words of a sentence (a question or answer in the case of dialogue), and at a higher level to sequence these sentences together into a dialogue.

Recently however, there has been considerable interest in convolutional models of language [3, 11, 14, 21], which have shown to perform at least as well as recurrent models, if not better, on a number of different tasks. They are also computationally more efficient, and typically suffer less from issues relating to exploding or vanishing gradients for which recurrent networks are known [19].

In modelling sentences with convolutions, the tokens (words) of the sentence are transformed into a stack of fixed-dimensional embeddings (e.g. using word2vec [18] or Glove [20], or those learned for a specific task). For a given sentence, say question $\boldsymbol{q}_t$, this results in an embedding $\mathring{\boldsymbol{q}}_t \in \mathbb{R}^{E \times L}$ for embedding size E and sentence length L, where L can be bounded by the maximum sentence length in the corpus, with padding tokens employed where required. This two-dimensional stack is essentially a single-channel 'image' on which convolutions can be applied in the standard manner in order to encode the entire sentence. Note this similarly applies to the answer $\boldsymbol{a}_t$ and caption $\boldsymbol{c}$, producing embedded $\mathring{\boldsymbol{a}}_t$ and $\mathring{\boldsymbol{c}}$, respectively.

We then extend this idea of viewing sentences as 'images' to whole dialogues, producing a *multi-channel* language embedding. Here, the sequence of sentences itself can be seen as a stack of (a stack of) word embeddings $\mathring{\boldsymbol{d}} \in \mathbb{R}^{E \times L \times 2T}$, where now the number of channels accounts for the number of questions and answers in the dialogue. We refer to this process as "colouring" dialogue, by analogy to the most common meaning given to image channels – colour.

Our primary motivation for adopting a convolutional approach here is to explore its efficacy in extending from simpler language tasks [11, 14] to full visual dialogue. We hence instantiate the following models for 1VD and 2VD:

Answer [1VD]: We employ the CVAE formulation from Eq. (4) and Fig. 3 to iteratively generate answers, conditioned on the image, caption and current dialogue history.

Block [1VD, 2VD]: Using the CVAE formulation from Eq. (3) and Fig. 2 we generate entire *blocks* of dialogue directly (i.e. $\boldsymbol{h} = \emptyset$ since dialogue context is implicit rather than explicit). We allow the convolutional model to *implicitly* supply the context instead. We consider this 2VD, although this block architecture can also generate iteratively, and can be evaluated on 1VD (see §4.2).

Block Auto-Regressive [1VD, 2VD]: We introduce an auto-regressive component to our generative model in the same sense as recent auto-regressive generative models for images [9, 25]. We augment the **Block** model by feeding its output through an auto-regressive (AR) module which explicitly enforces sequentiality in the generation of the dialogue blocks. This effectively factorises the likelihood in Eq. (3) as $p_\theta(\boldsymbol{d} \mid \boldsymbol{z}, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}) = p_\theta\big(\boldsymbol{d}^1 \mid \boldsymbol{z}, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}\big) \prod_{n=2}^{N} p_\theta\big(\boldsymbol{d}^n \mid \boldsymbol{d}^{1:n-1}\big)$ where N is the number of AR layers, and $\boldsymbol{d}^1$ is the (intermediate) output from the standard **Block** model. Note, again $\boldsymbol{h} = \emptyset$, and $\boldsymbol{d}^n$ refers to an entire dialogue at the n-th AR layer (rather than the t-th dialogue exchange as is denoted by $\boldsymbol{d}_t$).

4. Experiments

We present an extensive quantitative and qualitative analysis of our models' performance in both 1VD, which requires answering a sequence of image-contextualised questions, and full 2VD, where both questions *and* answers must be generated given a specific visual context. Our proposed generative models are denoted as follows:

A – **a**nswer architecture for 1VD
B – **b**lock dialogue architecture for 1VD & 2VD
$\mathbf{B_{AR}}$ – auto-regressive extension of **B** for 1VD & 2VD

A is a generative convolutional extension of our baseline [6] and is used to validate our methods against a standard benchmark in the 1VD task. **B** and $\mathbf{B_{AR}}$, like **A**, are generative, but are extensions capable of doing full dialogue generation, a much more difficult task. Importantly, **B** and $\mathbf{B_{AR}}$ are flexible in that despite being trained to generate a block of questions *and* answers ($\boldsymbol{h} = \emptyset$), they can be *evaluated* iteratively for both 1VD and 2VD (see §4.2). We summarise the data and condition variables for all models in Tab. 1. To evaluate performance on both tasks, we propose novel evaluation metrics which augment those of our baseline [6]. To the best of our knowledge, we are the first to report models

Table 1: Data ($\boldsymbol{x}$) and condition ($\boldsymbol{y}$) variables for models **A** and **B**/$\mathbf{B_{AR}}$ for 1VD and 2VD. Models **B**/$\mathbf{B_{AR}}$ can be evaluated as a block or iteratively (see §4.2), accepting ground-truth ($\boldsymbol{q}/\boldsymbol{a}$) or predicted ($\hat{\boldsymbol{q}}/\hat{\boldsymbol{a}}$) dialogue history (see Tab. 2).

