VARIATIONAL DISENTANGLED ATTENTION FOR REGULARIZED VISUAL DIALOG

Anonymous authors
Paper under double-blind review

ABSTRACT

One of the most important challenges in a visual dialog is to effectively extract the information from a given image and its historical conversation which are related to the current question. Many studies adopt the soft attention mechanism in different information sources due to its simplicity and ease of optimization. However, some of visual dialogs are observed in a single round. This implies that there is no substantial correlation between individual rounds of questions and answers. This paper presents a unified approach to disentangled attention to deal with context-free visual dialogs. The question is disentangled in latent representation. In particular, an informative regularization is imposed to strengthen the dependence between vision and language by pretraining on the visual question answering before transferring to visual dialog. Importantly, a novel variational attention mechanism is developed and implemented by a local reparameterization trick which carries out a discrete attention to identify the relevant conversations in a visual dialog. A set of experiments are evaluated to illustrate the merits of the proposed attention and regularization schemes for context-free visual dialogs.

1 INTRODUCTION

With the advances in deep learning and the abundant human conversations on social media, the conversational agent systems have drawn increasing attention from the research community of natural language processing for artificial intelligence (AI). Visual dialog (Das et al., 2017) is a conversational question answering task that has recently received lots of research efforts. Given an image, a conversation history consisting of a series of question-and-answer pairs, and a follow-up question about the image, the system has to predict the natural language answer to the question. But, in context-free dialogs, parts of the visual dialog are not consistent and continuous as shown in Figure 1). The challenges of this task are basically due to the situation that the learner needs to involve the modeling, understanding, and participating in information seeking conversations. In order to learn the relevance of conversations in visual dialog, many studies (Kang et al., 2019; Murahari et al., 2020; Agarwal et al., 2020; Park et al., 2021) have focused on designing the delicate attention modules and learning representations through the stacking of attention modules. Most of these studies adopt a deterministic soft attention mechanism that calculates the outputs from all source components no matter they are relevant or irrelevant to questions. Context-free questions make a deterministic soft attention network to receive too much irrelevant information. The resulting attention network likely hurts the robustness and performance of the model.

This paper explores a unified approach to handle the aforementioned challenges with two stages. First, in order to respond effectively to the context-free questions, certain parts of the model are pretrained from the related task on visual question answering (VQA) (Antol et al., 2015). The model turns the visual dialog task into a VQA task based on a single round question answering (QA) so as to cope with the context-free questions. However, recent researches (Agrawal et al., 2018; Cadene et al., 2019) show that most of VQA models suffer from the *language prior* problem. This problem results in the circumstances that the model will fail to ground questions in image content and perform poorly in real-world settings. To alleviate the issue of language priors, this study introduces the information-theoretic latent disentanglement which learns the decomposed linguistic representation from questions. In addition, an informative regularization scheme is developed for latent disentanglement by imposing the constraint in learning representation. This paper further proposes a novel discrete variational attention mechanism that can be combined with pre-trained model to handle the

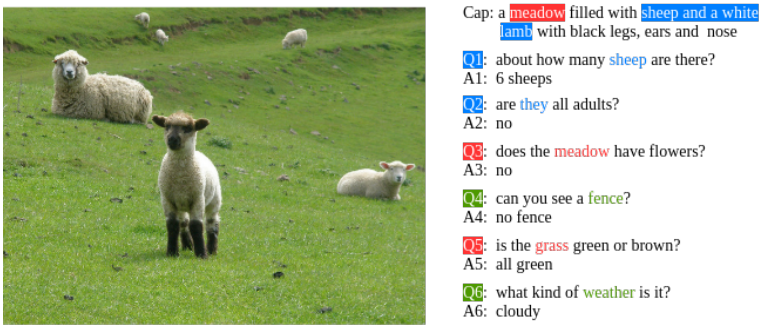


Figure 1: An example of visual dialog task. Red and blue highlights show two context-relevant questions, and green highlights denote the context-free questions. This example demonstrates that the questioner likely dives into a different topic in a conversation.

context-free issues in visual dialog. Inspired by the discrete stochastic neural networks (Shayer et al., 2018; Peters & Welling, 2018), this work presents a new variational attentive distribution based on Bernoulli distribution which is implemented by using a local reparameterization trick to carry out a differentiable training process. The architecture is constructed by Bayesian framework which is optimized by using a learned prior for regularization.

2 Background Survey

2.1 Visual question answering and visual dialog

Visual question answering (VQA) is receiving the popularity as an AI-complete challenge where the task is to autonomously answer natural language questions based on the visuals. VQA is different from visual dialog for the reasons that VQA does not need to consider additional historical conversations, and VQA only needs to answer a series of independent questions based on image contents. However, many VQA models suffer from the language priors problem, which is caused by superficial correlations between a question and some answer words learned during training. It is because that the mapping or relation between question and answer is too strong so that the model overemphasizes this mapping and ignores the high-level relations between different interactions. Correspondingly, the true information in an image is not really captured. The so-called language prior problem happens. To deal with this issue, a number of works have been proposed. Agrawal et al. (2018) provided a methodology that explicitly separated the visual recognition from an answer prediction for various question types. Ramakrishnan et al. (2018) proposed an adversarial learning strategy to enhance the dependence between question and image by degrading the performance of question-only branch. Cadene et al. (2019) dynamically modified the weights of training instance based on the prior masks learned by the question-only branch so that the influence of the large biased cases is decreased while increasing the influence of the small biased instances. In visual dialog, there have been many recent studies focused on using the basis of visual grounding tasks to impair the performance of the model. Murahari et al. (2020) strengthen the dependency of visual and language through pretraining image caption and VQA. Cogswell et al. (2020) propose a new framework to transfer to the dialog task through the trained VQA model.

