Modeling Coreference Relations in Visual Dialog

Mingxiao Li
KU Leuven
mingxiao.li@kuleuven.be

Marie-Francine Moens
KU Leuven
sien.moens@cs.kuleuven.be

Abstract

Visual dialog is a vision-language task where an agent needs to answer a series of questions grounded in an image based on the understanding of the dialog history and the image. The occurrences of coreference relations in the dialog makes it a more challenging task than visual question-answering. Most previous works have focused on learning better multi-modal representations or on exploring different ways of fusing visual and language features, while the coreferences in the dialog are mainly ignored. In this paper, based on linguistic knowledge and discourse features of human dialog we propose two soft constraints that can improve the model's ability of resolving coreferences in dialog in an unsupervised way. Experimental results on the VisDial v1.0 dataset shows that our model, which integrates two novel and linguistically inspired soft constraints in a deep transformer neural architecture, obtains new state-of-the-art performance in terms of recall at 1 and other evaluation metrics compared to current existing models and this without pretraining on other vision-language datasets. Our qualitative results also demonstrate the effectiveness of the method that we propose. [1]

1 Introduction

Recently, with the unprecedented advances in computer vision and natural language processing, we have seen a considerable effort in developing artificial intelligence (AI) agents that can jointly understand visual and language information. Visual-language tasks, such as image captioning (Xu et al., 2015) and visual question-answering (VQA) (Antol et al., 2015), have achieved inspiring progress over the past few years. However, the applications of these agents in real-life are still quite limited, since they cannot handle the situation when continuous information exchange with a human is necessary, such as in visual-language navigation (Anderson et al., 2018b) and visual dialog (Das et al., 2017). The visual dialog task can be seen as a generalization of VQA. Both tasks require the agent to answer a question expressed in natural language about a given image. A VQA agent needs to answer a single question, while a dialog agent has to answer a series of language questions based on its understanding of visual content and dialog history. Compared to VQA, the visual dialog task is more difficult because it demands the agent to resolve visual coreferences in the dialog. Considering the example in Figure 1. when the agent encounters question 6 "do they look old or new ?" and questions 9 "are they tall or short ?", it has to infer that the pronoun "they" in these two questions refers to different entities in the image or the dialog history.

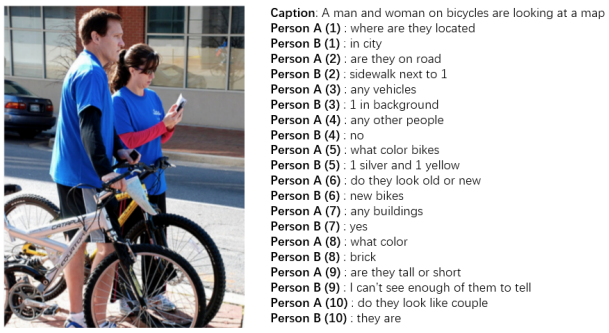

Figure 1: An example taking from the VisDial v1.0 dataset. The questioner (Person A) sees the caption and tries to understand the whole scene of the image by asking questions to the answerer (Person B) who can see the whole image.

This paper studies how we can improve the results of a visual dialog task by better resolving the coreference relations in the dialog. In this work we restrict coreference resolution to pronouns. We

[1] Our code are released on: https://github.com/Mingxiao-Li/Modeling-Coreference-Relations-in-Visual-Dialog

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 3306–3318
April 19 - 23, 2021.

use a multi-layer transformer encoder as our baseline model. Based on the assumption that the output contextual embedding of a pronoun and its antecedent should be close in the semantic space, we propose several soft constraints that can improve the model's capability of resolving coreferences in the dialog in an unsupervised way (i.e., without ground truth coreference annotations in the training data). Our first soft constraint is based on the linguistic knowledge that the antecedent of a pronoun can only be a noun or noun phrase. To integrate this constraint in the baseline model, we introduce a learnable part-of-speech (POS) tag embedding and a part-of-speech tag prediction loss. Inspired by the observation that in human dialog the referents of pronouns often occur in nearby dialog utterances, we propose a second soft constraint using a sinusoidal sentence position embedding, which aims to enhance local interactions between nearby sentences.

Our contributions are as follows: First, as a baseline we adapt the multi-layer transformer encoder to the visual dialog task and obtain results comparable to the state of the art. Second, we propose two soft constraints to improve the model's ability of resolving coreference relations in an unsupervised way. We also perform an ablation study to demonstrate the effectiveness of the introduced soft constraints. Third, we conduct a qualitative analysis and show that the proposed model can resolve pronoun coreferents by making sure that in the neural architecture the pronoun mostly attends to its antecedent.