Task	Model	Train		Evaluate		Eval method
		$\boldsymbol{x}$	$\boldsymbol{y}$	$\boldsymbol{x}$	$\boldsymbol{y}$	
1VD	**A**	$\boldsymbol{a}_t$	$\boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+$	$\emptyset$	$\boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+$	–
	B, $\mathbf{B_{AR}}$	$\boldsymbol{d}$	$\boldsymbol{i}, \boldsymbol{c}$	$\{\boldsymbol{d}\text{–}\boldsymbol{q}\boldsymbol{a}, \boldsymbol{d}\text{–}\boldsymbol{q}\hat{\boldsymbol{a}}\}$	$\boldsymbol{i}, \boldsymbol{c}$	iterative
2VD	**B**, $\mathbf{B_{AR}}$	$\boldsymbol{d}$	$\boldsymbol{i}, \boldsymbol{c}$	$\emptyset$ $\boldsymbol{d}\text{–}\hat{\boldsymbol{q}}\hat{\boldsymbol{a}}$	$\boldsymbol{i}, \boldsymbol{c}$	block iterative

that can generate both questions and answers given an image and caption, a necessary step toward a truly conversational agent. Our key results are:

- We set state-of-the-art results in the 1VD task on the *VisDial* dataset, improving the mean rank of the generated answers by 5.66 (Tab. 3, $\mathcal{S}_{w2v}$) compared to Das *et al*. [6].
- Our block models are able to generate both questions and answers, a more difficult but more realistic task (2VD).
- Since our models are generative, we are able to show highly diverse and plausible question and answer generations based on the provided visual context.

Datasets: We use the *VisDial* [6] dataset (v0.9) which contains Microsoft COCO images each paired with a caption and a dialogue of 10 question-answer pairs. The train/test split is $82,783/40,504$ images, respectively.

Baseline: Das *et al*. [6]'s best model, `MN-QIH-G`, is a recurrent encoder-decoder architecture which encodes the image $\boldsymbol{i}$, the current question $\boldsymbol{q}_t$ and the attention-weighted *ground truth* dialogue history $\boldsymbol{d}_{1:t-1}$. The output conditional likelihood distribution is then used to (token-wise) predict an answer. Our **A** model is a generative and convolutional extension, evaluated using existing ranking-based metrics [6] on the generated and candidate answers. We also (iteratively) evaluate our **B**/$\mathbf{B_{AR}}$ for 1VD as detailed in §4.2 (see Tab. 3).

4.1. Network architectures and training

Following the CVAE formulation (§3) and its convolutional interpretation (§3.1), all our models (**A**, **B** and $\mathbf{B_{AR}}$) have three core components: an encoder network, a prior network and a decoder network. Fig. 4 (top) shows the encoder and prior networks, and Fig. 4 (middle, bottom) show the standard and auto-regressive decoder networks.

Prior network The prior neural network, parametrised by θ, takes as input the image $\boldsymbol{i}$, the caption $\boldsymbol{c}$ and the dialogue context. Referring to Table 1, for model **A**, recall $\boldsymbol{y} = \{\boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+\}$ where the context $\boldsymbol{h}_t^+$ is the dialogue history up to t-1 and the current question $\boldsymbol{q}_t$. For models **B**/$\mathbf{B_{AR}}$, $\boldsymbol{y} = \{\boldsymbol{i}, \boldsymbol{c}\}$ (note $\boldsymbol{h} = \emptyset$). To obtain the image representation, we pass $\boldsymbol{i}$ through *VGG-16* [23] and extract the penultimate (4096-d) feature vector. We pass caption $\boldsymbol{c}$ through a pre-trained *word2vec* [18] module (we do not learn these word embeddings). If $\boldsymbol{h} \neq \emptyset$, we pass the one-hot encoding of