2.2 Learning for disentanglement and attention

Disentangled learning can be divided into two aspects according to different situations, one is the feature disentanglement, the other is the cross-domain disentanglement. This paper focuses on the feature disentanglement, which extracts the shared information in the database and then decomposes the attribute representations of different style meanings. This disentanglement learns the decomposition through the differences between local features, which is often used in attribute extraction and attribute transfer (Mathieu et al., 2019; Chen et al., 2019; Huang & Chien, 2021). To effectively learn the disentangled representations, mutual information (MI) has been introduced to carry out different approaches to achieve desirable performance (Alemi et al., 2016; Chen et al., 2016; Cheng

et al., 2020b; Hwang et al., 2020). However, calculating the exact MI value is difficult because the integral in the calculation $I(\mathbf{x};\mathbf{z}) \equiv \mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[\log \frac{p(\mathbf{x},\mathbf{z})}{p(\mathbf{x})p(\mathbf{z})}\right]$ is intractable in most cases of deep learning using neural network models. Various MI estimation approaches have been proposed to address this issue. Agakov (2004) proposed the classic MI upper and lower bounds, which opened an avenue to estimate MI through different approximate distributions. Cheng et al. (2020a) introduced the approximated network and estimated the relation between the joint probability distribution and the respective marginal probability distribution. Some approaches introduced the critic function or the similarity calculations to estimate the MI instead of inferring the probability density function. Donsker & Varadhan (1975) derived the MI lower bound based on introducing energy-based function. To overcome the intractable formulation, Belghazi et al. (2018) adopted the Monte-Carlo approximation of the expectations, so that this method could estimate the gradient and find the MI estimator. On the other hands, Hjelm et al. (2018) introduced the adversarial learning, which applied the discriminative network to maximize the objective directly instead of calculating its exact value.

In addition to the disentangled learning, the attention-based methods play an important role to spotlight the regions in images or words in sentences for a target task such as visual dialog. Attentive representation learning was first introduced in machine translation task (Bahdanau et al., 2015), and has been developed for many other tasks. Attention method was introduced to improve the recurrent sequence-to-sequence (seq2seq) model. To identify the specialized local features from source sequences, attention mechanism is based on finding the similarity between source and target sequences to know which part of source sequence the decoder would like to emphasize at each time step. Furthermore, in order to reflect the complicated dependencies and improve the robustness of attention mechanism, many studies combined the attention weights with Bayesian framework. Bahuleyan et al. (2018) proposed the variational attention to improve the diversity of attention in seq2seq models by sampling the attention weights from Gaussian distribution. But, in some works, the property of probability in the estimated distribution of attention weights was not really satisfied. Deng et al. (2018) proposed the stochastic version of hard attention mechanism, where the attention weights were seen as the discrete variables which were sampled from categorical distributions. Backpropagation was no longer suitable for optimization because the parameters of network were non-differentiable. Fan et al. (2020) proposed a differentiable way to form the attention distribution by normalizing the reparameterizable distributions.

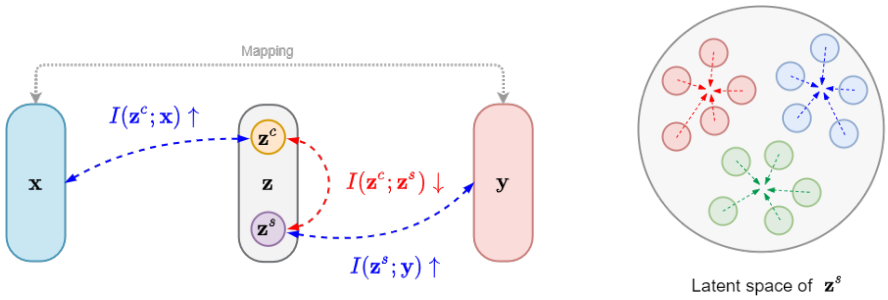
3 VARIATIONAL DISENTANGLED ATTENTION

This work presents a unified way to visual dialog which is decomposed into two stages. The first one is to build a pre-trained model. In order to strengthen the dependence between the image and the question, in the pre-training stage, the issue of language priors is handled through the disentanglement with regularization. Secondly, in order to effectively provide the context-relevant information to the pre-trained model, a discrete variant of variational attention mechanism with local reparameterization trick is developed. The differentiable training process is exploited for ease of implementation.

3.1 INFORMATIVE LATENT DISENTANGLEMENT

This paper addresses an VQA model which is seen as a pre-trained task before transferring it to visual dialog. This model makes use of a large-scale VQA dataset for training the powerful visually-grounded representations by disentangling and regularizing the linguistic representation to alleviate the issue of language priors. Specifically, a disentangled representation should separate the distinct and informative factors of variations behind the data (Desjardins et al., 2012). This study incorporate the disentangled learning in VQA task. The attribute of questions is observed and decomposed into two different meanings or representations through information-theoretic learning. An informative regularization is developed for latent disentanglement with some constraint in learning procedure.

In general, the strategies of disentangled learning is to identify the data attributes and specify the learning goals. The language prior problem, caused by superficial correlations between questions and answers learned during training, can be tackled by strengthening the mapping between question and answer. In order to reflect the attributes in the question, two distinct types including the *content* representation $\mathbf{z}^c$ and the *style* representation $\mathbf{z}^s$ are considered. Specifically, the question $\mathbf{x}$ is



(a) Latent decomposition using mutual information

(b) Clustering phenomenon

Figure 2: (a) Illustration for decomposition of a question $\mathbf{x}$ based on maximization and minimization of mutual information and (b) Illustration for clustering phenomenon in latent space of style representation where red, blue, and green clusters denote different types of questions, and dashed arrows indicate the direction of regularization. In an ideal learning process, if the same type of questions is clustered in the latent space, it may contain unknown disturbing factors. An informative regularization is helpful to alleviate such a disturbance.

transformed and factorized into two different representations (i.e. $\{\mathbf{z}^c, \mathbf{z}^s\} = f_\theta(\mathbf{x})$) through a disentanglement network f_θ which is trained in accordance with information theory. This paper basically constructs the following goals to carry out learning strategy for disentanglement.

- Decompose the content and the style information from the given question. Specifically, content representation should extract the basic significance from questions and style representation show different attributes of questions. Ideally, we expect that style representation can learn the main factor that cause language priors in questions.
- In order to make the relationship between content and style more independent and meaningful, we aim to minimize the mutual information between them.
- Ensure the content representation sufficiently contains information from the given question.
- Learn style representation to extract different attributes that cause the language priors in questions. This process demonstrates the extraction of attributes hidden in questions in the mapping between question and answer.