2 Related Work

Visual Dialog. The Visual Dialog task is proposed by Das et al. (2017), where a dialog agent has to answer questions grounded in an image based on its understanding of the dialog history and the image. Most of previous work focuses on using an attention mechanism to learn interactions between image, dialog history and question. Gan et al. (2019) use an attention network to conduct multi-step reasoning in order to answer a question. Niu et al. (2019) propose a recursive attention network, which selects relevant information from the dialog history recursively. Kang et al. (2019) apply a multi-head attention mechanism (Vaswani et al., 2017) to learn mutimodal representations. Schwartz et al. (2019) fuse information from all entities including question, answer, dialog history, caption and image using a factor graph. Murahari et al. (2019) propose two-stage training. They first pretrain their transformer based two-stream attention network on other visual-language datasets, then finetune it on the visual dialog dataset. Other approaches consider different learning methodologies to model the visual dialog task, for example, Lu et al. (2017) use adversarial learning and Yang et al. (2019) apply reinforcement learning.

Coreference Resolution. Coreference resolution aims at detecting linguistic expressions referring to the same entities in the context of the discourse. The task has been dominated by machine learning approaches since the first learning based coreference resolution system was proposed by Connolly et al. (1997). Before Lee et al. (2017) proposed the first end-to-end neural network based coreference resolution system, most of the learning-based systems have been built with hand engineered linguistic features. Durrett and Klein (2013) use surface linguistic features, such as mention type, the semantic head of a mention, etc., and their combinations to build a classifier to determine if two mentions refer to the same entity. Do et al. (2015) adopt integer linear programming (ILP) to introduce coreference constraints including centering theory constraints, direct speech constraints and definite noun phrase and exact match constraints in the inference step in order to adapt an existing coreference system trained on the newswire domain to short narrative stories without any retraining. Recently, Joshi et al. (2019) apply a BERT model to coreference resolution and achieve promising results on the OntoNotes corpus (Pradhan et al., 2012) and the GAP dataset (Webster et al., 2018). Different from all coreference systems mentioned above, which rely on supervised learning and on a dataset annotated with coreference links, our work focuses on applying soft linguistic constraints to improve the model's ability of resolving coreferents in an implicit and unsupervised way. Similar to the work of Venkitasubramanian et al. (2017) that operates on language and vision information, our model uses attention to jointly learn multi-modal representations.

3 Methodology

In this section, we formally describe the visual dialog task (Das et al., 2017) and the approaches we propose. In visual dialog, given an image I, the image caption C and the dialog history until round $t-$

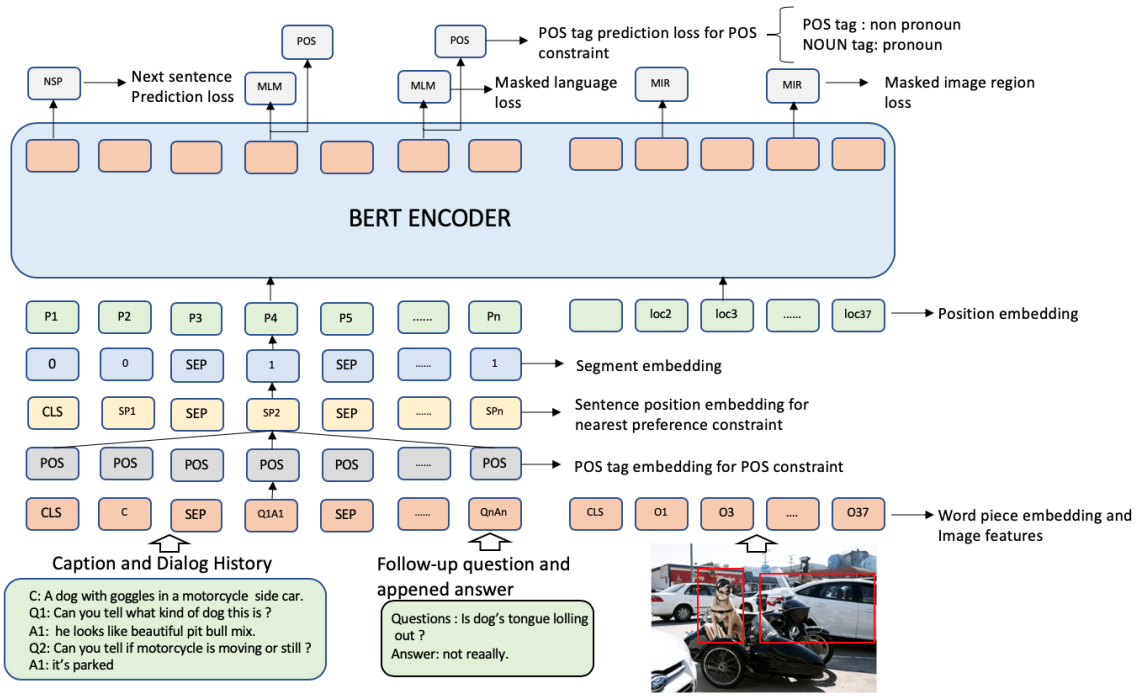

Figure 2: The model architecture with the two soft constraints that we propose. The baseline model takes the image feature I, image caption C, dialog history H_t, follow-up question Q_t and appended candidate answer A_t as input and is trained by using a masked language model (MLM), masked image region (MIR) and next sentence prediction (NSP) losses. Two soft constraints are integrated into the model by adding POS tag embedding and sentence position embedding to the input language sequence embedding and by introducing a new POS tag prediction objective during training.