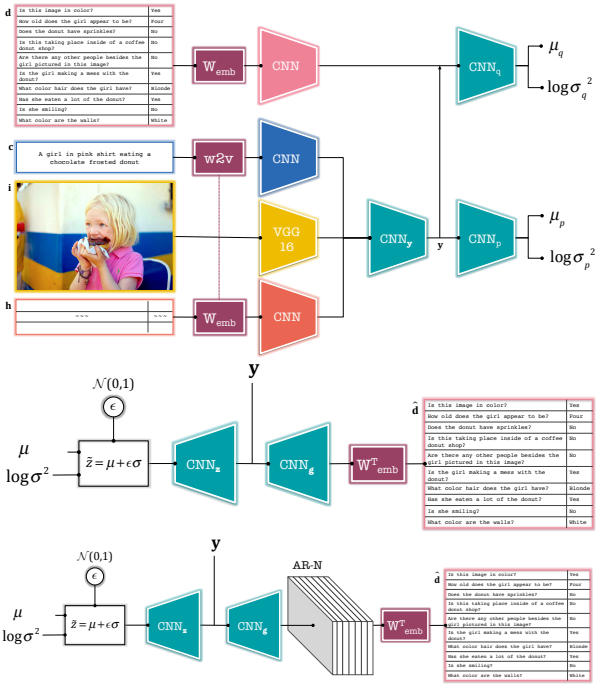

Figure 4: Convolutional **(top)** conditional encoder and prior architecture, **(middle)** conditional decoder, and **(bottom)** auto-regressive conditional decoder architectures, applying to both one- and two-way visual dialogue (1VD and 2VD).

each word through a *learnable* word embedding module and stack these embeddings as described in §3.1. We encode these condition variables convolutionally to obtain $\boldsymbol{y}$, and pass this through a convolutional block to obtain $\boldsymbol{\mu}_p$ and $\log \boldsymbol{\sigma}_p^2$, the parameters of the conditional prior $p_\theta(\boldsymbol{z} \mid \boldsymbol{y})$.

Encoder network The encoder network, parametrised by ϕ, takes $\boldsymbol{x}$ and the encoded condition $\boldsymbol{y}$ (obtained from the prior network) as input. For model **A**, $\boldsymbol{x} = \boldsymbol{a}_t$ while for **B**/$\mathbf{B_{AR}}$, $\boldsymbol{x} = \boldsymbol{d} = \langle (\boldsymbol{q}_t, \boldsymbol{a}_t) \rangle_{t=1}^{T}$. In all models, $\boldsymbol{x}$ is transformed through a word-embedding module into a single-channel answer 'image' for **A**, or a multi-channel image of alternating questions and answers for **B**/$\mathbf{B_{AR}}$. The embedded output is then combined with $\boldsymbol{y}$ to obtain $\boldsymbol{\mu}_q$ and $\log \boldsymbol{\sigma}_q^2$, the parameters of the conditional latent posterior $q_\phi(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{y})$.

Decoder network The decoder network takes as input a latent $\boldsymbol{z}$ and the encoded condition $\boldsymbol{y}$. The sample is transpose-convolved, combined with $\boldsymbol{y}$ and further transformed to obtain an intermediate output volume of dimension $E \times L \times M$, where E is the word embedding dimension, L is the maximum sentence length and M is the number of dialogue entries in $\boldsymbol{x}$ ($M = 1$ for **A**, $M = 2T$ for **B** variants). Following this, **A** and **B** employ a standard linear layer, projecting the E dimension to the vocabulary size V (Fig. 4 (middle)), whereas $\mathbf{B_{AR}}$ employs an autoregressive module followed by this standard linear layer (Fig. 4 (bottom)). At train time, the V-dimensional output is *softmax*ed and the CE term of the ELBO computed. At test time, the

Table 2: Iterative evaluation of **B/B$_{\mathbf{AR}}$** for 1VD and 2VD. Under each condition, the input dialogue block is filled with ground-truth or predicted history ($\boldsymbol{q}/\boldsymbol{a}$ or $\hat{\boldsymbol{q}}/\hat{\boldsymbol{a}}$, respectively), while future entries are filled with the PAD token.

	1VD		2VD
	$\boldsymbol{d}$–$\boldsymbol{qa}$	$\boldsymbol{d}$–$\boldsymbol{q}\hat{\boldsymbol{a}}$	$\boldsymbol{d}$–$\hat{\boldsymbol{q}}\hat{\boldsymbol{a}}$
$< t$	$(\boldsymbol{q}, \boldsymbol{a})$	$(\boldsymbol{q}, \hat{\boldsymbol{a}})$	$(\hat{\boldsymbol{q}}, \hat{\boldsymbol{a}})$
$= t$	$(\boldsymbol{q}$, PAD)	$(\boldsymbol{q}$, PAD)	(PAD, PAD) / $(\hat{\boldsymbol{q}}$, PAD)
$> t$	(PAD, PAD)	(PAD, PAD)	(PAD, PAD)

$argmax$ of the output provides the predicted word index. The weights of the encoder and prior's learnable word embedding module and the decoder's final linear layer are shared.