According to these goals for latent disentanglement, the learning criteria can be formulated in the form of using mutual information terms as follows

$$\mathcal{J}_{\text{Dis}} = I(\mathbf{z}^c;\mathbf{z}^s) - I(\mathbf{z}^c;\mathbf{x}) - I(\mathbf{z}^s;\mathbf{y}) \tag{1}$$

where $\mathbf{y}$ denotes the ground-truth answer corresponding to question $\mathbf{x}$. Figure 2a illustrates the interaction between different components, the parameters of disentanglement network are learned by minimizing the objective $\mathcal{J}_{\text{Dis}}$. However, original mutual information is intractable. An alternative calculation based on neural networks is required. To handle the issue, this paper introduces the MI estimation (Agakov, 2004; Cheng et al., 2020a;b) to implement the tractable criteria instead of using intractable MI. To maximize $I(\mathbf{z}^c;\mathbf{x})$ and $I(\mathbf{z}^s;\mathbf{y})$, this paper introduces their MI lower bounds as the alternative objectives. Specifically, MI estimators are calculated by using feedforward neural network with parameters ϕ_c and ϕ_s to find lower bounds by

$$I(\mathbf{z}^c;\mathbf{x}) \geq H(\mathbf{x}) + \mathbb{E}_{p(\mathbf{x},\mathbf{z}^c)}[\log p_{\phi_c}(\mathbf{x}|\mathbf{z}^c)] \tag{2}$$

$$I(\mathbf{z}^s;\mathbf{y}) \geq H(\mathbf{y}) + \mathbb{E}_{p(\mathbf{y},\mathbf{z}^s)}[\log p_{\phi_s}(\mathbf{y}|\mathbf{z}^s)]. \tag{3}$$

Note that both $H(\mathbf{x})$ and $H(\mathbf{y})$ are entropy terms which are constants with respect to ϕ_c and ϕ_s, which are ignored during optimization. On the other hand, in order to minimize the intractable MI $I(\mathbf{z}^c;\mathbf{z}^s)$, this paper introduces an approximate network q_σ to approximate conditional distribution, so that the upper bound of $I(\mathbf{z}^c;\mathbf{z}^s)$ can be obtained. Note that the parameter σ is fixed when estimating the MI upper bound by

$$I(\mathbf{z}^c;\mathbf{z}^s) \leq \mathbb{E}_{p(\mathbf{z}^c,\mathbf{z}^s)}[\log q_\sigma(\mathbf{z}^s|\mathbf{z}^c)] - \mathbb{E}_{p(\mathbf{z}^c)p(\mathbf{z}^s)}[\log q_\sigma(\mathbf{z}^s|\mathbf{z}^c)]. \tag{4}$$

In training process, the approximate network is trained by minimizing the negative log-likelihood $\mathbb{E}_{p(\mathbf{z}^c,\mathbf{z}^s)}[-\log q_\sigma(\mathbf{z}^s|\mathbf{z}^c)]$. With the MI estimators, the objective function of disentanglement is integrated into $\mathcal{I}_{\text{Dis}}$, which is the upper bound of $\mathcal{J}_{\text{Dis}}$ so that it can be optimized by the the following objective by using MI estimators

$$\begin{aligned}\mathcal{I}_{\text{Dis}} = &\underbrace{\mathbb{E}_{p(\mathbf{z}^c,\mathbf{z}^s)}[\log q_\sigma(\mathbf{z}^s|\mathbf{z}^c)] - \mathbb{E}_{p(\mathbf{z}^c)p(\mathbf{z}^s)}[\log q_\sigma(\mathbf{z}^s|\mathbf{z}^c)]}_{\text{upper bound of } I(\mathbf{z}^c;\mathbf{z}^z)} \\ &- \underbrace{\mathbb{E}_{p(\mathbf{x},\mathbf{z}^c)}[\log p_{\phi_c}(\mathbf{x}|\mathbf{z}^c)]}_{\text{lower bound of } I(\mathbf{z}^c;\mathbf{x})} - \underbrace{\mathbb{E}_{p(\mathbf{y},\mathbf{z}^s)}[\log p_{\phi_s}(\mathbf{y}|\mathbf{z}^s)]}_{\text{lower bound of } I(\mathbf{z}^s;\mathbf{y})} \geq \mathcal{J}_{\text{Dis}}\end{aligned} \tag{5}$$

3.2 Informative model regularization

After decomposing the attributes from question, the style representations $\mathbf{z}^s$ have an unpredictable phenomenon in the visualized latent space. The unknown disturbing factors make the style representations clustered in the latent space and also make them sensitive in representing the attributes of question. Based on the concept of learning criteria, the clustering phenomenon in the style representations[1]is seen as the main factor which causes the language priors. Such a representation cannot effectively express the general meaning of the question, but can only reflect the mapping or relation between each question and the corresponding answer. To alleviate this phenomenon, this paper introduces another regularization term corresponding to the entropy of model predicted answer and style predicted answer. The entropy represents the *perturbation* of a probability distribution, and the learning of representation can be constrained by controlling the entropy (Figure 2b). Therefore, there are two goals in the regularization term.

- Ensure the model distribution $p(\mathbf{y}|f_\theta(\mathbf{x}),\mathbf{x}^v)$ which considers that image $\mathbf{x}^v$ and question $\mathbf{x}$ are sharper than the style distribution $p_{\phi_s}(\mathbf{y}|f_\theta(\mathbf{x}))$ which doesn't consider image contents.
- Make the style predicted distribution smoother so that it can generally represents the style attributes of question.

The regularization is derived with two terms where the first term is the model predicted entropy, and the second term is the style predicted entropy which constrains the learning of style representation

$$\begin{aligned}\mathcal{R}_{\text{H}} &= H(\mathbf{y}|f_\theta(\mathbf{x}),\mathbf{x}^v) - \gamma H(\mathbf{y}|f_\theta(\mathbf{x})) \\ &= \mathbb{E}_{p(\mathbf{x},\mathbf{x}^v,\mathbf{y})}[-\log p(\mathbf{y}|f_\theta(\mathbf{x}),\mathbf{x}^v)] - \gamma\mathbb{E}_{p(\mathbf{x},\mathbf{y})}[-\log p_{\phi_s}(\mathbf{y}|f_\theta(\mathbf{x}))]\end{aligned} \tag{6}$$

where the hyperparameter $\gamma \geq 0$ controls the balance of how the features of language priors expressed in style representations. For lower γ, slight constraint occurs and the style representation continues to learn the mapping sensitively. On the contrary, larger γ forces the style representation to generally learn the attributes of question. It is then possible to restrict the representation capability.

To optimize the pre-trained task, minimizing the negative log-likelihood loss is introduced. For minimizing the conditional negative log-likelihood $-\log p(\mathbf{y}|f_\theta(\mathbf{x}),\mathbf{x}^v)$ of VQA task, we introduce the disentanglement network f_θ. The learning objective is formulated as a context-irrelevant loss

$$\mathcal{L}_{\text{I}}(\mathbf{x},\mathbf{x}^v,\mathbf{y}) = -\log p(\mathbf{y}|f_\theta(\mathbf{x}),\mathbf{x}^v) + \beta\mathcal{I}_{\text{Dis}} \tag{7}$$

where first term is the conditional negative log-likelihood, second one is the MI upper bound and β in loss function is a formal expression of reweighting the two objectives of disentanglement and answer prediction. Furthermore, in order to alleviate the impact of language priors on the task, this paper adopts the regularization $\mathcal{R}_{\text{H}}$ to constrain the disentanglement network. The overall pre-trained VQA task is optimized by minimizing the regularized objective $\mathcal{L}_{\text{I}} + \mathcal{R}_{\text{H}}$.