Autoregressive module Inspired by *PixelCNN* [26] which sequentially predicts image pixels, and similar to [9], we apply $N = \{8, 10\}$ size-preserving autoregressive layers to the intermediate output of model **B** (size $E \times L \times 2T$), and then project E to vocabulary size V. Each layer employs masked convolutions, considering only 'past' embeddings, sequentially predicting $2T * L$ embeddings of size E, enforcing sequentiality at both the sentence- and dialogue-level.

KL annealing Motivated by [4] in learning continuous latent embedding spaces for language, we employ KL annealing in the loss objectives of Eq. (3) and Eq. (4). We weight the KL term by $\alpha \in [0, 1]$ linearly interpolated over 100 epochs, and then train for a further 50 epochs ($\alpha = 1$).

Network and training hyper-parameters In embedding sentences, we pad to a maximum sequence length of $L = 64$ and use a word-embedding dimension of $E = 256$ (for *word2vec*, $E = 300$). After pre-processing and filtering the vocabulary size is $V = 9710$ (see supplement for further details). We use the Adam optimiser [16] with default parameters, a latent dimensionality of 512 and employ batch normalisation with momentum$= 0.001$ and learnable parameters. For model **A** we use a batch size of 200, and 40 for **B/B$_{\mathbf{AR}}$**. We implement our pipeline using PyTorch [22].

4.2. Evaluation methods for block models

Although **B/B$_{\mathbf{AR}}$** generate whole blocks of dialogue directly ($\boldsymbol{h} = \emptyset$), they can be evaluated iteratively, lending them to both 1VD and 2VD (see supplement for descriptions of generation/reconstruction pipelines).

- **Block evaluation [2VD]**. The generation pipeline generates whole blocks of dialogue directly, conditioned on the image and caption (i.e. $\boldsymbol{x} = \emptyset$ and $\boldsymbol{y} = \{\boldsymbol{i}, \boldsymbol{c}\}$ for **B/B$_{\mathbf{AR}}$** evaluation in Tab. 1). This is 2VD since the model must generate a coherent block of both questions *and* answers.
- **Iterative evaluation**. The reconstruction pipeline can generate dialogue items iteratively. At time t, the input dialogue block is filled with zeros (PAD token) and the ground-truth/predicted dialogue history to $< t$ is slotted in (see below and Tab. 2). This future-padded block is then encoded with the condition inputs, and then reconstructed. The t-th dialogue item is extracted (whether an answer if 1VD or a question/answer if 2VD), and this is repeated T (for 1VD) or $2T$ (for 2VD) times. Variations are:
  - $\boldsymbol{d}$–$\boldsymbol{qa}$ **[1VD]**. At time t, the input dialogue block is filled with the history of *ground-truth* questions and answers up to t-1, along with the current ground-truth question. All future entries are padded – equivalent to [6] using the ground-truth dialogue history.
  - $\boldsymbol{d}$–$\boldsymbol{q}\hat{\boldsymbol{a}}$ **[1VD]**. Similar to $\boldsymbol{d}$–$\boldsymbol{qa}$, except that the input block is filled with the history of ground-truth questions and *previously predicted* answers along with the current ground-truth question. This is a more realistic 1VD.
  - $\boldsymbol{d}$–$\hat{\boldsymbol{q}}\hat{\boldsymbol{a}}$ **[2VD]**. The most challenging and realistic condition in which the input block is filled with the history of previously predicted questions *and* answers.

Table 3: 1VD evaluation of **A** and **B/B$_{\mathbf{AR}}$** on *VisDial* (v0.9) test set. Results show ranking of answer candidates based on the score functions $\mathcal{S}_M$ and $\mathcal{S}_{w2v}$.

Score function		Method	MR	MRR	R@1	R@5	R@10
$\mathcal{S}_M$		RL-QAbot [7]	21.13	0.4370	-	53.67	60.48
		MN-QIH-G [6]	17.06	0.5259	42.29	62.85	68.88
		A (LW)	23.87	0.4220	30.48	53.78	57.52
		A (ELBO)	20.38	0.4549	34.08	56.18	61.11
$\mathcal{S}_{w2v}$		MN-QIH-G [6]	31.31	0.2215	16.01	22.42	34.76
		A (RECON)	15.36	0.4952	41.77	54.67	66.90
		A (GEN)	**25.65**	0.3227	25.88	33.43	47.75
	$\boldsymbol{d}$–$\boldsymbol{qa}$	**B**	28.45	0.2927	23.50	29.11	42.29
		B$_{\mathbf{AR}}$8	25.87	**0.3553**	**29.40**	**36.79**	**51.19**
		B$_{\mathbf{AR}}$10	26.30	0.3422	28.00	35.34	50.54
	$\boldsymbol{d}$–$\boldsymbol{q}\hat{\boldsymbol{a}}$	**B**	30.57	0.2188	16.06	20.88	35.37
		B$_{\mathbf{AR}}$8	29.10	0.2864	22.52	29.01	48.43
		B$_{\mathbf{AR}}$10	29.15	0.2869	22.68	28.97	46.98

4.3. Evaluation and Analysis

We evaluate our **A**, **B**, and **B$_{\mathbf{AR}}$** models on the 1VD and 2VD tasks. Under 1VD, we predict an answer with each time step, given an image, caption and the current dialogue history (§4.3.1 and Tab. 3), while under 2VD, we predict both questions *and* answers (§4.3.2 and Tab. 4). All three models are able to perform the first task , while only **B** and **B$_{\mathbf{AR}}$** are capable of the second task.

4.3.1 One-Way Visual Dialogue (1VD) task

We evaluate the performance of **A** and **B/B$_{\mathbf{AR}}$** on 1VD using the candidate ranking metric of [6] as well as an extension of this which assesses the *generated* answer quality (Tab. 3). Fig. 1 and Fig. 5 show our qualitative results for 1VD.

Candidate ranking by model log-likelihood [$\mathcal{S}_{\mathrm{M}}$] The *VisDial* dataset [6] provides a set of 100 candidate answers $\{\boldsymbol{a}_t^c\}_{c=1}^{100}$ for each question-answer pair at time t per image. The set includes the ground-truth answer $\boldsymbol{a}_t$ as well as similar, popular, and random answers. Das *et al*. [6] rank these candidates using the log-likelihood value of each under

Question	Ground-truth answer	z_1	z_2	z_3
How old is the girl?	Maybe three	Looks about six	I can't tell	Yes
What race is the girl?	White	Yes	White	Caucasian
Is she outside?	Yes	No	Yes	Yes
Is her hair long or short?	Short	Short	Short	Short
What color is her hair?	Blonde	Blonde	Brown	Brown
Is her hair curly or straight?	It's straight	Straight	Straight	Straight
What is she wearing?	Pink shirt, white pants	T-shirt and jacket and pants	Shirt and pants	Jeans like a pajamas
Is the teddy bear in her lap?	Yes	No	Yes	Yes
What color is the teddy bear?	White	Brown	Tan	Tan and white
Is it nice outside?	Yes sunny	Yes	It looks lovely	Yes

Question	Ground-truth answer	z_1	z_2	z_3
How old does she look?	Around seven or eight	I cannot tell about her	Looks about six	She is about teenager around
Any other people?	No	No	Yes	Yes
Any buildings?	No	Yes	No	No
Is it day or night?	Day	It looks like it in image is in so	Daytime	Day
Is it raining?	No	Yes	Yes	No
What color umbrella?	Pink and clear	Dark colored color	White	White
Is it open or closed?	Open	Yes	Open	Open
Is it sunny?	I can't tell	Yes	No	Yes
What color is her hair?	Dark brown	Brown	Brown	Brown
Is it long or short?	Long	Short	Short	I'd say long

Figure 5: Example generated answers from **A**'s conditional prior – conditioned on an image, caption, question and dialogue history. See supplement for further examples.

their model (conditioned on the image, caption and dialogue history, including the current question), and then observe the position of the ground-truth answer (closer to 1 is better). This position is averaged over the dataset to obtain the Mean Rank (MR). In addition, the Mean Reciprocal Rank (MRR; 1/MR) and recall rates at $k = \{1, 5, 10\}$ are computed.

To compare against their baseline, we rank the 100 candidates answers by estimates of their *marginal* likelihood from **A**. This can be done with i) the conditional ELBO (Eq. (4)), and by ii) likelihood weighting (LW) in the conditional generative model $p_\theta(\boldsymbol{a}_t \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+) = \int p_\theta(\boldsymbol{a}_t, \boldsymbol{z} \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+) dz = \int p_\theta(\boldsymbol{z} \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+) p_\theta(\boldsymbol{a} \mid \boldsymbol{z}, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+)\, dz$. Ranking by both these approaches is shown in the $\mathcal{S}_{\text{M}}$ section of Tab. 3, indicating that we are comparable to the state of the art in discriminative models of sequential VQA [6, 7].