3.3 Variational attention with local reparameterization

In this section, we expand VQA task to visual dialog task. Specifically, we treat regularized model is a pre-trained model and present a novel variational attention mechanism to produce the context-relevant (condition-relevant) information by considering question and historical conversations. With

[1]The finding from the experiments shows that the style representations are the main factor affecting the language priors.

both pre-trained model and variational attention mechanism, the visual dialog system can effectively respond to various questions in this task.

Consider a supervised learning problem with training data $\mathcal{D} := \{\mathbf{x}, \mathbf{y}\}$, where $\mathbf{x}$ is input data, $\mathbf{y}$ is corresponding target. Conditional likelihood $p(\mathbf{y}|\mathbf{x})$ constructed with variational Bayesian attention framework, where the key concept of variational attention is to convert attention weights that obtained from query and keys to stochastic weights $\mathbf{w} \in \mathbb{R}^k$, where k is determined by the number of keys. Consider a Bayesian framework, similar to (Kingma & Welling, 2014), two data-dependent distributions were considered: *posterior distribution* $p(\mathbf{w}|\mathbf{x}, \mathbf{y})$ and *prior distribution* $p(\mathbf{w}|\mathbf{x})$. In practical, variational attention modules the attentive distribution that share a common structure with deterministic soft attention instead of adopting the deterministic attention weights directly. With variational inference (Hoffman et al., 2013; Kingma & Welling, 2014), it is equivalent to maximizing an evidence lower bound (ELBO) of the intractable log marginal likelihood $\log p(\mathbf{y}|\mathbf{x})$, where q_ϕ is approximate posterior distribution.

$$\mathbb{E}_{\mathbf{w}\sim q_\phi}[\log p(\mathbf{y}|\mathbf{x}, \mathbf{w})] - D_{\mathrm{KL}}(q_\phi(\mathbf{w}|\mathbf{x}, \mathbf{y})\|p(\mathbf{w}|\mathbf{x})) \tag{8}$$

In learning procedure, the model is trained by maximizing the objective function Eq. 8. Learning the distribution q_ϕ by minimizing $D_{\mathrm{KL}}(q_\phi(\mathbf{w}|\mathbf{x}, \mathbf{y})\|p(\mathbf{w}|\mathbf{x}))$. With the learning criterion, stochastic weights can be modeled by variational distributions.

In order to identify context-relevant conversations in visual dialog (Figure 1), we think about how to effectively and robustly identify the relevant components. Inspired by (Shayer et al., 2018; Peters & Welling, 2018), this paper proposes a variational attention mechanism that adopting Bernoulli distribution to sample weights discretely. Specifically, the variational attention mechanism constructed with Bernoulli distribution can ignore conditional-irrelevant information and provide conditional-relevant information to the pre-trained model. In probability theory and statistics, Bernoulli distribution is a discrete probability distribution of a random variable $\mathrm{w} \sim P_W(\mathrm{w}; p)$ which takes the value 1 with probability p and the value 0 with probability $1 - p$. Connecting with attention mechanism, there are a query $\mathbf{q}$ and keys $\boldsymbol{K}$. This paper define $\mathbf{p} = g(\boldsymbol{\alpha})$, where $\mathbf{p}$ is a vector that decide the parameters of the Bernoulli distribution, $\boldsymbol{\alpha}$ is deterministic attention weights obtained by standard attention mechanism (i.e. $\boldsymbol{\alpha} = f_{\text{attn}}(\mathbf{q}, \boldsymbol{K})$) and $g(\boldsymbol{\alpha}) = p_{\text{min}} + \boldsymbol{\alpha}(p_{\text{max}} - p_{\text{min}})$ is a scaling function which limits the output range between p_{min} and p_{max}. Note that $0 \leq \alpha_i \leq 1$ computed from deterministic attention calculation. The optimal solution would be $P_W(w_i = 1) = 1$ or 0 if equality is achieved. It is not desirable, as the Bernoulli distribution in such case would be too deterministic and always assign one or zero probability to the attentive weights. Therefore, the purpose of introducing the scaling function is to avoid the loss of randomness in the Bernoulli distribution.

Consider a variational attention mechanism. Since $\mathbf{w}$ is a random vector, $\mathbf{z}^a = \sum_i \mathrm{w}_i \mathbf{v}_i$ is also a random vector, where $\mathbf{v}_i$ is a value component of an attention mechanism. However, in practical, we cannot directly reparameterize the Bernoulli distribution, and discrete weights will cause training process to be non-differentiable. Therefore, this paper quotes Central Limit Theorem (CLT, details was provided in App. A.1) and reparameterize the probability distribution of sample average that sampled from Bernoulli. Specifically, we sample M independent and identically distributed (i.i.d.) samples from the Bernoulli distribution, so that we can compute the mean $\bar{\boldsymbol{\mu}} = \mathbb{E}_{\mathbf{w}_i\sim P_W}\left[\frac{1}{M}\sum_{i=1}^M \mathbf{w}_i\right]$ and variance $\bar{\boldsymbol{\sigma}}^2 = \mathrm{Var}_{\mathbf{w}_i\sim P_W}\left(\frac{1}{M}\sum_{i=1}^M \mathbf{w}_i\right)$ of sample average respectively. Following the linear computation of Gaussian distribution, it follows that $\mathbf{z}^a$ is a Gaussian distribution with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma}^2$, specifically:

$$\mathbf{z}^a \sim \mathcal{N}(\mathbf{z}^a; \boldsymbol{\mu}, \boldsymbol{\sigma}^2) = \mathcal{N}\left(\mathbf{z}^a; \sum_i \bar{\mu}_i \mathbf{v}_i, \sum_i \bar{\sigma}_i^2 \mathbf{v}_i^2\right) \tag{9}$$

Hence, we can obtain a variational distribution over the Bernoulli distribution. From the attentive distribution approximated by Gaussian we can easily sample $\mathbf{z}^a$ by reparameterization trick (Kingma & Welling, 2014). In practical, $\mathbf{z}^a = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon}$ is sampled from normal distribution (i.e. $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}; 0, I)$), where $\odot$ denotes element-wise multiplication. The combination of CLT approximation and reparameterization trick is also known as the local reparameterization trick. Given the sampled $\mathbf{z}^a$ we can proceed as usual and apply nonlinear or backprobagation. At test phase, instead of using the local reparameterization trick, a stochastic weights $\mathbf{w} \sim P_W(\mathbf{w}; g(\boldsymbol{\alpha}))$ is sampled and used.

In this task, we treat regularized VQA model as a pre-trained model and apply a variational attention mechanism to produce the context-relevant information. In practical, we let the current question as

		VQA v2				VQA-CP v2			
Model	γ	**Overall**	**Yes/No**	**Number**	**Other**	**Overall**	**Yes/No**	**Number**	**Other**
UpDn	–	62.85	**80.89**	**42.78**	54.44	39.49	45.21	11.96	42.98
AdvReg	–	62.75	79.84	42.35	55.16	41.17	65.49	15.48	35.48
Ours†	–	**63.07**	80.74	42.08	**55.73**	39.94	63.87	14.63	35.79
Ours	0.2	60.31	78.39	40.25	51.02	43.83	68.42	14.79	41.08
	0.4	58.81	76.27	39.38	49.82	47.99	70.51	15.69	**44.93**
	0.5	58.33	76.84	38.81	49.73	**48.33**	**71.03**	**18.25**	44.43

Table 1: Evaluations of our model on VQA v2 and VQA-CP v2. In the table, UpDn (Anderson et al., 2018) does not deal with the language priors problem, so does our model without introduce regularization (mark † in the table). AdvReg (Ramakrishnan et al., 2018) introduce adversarial learning to alleviate the language priors.

a query, and historical conversations as keys. Consider a variational Bayesian framework, we can integrate a context-relevant loss to visual dialog system.

$$\begin{aligned}\mathcal{L}_{\mathrm{R}}(\mathbf{x}, \mathbf{x}^c, \mathbf{x}^v, \mathbf{y}) = \mathbb{E}_{\mathbf{z}^a \sim q_\phi}[-\log p(\mathbf{y}|f_\theta(\mathbf{x}), \mathbf{x}^c, \mathbf{x}^v, \mathbf{z}^a)] \\ + D_{\mathrm{KL}}(q_\phi(\mathbf{z}^a|\mathbf{x}, \mathbf{x}^c, \mathbf{y})\|p(\mathbf{z}^a|\mathbf{x}, \mathbf{x}^c)) + \beta\mathcal{I}_{\mathrm{Dis}}\end{aligned} \tag{10}$$

where $\mathbf{x}^c$ denotes the conversations in visual dialog. The details of overall model architecture was provide in App. A.2.

4 Experiments

4.1 Datasets and evaluation metrics

For pre-trained VQA task, our approach is evaluated on the most commonly used benchmark dataset: Visual Question Answering under Changing Priors (VQA-CP v2) (Agrawal et al., 2018). If a well-designed model has language priors problem during training, there will be no way to escape during testing. The dataset are created by reorganizing the training and validation splits of the VQA v2 dataset (Goyal et al., 2017). For comparison, we also evaluate our approach on the VQA v2 dataset which consists of a large number of real images collected from COCO dataset (Lin et al., 2014), free-form natural language questions and concise answers. For visual dialog task, we use the VisDial v1.0 dataset (Das et al., 2017) to evaluate the proposed method. Images for the training split are also from COCO dataset. Each dialog is crowd-sourced on a distinct image and consists of 10 rounds of dialog turns. Each question is also accompanied by a list of 100 randomly generated potential responses, which the model requested to rank them. In addition, in order to evaluate the consistency of the ranking, "dense annotations" was provided with a relevant scores between zero and one containing 100 candidate response.

For evaluation metrics, this paper follows standard evaluation method (Antol et al., 2015; Agrawal et al., 2018) to evaluate the pre-trained VQA model. According to the catalog provided on dataset, we calculated the accuracy under different catalogs, namely "Yes/No", "Number" and "Other" catalogs. On the other hand, retrieval metrics (Das et al., 2017) were used to evaluate the effectiveness of variational attention in visual dialog. Specifically, Mean, R@k, MRR and NDCG were introduced to evaluation. Among them, the first three evaluation metrics evaluate the ranking position of human responses, and the last one focuses on evaluating the relevance of the overall sorted responses.

4.2 Quantitative results

Visual question answering. For the quantitative evaluation, we evaluated the accuracy on VQA v2 and VQA-CP v2 datasets. For fair comparison, we compared the methods constructed based on object detection (App. A.2) framework and reported the results in Table 1, which shows that UpDn and the model we propose that without introducing regularization get the better performance in VQA v2, it shows that when the model suffers the language priors problem, it can perform well on VQA v2. On the contrary, both of them have a dramatic decline on VQA-CP v2 evaluation. In addition, through the adjustment of the hyperparameter γ, we could adjust the degree of the language

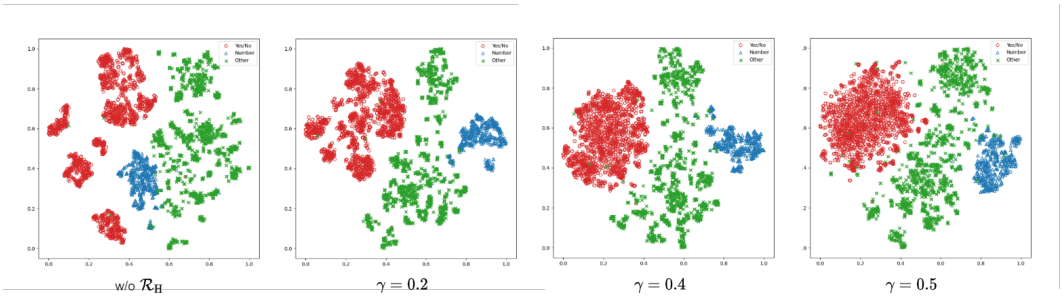


Figure 3: Latent space visualization for different constraints of style representation, where the red, blue and green points respectively represent the "Yes/No" questions, "Number" questions and "Other" questions.