Candidate ranking by *word2vec* cosine distance [$\mathcal{S}_{w2v}$] The evaluation protocol of [6] scores and ranks a given set of candidate answers, without being a function of the actual answer *predicted* by the model, $\hat{\boldsymbol{a}}_t$. This results in the rank of the ground-truth answer candidate reflecting its score under the model *relative* to the rest of the candidates' scores, rather than capturing the quality of the answer output by the model, which is left unobserved. To remedy this, we instead score each candidate by the cosine distance between the *word2vec* embedding of the predicted answer $\hat{\boldsymbol{a}}_t$ and that candidate's *word2vec* embedding. We take the embedding of a sentence to be the average embedding over word tokens following Arora *et al*. [2]. In addition to accounting for the predicted answer, this method also allows semantic similarities to be captured such that if the predicted answer is similar (in meaning and/or words generated) to the ground-truth candidate answer, then the cosine distance will be small, and hence the ground-truth candidate's rank closer to 1.

We report these numbers for **A**, iteratively-evaluated **B/B$_{\text{AR}}$**, and also our baseline model MN-QIH-G [6], which we re-evaluate using the *word2vec* cosine distance ranking (see $\mathcal{S}_{w2v}$ in Tab. 3). In the case of **A** (GEN), we evaluate answer *generations* from **A** whereby we condition on $\boldsymbol{i}, \boldsymbol{c}$ and $\boldsymbol{h}_t^+$ via the prior network, sample $\boldsymbol{z} \sim \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_p, \boldsymbol{\sigma}_p^2)$ and generate an answer via the decoder network. Here we show an improvement of 5.66 points in MR over the baseline. On the other hand, **A** (RECON) evaluates answer *reconstructions* in which $\boldsymbol{z}$ is sampled from $\mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_q, \boldsymbol{\sigma}_q^2)$ (where ground-truth answer $\boldsymbol{a}_t$ is provided). We include **A** (RECON) merely as an "oracle" autoencoder, observing its good ranking performance, but do not explicitly compare against it.

We also note that the ranking scores of the block models are worse (by 3-4 MR points) than those of **A**. This is expected since **A** is explicitly trained for 1VD which is not the case for **B/B$_{\text{AR}}$**. Despite this, the performance gap between **A** (GEN) and **B/B$_{\text{AR}}$** (with $\boldsymbol{d}$–$\boldsymbol{qa}$) is not large, bolstering our iterative evaluation method for the block architectures. Note finally that the **B/B$_{\text{AR}}$** models perform better under $\boldsymbol{d}$–$\boldsymbol{qa}$ than under $\boldsymbol{d}$–$\boldsymbol{q}\hat{\boldsymbol{a}}$ (by 2-3 MR points). This is also expected as answering is easier with access to the ground-truth dialogue history rather than when only the previously *predicted* answers (and ground-truth questions) are provided.

4.3.2 Two-way Visual Dialogue (2VD) task

Our flexible CVAE formulation for visual dialogue allows us to move from 1VD to the generation of both questions *and* answers (2VD). Despite this being inherently more challenging, **B/B$_{\text{AR}}$** are able to generate diverse sets of questions and answers contextualised by the given image and caption. Fig. 6 shows snippets of our two-way dialogue generations.

In evaluating our models for 2VD, the candidate ranking protocol of [6] which relies on a *given* question to rank the answer candidates, is no longer usable when the questions themselves are being generated. This is the case for **B/B$_{\text{AR}}$** block evaluation, which has no access to the ground-truth dialogue history, and the $\boldsymbol{d}$–$\hat{\boldsymbol{q}}\hat{\boldsymbol{a}}$ iterative evaluation, when the full predicted history of questions and answers is provided (Tab. 2). We therefore look directly to the CE and KL terms of the ELBO as well as propose two new metrics, $sim_{\boldsymbol{c},\boldsymbol{q}}$ and $sim_{\circlearrowleft}$, to compare our methods in the 2VD task:

- **Question relevance** ($sim_{\boldsymbol{c},\boldsymbol{q}}$). We expect a generated question to query an aspect of the image, and we use the presence of semantically similar words in both the question and image caption as a proxy of this. We compute the cosine distance between the (average) *word2vec* embedding of each predicted question $\boldsymbol{q}_t$ and that of the caption