Model	NDCG↑	MRR↑	R@1↑	R@5↑	R@10↑	Mean↓
RvA	55.86	64.42	50.71	81.50	90.15	4.06
MCA-I-H	**60.27**	64.33	51.12	80.91	89.65	4.24
MVAN	60.17	65.33	51.86	82.40	**90.90**	3.88
Ours (soft)	60.23	65.18	51.74	82.38	90.88	3.72
Ours	60.21	**65.74**	**53.46**	**82.63**	90.79	**3.69**

Table 2: Evaluate on VisDial v1.0. This paper compare the models which constructed based on soft attention mechanism in different manners. RvA (Niu et al., 2019) recursively calculate vision attention based on dialog context. MCA-I-H (Agarwal et al., 2020) stacked the self attention and cross attention in image representations and text representations. MVAN (Park et al., 2021) considers the multi view such as word level attention and sentence level attention.

priors alleviation. Experiments show that when the value of γ is set at 0.5, there would be better performance. Furthermore, in order to investigate the effect of the proposed method on language priors under different constraints, we visualized the pattern of the regularization term for the representation learning of the style representations by t-SNE plots (Van der Maaten & Hinton, 2008). In Figure 3 we could see that style representations are divided the same type of questions into many clusters. Through the corroboration of numerical report and the visualized latent distribution, it can be verified that the regularization adjustment can effectively alleviate the clustering phenomenon and language priors.

Visual dialog. For visual dialog task, we compared models (Niu et al., 2019; Agarwal et al., 2020; Park et al., 2021) that introduce soft attention to vision information and text information in different manner. On the other hand, in order to compare the quality of the proposed method, we implement a soft attention version of the proposed method and evaluate it under same setting. The results are reported in Table 2. The proposed method get the better performence than other methods in MRR, R@1 and Mean, which shows that it is more specialized in predicting human responses than sorting candidate responses. Furthermore, in order to investigate the robustness of the proposed method to historical conversations, we randomly selected a small number of context sentences (not including captions) in the validation split of VisDial v1.0 and randomly select less of three rounds of dialog in each conversation and modify less of three words, which is to demonstrate when the historical dialog is wrong or irrelevant to current question. We report the results in Table 3. It could be seen from the results that most models would be affected, but the proposed method alleviate the decline. Similarly, we could also find that the soft attention version has a lot of decline compared to the variational attention version.

4.3 Qualitative Results

Visual question answering. For the qualitative results, we quantitatively examine the effectiveness of the proposed strategy. We reported the results in App. A.3, this paper show the results of ablation study include the effect of not introducing regularization and regularization under different

Model	NDCG↑	MRR↑	R@1↑	R@5↑	R@10↑	Mean↓
RvA	53.24	60.58	49.31	80.67	86.92	4.23
MVAN	57.83	61.43	50.03	80.89	87.69	4.09
Ours (soft)	58.75	61.31	49.89	80.74	87.53	4.17
Ours	**60.03**	**62.82**	**51.12**	**81.48**	**88.23**	**3.92**

Table 3: Robustness test with the noise context

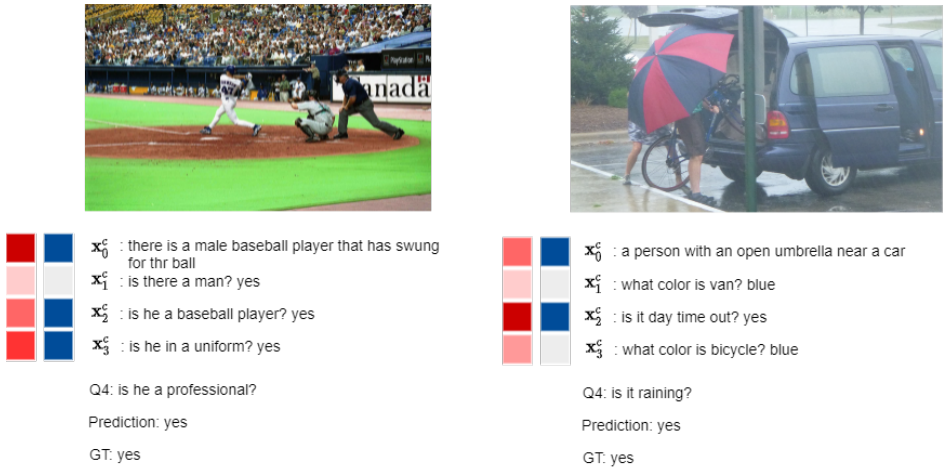


Figure 4: Qualitative results on the visual dialog task. Visualization of the attention map with the different algorithm. The red (soft attention) considers all components, the blue (discrete variational attention) identifies whether it is related to the question. The darker the color indicates that the model determines that it is more important to the current question.

weights on the results. The experiment results show that adjusting γ properly can effectively reduce the occurrence of language priors in VQA.

Visual dialog. For visual dialog, we visualize the attention map with different approaches based on the same architecture as shown in Figure 4. The red one is the results of soft attention mechanism, the blue one is the results of proposed variational attention mechanism, we can see that the variational attention is discretely consider relevant contextual information instead of considering all history no matter they are relevant or not. Furthermore, we analyzed error examples on visual dialog for the proposed method, the results show that when it is the first round of dialog, there are not many bases that can calculate context weights, which leads to limiting on the randomness of the model and some unreasonable situations, the detail was provided in App. A.4.

5 Conclusions

In this paper, this paper present a unified approach to deal with context-free issue in visual dialog. First of all, we introduce VQA task for pre-trained so that the model can effectively respond context-free questions. In addition, this paper present a regularization scheme and apply it to a disentanglement network that constructed with mutual information, which aims to alleviate language priors and to strengthen the dependence of vision and language. Secondly, in order to effectively identify context-relevant conversations, this paper present a discrete variational attention mechanism to produce context-relevant condition, so that we can transfer VQA to visual dialog. The experiment results show that the proposed method can effectively alleviate language priors and improve the robustness of the system.

REFERENCES

David Barber Felix Agakov. The IM algorithm: a variational approach to information maximization. *Advances in Neural Information Processing Systems*, 16(320):201, 2004.

Shubham Agarwal, Trung Bui, Joon-Young Lee, Ioannis Konstas, and Verena Rieser. History for visual dialog: Do we really need it? In *Proc. of Annual Meeting of Association for Computational Linguistics*, pp. 8182–8197, 2020.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4971–4980, 2018.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6077–6086, 2018.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proc. of IEEE International Conference on Computer Vision*, pp. 2425–2433, 2015.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proc. of International Conference on Learning Representations*, 2015.

Hareesh Bahuleyan, Lili Mou, Olga Vechtomova, and Pascal Poupart. Variational attention for sequence-to-sequence models. In *Proc. of International Conference on Computational Linguistics*, pp. 1672–1682, 2018.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *Proc. of International Conference on Machine Learning*, pp. 531–540, 2018.

Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. Rubi: Reducing Unimodal Biases for visual question answering. *Advances in Neural Information Processing Systems*, 32: 841–852, 2019.