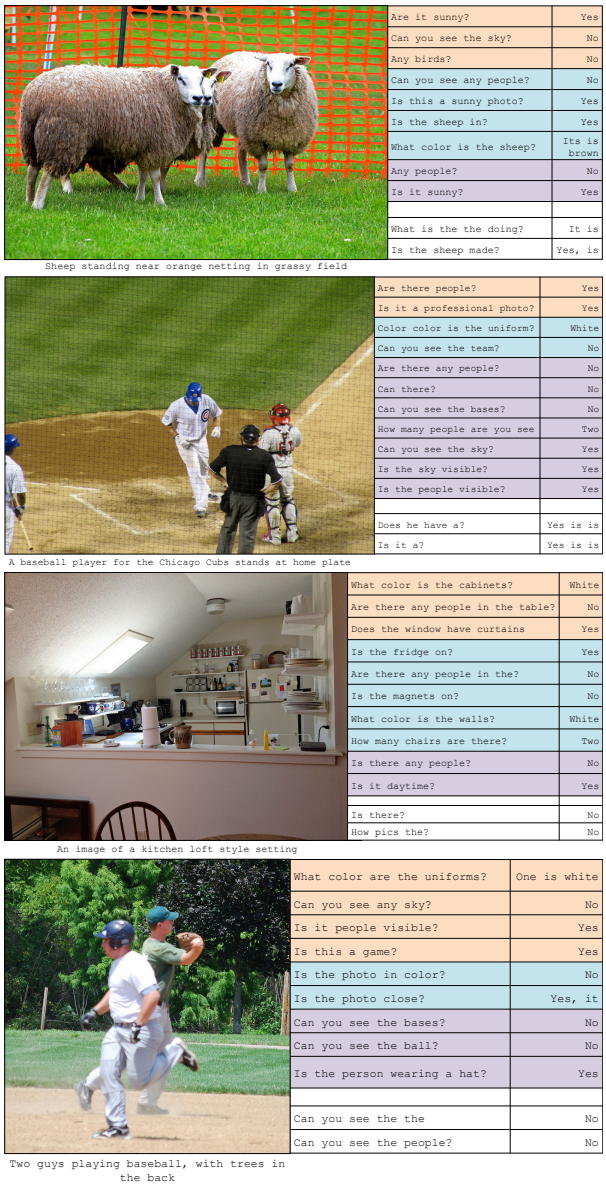

Figure 6: Examples of two-way dialogue generation from the **B/B$_{\text{AR}}$** models. Different colours indicate different generations – coherent sets with a single colour, and failures in white. See supplement for further examples.

$\boldsymbol{c}$, and average over all T questions in the dialogue (closer to 1 indicates higher semantic similarity).

- **Latent dialogue dispersion** ($sim_{\circlearrowleft}$). For a generated dialogue block $\boldsymbol{d}^g$, $sim_{\circlearrowleft}$ computes the KL divergence $\mathbb{D}_{\text{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{d}^g, \boldsymbol{i}, \boldsymbol{c}) \,\|\, q_\phi(\boldsymbol{z}|\boldsymbol{d}, \boldsymbol{i}, \boldsymbol{c}))$, measuring how close the generated dialogue is to the true dialogue $\boldsymbol{d}$ in the latent space, given the same image $\boldsymbol{i}$ and caption $\boldsymbol{c}$.

From Tab. 4, we observe a decrease in the loss terms as the auto-regressive capacity of the model increases (none $\rightarrow$ 8 $\rightarrow$ 10), suggesting that explicitly enforcing sequentiality in the dialogue generations is useful. For sim$_{\circlearrowleft}$ within a particular model, the dispersion values are typically larger for the harder task (without dialogue context). We also observe that dispersion increases with number of AR layers, suggesting AR improves the diversity of the model outputs, and avoids simply recovering data observed at train time.

Table 4: 2VD evaluation on *VisDial* (v0.9) test set for **B/B$_{\text{AR}}$** models. For $\boldsymbol{d}$, '$\emptyset$' indicates block evaluation, and '$\boldsymbol{d}$–$\hat{\boldsymbol{q}}\hat{\boldsymbol{a}}$' indicates iterative evaluation (see §4.2).

Method	$\boldsymbol{d}$	CE	KLD	sim$_{\boldsymbol{c},q}$	sim$_{\circlearrowleft}$
B	$\emptyset$	31.18	4.34	0.4931	14.20
	$\boldsymbol{d}$–$\hat{\boldsymbol{q}}\hat{\boldsymbol{a}}$	25.40	4.01	0.4091	1.86
B$_{\text{AR}}$8	$\emptyset$	28.81	2.54	0.4878	31.50
	$\boldsymbol{d}$–$\hat{\boldsymbol{q}}\hat{\boldsymbol{a}}$	26.60	2.29	0.3884	2.39
B$_{\text{AR}}$10	$\emptyset$	28.49	1.89	0.4927	44.34
	$\boldsymbol{d}$–$\hat{\boldsymbol{q}}\hat{\boldsymbol{a}}$	24.93	1.80	0.4101	2.35

While the proposed metrics provide a novel means to evaluate dialogue in a generative framework, like all language-based metrics, they are not complete. The question-relevance metric, sim$_{\boldsymbol{c},\boldsymbol{q}}$, can stagnate, and neither metric precludes redundant or nonsensical questions. We intend for these metrics to *augment* the bank of metrics available to evaluate dialogue and language models. Further evaluation, including i) using auxiliary tasks, as in the image-retrieval task of [7], to drive and evaluate the dialogues, and ii) turning to human evaluators to rate the generated dialogues, can be instructive in painting a more complete picture of our models.

5. Conclusion

In this work we propose FLIPDIAL, a generative convolutional model for visual dialogue which is able to generate answers (1VD) as well as generate both questions *and* answers (2VD) based on a visual context. In the 1VD task, we set new state-of-the-art results with the answers generated by our model, and in the 2VD task, we are the first to establish a baseline, proposing two novel metrics to assess the quality of the generated dialogues. In addition, we propose and evaluate our models under a much more realistic setting for both visual dialogue tasks in which the *predicted* rather than ground-truth dialogue history is provided at test time. This challenging setting is more akin to real-world situations in which dialogue agents must be able to evolve with their predicted exchanges. We emphasize that research focus must be directed here in the future. Finally, under all cases, the sets of questions and answers generated by our models are qualitatively good: diverse and plausible given the visual context. Looking forward, we are interested in exploring additional methods for enforcing diversity in the generated questions and answers, as well as extending this work to explore *recursive* models of reasoning for visual dialogue.

Acknowledgements This work was supported by the EPSRC, ERC grant ERC-2012-AdG 321162-HELIOS, EPSRC grant Seebibyte EP/M013774/1, EPSRC/MURI grant EP/N019474/1 and the Skye Foundation.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual Question Answering. In *ICCV*, 2015. 2

[2] S. Arora, Y. Liang, and T. Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017. 7

[3] S. Bai, J. Kolter, and V. Koltun. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling . *CoRR*, abs/1803.01271, 2018. 4

[4] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015. 6

[5] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 4

[6] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra. Visual Dialog. In *CVPR*, 2017. 2, 3, 4, 5, 6, 7

[7] A. Das, S. Kottur, J. M. Moura, S. Lee, and D. Batra. Learning cooperative visual dialog agents with deep reinforcement learning. *arXiv preprint arXiv:1703.06585*, 2017. 2, 4, 6, 7, 8

[8] A. Das, S. Kottur, J. M. Moura, S. Lee, and D. Batra. Learning cooperative visual dialog agents with deep reinforcement learning. In *ICCV*, 2017. 4

[9] I. Gulrajani, K. Kumar, F. Ahmed, A. A. Taiga, F. Visin, D. Vazquez, and A. Courville. Pixelvae: A latent variable model for natural images. *arXiv preprint arXiv:1611.05013*, 2016. 4, 6

[10] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, Nov. 1997. 4

[11] B. Hu, Z. Lu, H. Li, and Q. Chen. Convolutional neural network architectures for matching natural language sentences. In *NIPS*, pages 2042–2050, 2014. 2, 4

[12] U. Jain, Z. Zhang, and A. Schwing. Creativity: Generating diverse questions using variational autoencoders. *arXiv preprint arXiv:1704.03493*, 2017. 2, 4

[13] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[14] N. Kalchbrenner, E. Grefenstette, P. Blunsom, D. Kartsaklis, N. Kalchbrenner, M. Sadrzadeh, N. Kalchbrenner, P. Blunsom, N. Kalchbrenner, and P. Blunsom. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 212–217. Association for Computational Linguistics, 2014. 2, 4

[15] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015. 2

[16] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2014. 6

[17] D. P. Kingma and M. Welling. Auto-encoding Variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, 2014. 2

[18] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119, 2013. 4, 5

[19] R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning*, pages 1310–1318, 2013. 4

[20] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014. 4

[21] N.-Q. Pham, G. Kruszewski, and G. Boleda. Convolutional neural network language models. In *EMNLP*, pages 1153–1162, 2016. 2, 4

[22] PyTorch, 2017. 6

[23] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR*, abs/1409.1556, 2014. 5

[24] K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models. In *NIPS*, pages 3483–3491, 2015. 2, 3

[25] A. van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves, et al. Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems*, pages 4790–4798, 2016. 4

[26] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu. Pixel recurrent neural networks. In *NIPS*, 2016. 6

[27] T. Zhao, R. Zhao, and M. Eskenazi. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. *arXiv preprint arXiv:1703.10960*, 2017. 2, 4