Hung-Jen Chen, Ka-Ming Hui, Szu-Yu Wang, Li-Wu Tsao, Hong-Han Shuai, and Wen-Huang Cheng. Beautyglow: On-demand makeup transfer framework with reversible generative network. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10042–10050, 2019.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Proc. of International Conference on Neural Information Processing Systems*, pp. 2180–2188, 2016.

Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. Club: A Contrastive log-ratio upper bound of mutual information. In *Proc. of International Conference on Machine Learning*, pp. 1779–1788, 2020a.

Pengyu Cheng, Martin Renqiang Min, Dinghan Shen, Christopher Malon, Yizhe Zhang, Yitong Li, and Lawrence Carin. Improving disentangled text representation learning with information-theoretic guidance. In *Proc. of Annual Meeting of the Association for Computational Linguistics*, pp. 7530–7541, 2020b.

Michael Cogswell, Jiasen Lu, Rishabh Jain, Stefan Lee, Devi Parikh, and Dhruv Batra. Dialog without dialog data: Learning visual dialog agents from vqa data. *Advances in Neural Information Processing Systems*, 33, 2020.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 326–335, 2017.

Yuntian Deng, Yoon Kim, Justin Chiu, Demi Guo, and Alexander Rush. Latent alignment and variational attention. *Advances in Neural Information Processing Systems*, 31:9712–9724, 2018.

Guillaume Desjardins, Aaron Courville, and Yoshua Bengio. Disentangling factors of variation via generative entangling. *arXiv preprint arXiv:1210.5474*, 2012.

Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time, i. *Communications on Pure and Applied Mathematics*, 28(1):1–47, 1975.

Xinjie Fan, Shujian Zhang, Bo Chen, and Mingyuan Zhou. Bayesian attention modules. *Advances in Neural Information Processing Systems*, 33, 2020.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6904–6913, 2017.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *Proc. of International Conference on Learning Representations*, 2018.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(5), 2013.

Sheng-Jhe Huang and Jen-Tzung Chien. Attribute decomposition for flow-based domain mapping. In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1710–1714. IEEE, 2021.

HyeongJoo Hwang, Geon-Hyeong Kim, Seunghoon Hong, and Kee-Eung Kim. Variational interaction information maximization for cross-domain disentanglement. *Advances in Neural Information Processing Systems*, 33, 2020.

Gi-Cheon Kang, Jaeseo Lim, and Byoung-Tak Zhang. Dual attention networks for visual reference resolution in visual dialog. In *Proc. of Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing*, pp. 2024–2033, 2019.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *stat*, 1050:1, 2014.

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 *Proc. of European Conference on Computer Vision*, pp. 740–755. Springer, 2014.

Emile Mathieu, Tom Rainforth, Nana Siddharth, and Yee Whye Teh. Disentangling disentanglement in variational autoencoders. In *Proc. of International Conference on Machine Learning*, pp. 4402–4412, 2019.

Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. In *European Conference on Computer Vision*, pp. 336–352. Springer, 2020.

Yulei Niu, Hanwang Zhang, Manli Zhang, Jianhong Zhang, Zhiwu Lu, and Ji-Rong Wen. Recursive visual attention in visual dialog. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6679–6688, 2019.

Sungjin Park, Taesun Whang, Yeochan Yoon, and Heuiseok Lim. Multi-view attention network for visual dialog. *Applied Sciences*, 11(7):3009, 2021.

Jorn W. T. Peters and Max Welling. Probabilistic binary neural networks. *arXiv preprint arXiv:1809.03368*, 2018.

Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. Overcoming language priors in visual question answering with adversarial regularization. *Advances in Neural Information Processing Systems*, 31:1541–1551, 2018.

Oran Shayer, Dan Levi, and Ethan Fetaya. Learning discrete weights using the local reparameterization trick. In *Proc. of International Conference on Learning Representations*, 2018.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11):2579–2605, 2008.

A APPENDIX

A.1 CENTRAL LIMIT THEOREM

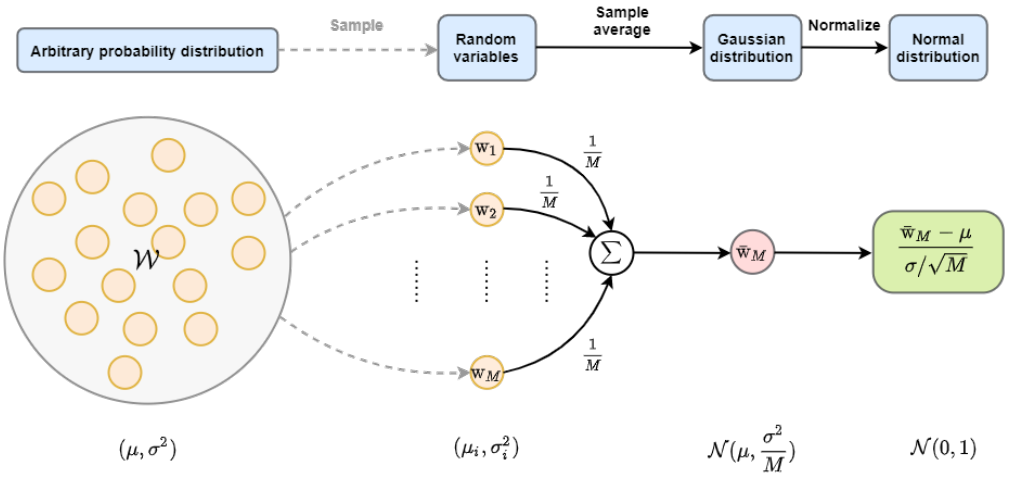


Figure 5: Concept of the Central Limit Theorem.

Central Limit Theorem (CLT) is one of the most important and commonly used methods in probability theory and statistics. The content of the theorem shows that when $\{\mathrm{w}_1, \mathrm{w}_2, \ldots \mathrm{w}_M\}$ are independent and identically distributed (i.i.d.) random variables randomly sampled from arbitrary probability distribution with mean μ and variance σ^2, the sample average would follows

$$\lim_{M\to\infty} \frac{\bar{\mathrm{w}}_M - \mu}{\sigma/\sqrt{M}} \sim \mathcal{N}(0,1) \quad ; \quad \bar{\mathrm{w}}_M = \frac{\mathrm{w}_1 + \mathrm{w}_2 + \cdots + \mathrm{w}_M}{M}$$

It states that if there are sufficiently large random samples from the population with replacement, then the distribution of the normalized sample average will be approximately normally distributed.

A.2 MODEL ARCHITECTURE OF VISUAL DIALOG SYSTEM

As described in Section 3.3, this paper expand the VQA task to visual dialog task. Specifically, we treat VQA model as a pre-trained model and we apply a variation attention mechanism to produce the context relevant (condition relevant) information. The overall model architecture shown in Figure 6. In visual dialog task, a system is given an image $\mathbf{x}^v$, the 'ground truth' dialog history $\mathbf{x}^c$ (including the image caption), the question $\mathbf{x}$, and a list of $N = 100$ candidate answers, and asked to return a sorting of the candidate answers. For pre-trained model, with disentanglement network f_θ, we adopt content embedding $\mathbf{z}^c$ as a query to compute the attention weights with object representations, where object representations are encoded by Faster R-CNN, which can identify objects in the image and encode them into respective representations. With attention weights between each proposals and content from question, we can take out the k highest attention weight values of proposals by gating function, it means that these k proposals in image are most relevant to the content representation of a given question. On the other hand, we treat the output of variational attention mechanism $\mathbf{z}^a$ is a condition of pre-trained model so that we can manipulate the model according to different conditions. In other words, the system can handle whether or not there is an associated contextual conversations. In decoder part, we encode sentence representations of answer candidates. In practical, we rank them according to the dot products of the candidate representations and output of overall architecture representation, then apply the softmax function to obtain the probability distribution of the candidates so that we can optimize model parameters.

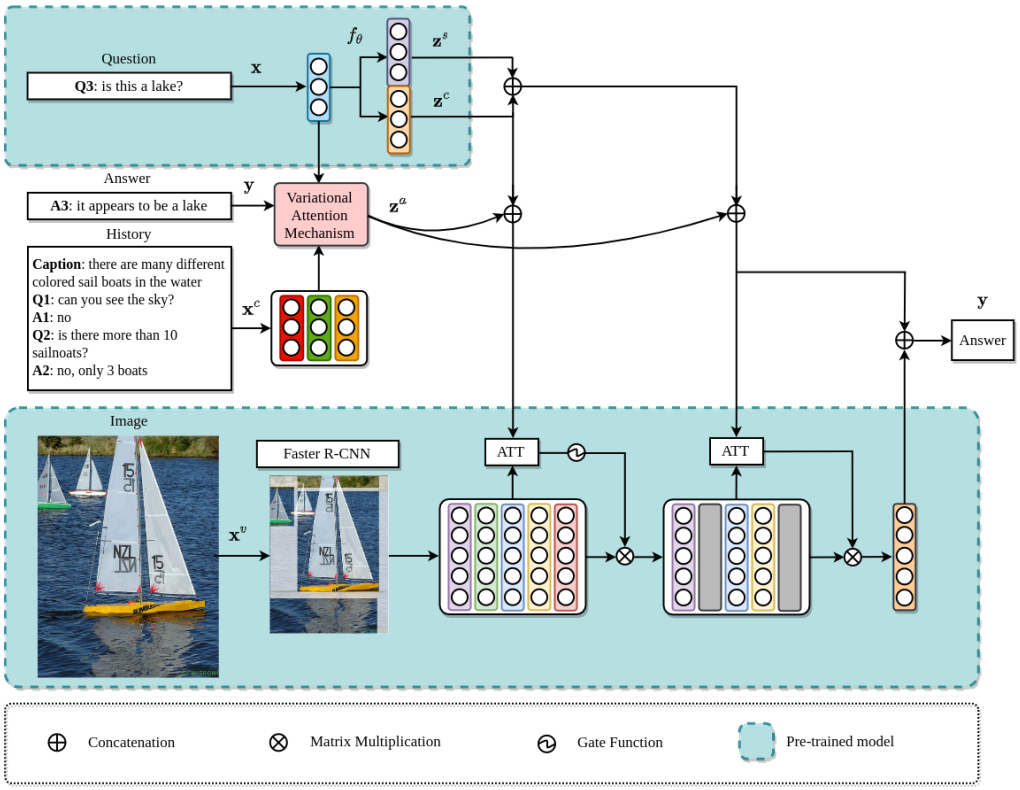


Figure 6: The overall model architecture of the visual dialog system.

A.3 QUALITATIVE RESULT OF VQA

For the qualitative results, this paper evaluate the effectiveness of the proposed method on VQA-CP v2 dataset and show that the effect of regularization under different setting. The results show that proper adjustment of γ can effectively alleviate the occurrence of the language priors in VQA.

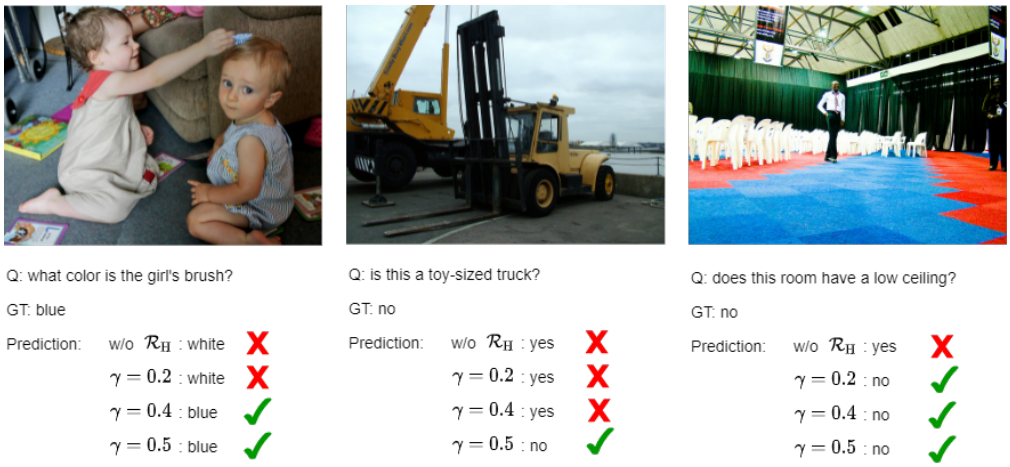


Figure 7: Qualitative results of language priors in VQA. In the figure, it can be seen that the adjustment of γ can transform the response originally affected by language priors into a response that correctly matches the image content.

A.4 ERROR ANALYSIS FOR DISCRETE VARIATIONAL ATTENTION MECHANISM

For error analysis, we visualized the attention map of the first round of conversation. Since it is the first round of dialog, there are not many historical conversations that can calculate attention weights, which leads to restrictions on the randomness of the model and some unreasonable situations. As shown in Figure 8, we can see that even if the information of the ground-truth response is hidden in the context, there is still a low probability that it will not be adopted by the model. Similarly, even if the question is not related to the context, the model still has a high probability of adopting its information and treating it as important information.

Figure 8: Error analysis on the proposed method.