1, $H = ((Q_1, A_1), (Q_2, A_2), \cdots, (Q_{t-1}, A_{t-1}))$, which is a sequence of question-answer pairs expressed in natural language and grounded in the image, a dialog agent is expected to correctly answer the question at round t by choosing the answer from 100 candidate answers $A_t = \{A_t^1, A_t^2, \cdots, A_t^{100}\}$.

We first introduce the baseline visual dialog model in Section 3.1, followed by the detailed explanation of our proposed soft coreference constraints and how we integrate these into the baseline model in Section 3.2. The coreference constraints are based on linguistic knowledge, so consequently our method is an example of how to integrate linguistic knowledge into a neural transformer architecture. Figure 2 shows the architecture of our proposed model.

3.1 Baseline Model

Transformer Encoder. We use a multi-layer transformer (Vaswani et al., 2017) encoder as our baseline model. The computation within a single layer transformer encoder is presented in appendix A.1, and the details of the input and training objective functions are illustrated in below subsections. The main idea of applying the transformer architecture to the visual dialog task is to use the multi-head self-attention mechanism to implicitly learn the intra and inter interactions within the single modality and between the different modalities (in this case language and vision), respectively.

Linguistic Representation. Following the monolingual BERT model (Devlin et al., 2019) and the multi-modal BERT model (Lu et al., 2019; Li et al., 2019; Su et al., 2020), we use the WordPiece (Wu et al., 2016) tokenization tool to tokenize each input sequence into word pieces sequence. Then the sum of the word piece embedding, position embedding and segment embedding, where the segment embedding is used to differentiate questions from answers and to delimit boundaries of question-answer pairs, are taken as the language sequence input of the model.

Image Representation. Following the multi-modal BERT model (Anderson et al., 2018a; Lu et al., 2019; Li et al., 2019; Su et al., 2020), we use Faster-RCNN (Ren et al., 2015) with ResNet (He et al., 2016) backbone to detect objects in the

image and keep the top-36 detected objects and their corresponding bounding boxes. The representation of each detected object is obtained by applying a mean-pooled convolution on the region of that object. We also project a $5-d$ geometrical representation of each box, including the normalized top-left and bottom-right coordinates of the detected objects and the fraction of the area they cover, to the same dimensions as the image feature vector. In this way we obtain a vector with the same dimensions as the feature representation of the image. The final input image representation is the sum of its geometrical and feature representations. To avoid missing image information that is not captured by the top-36 bounding boxes, we also concatenate the mean-pooled feature vector of the whole image to the beginning of the image region sequence.

Multi-modal Input. As the transformer encoder receives a sequence of tokens as input, to feed both image and language into the model, we simply concatenate the language sequence embedding and image region sequence representation to form a whole input sequence. Like in the BERT model, a special token [CLS] is added to the beginning of the input sequence to perform the next sentence prediction task. We also use another special token [SEP] to separate each question-answer pair and the two modalities. Our input sequence can be formulated as follows: $Input = \{[CLS], C, [SEP], Q_1 A_1, [SEP], Q_2 A_2, \cdots, Q_t A_t, [SEP], O_0, O_1, O_2, \cdots, O_{36}\}$, where C is the image caption. Q_i,A_i are question and answer at round i, and $O_{0\sim36}$ denote the input image region features.

Multitasks Training Objectives. To make the model learn a good alignment between different modalities, we utilize three losses: masked language model loss (MLM), masked image region loss (MIR), and next sentence prediction loss (NSP). Similar to the MLM in BERT, we randomly mask 15% word pieces in the language sequence by replacing the word piece with a special token [MASK], while in MIR, we randomly set 15% of the image region features to zero vectors. The model is trained to recover the masked words and predict the semantic category of the masked image regions:

$$L_{MLM} = -E_{(I,w)\sim D} log P(w_m | w_{\setminus m}, I) \quad (1)$$

$$L_{MIR} = \sum_{i}^{k} KL(P_m || P_g) \quad (2)$$

where $w_{\setminus m}$ and I denote the word sequence excluding the masked words w_m, and image regions, respectively. KL represents the KL divergence loss. P_m is the model output distribution and P_g is the ground truth classification distribution.

Recall that the visual dialog system aims to find the correct answer among the 100 candidate answers. We realize this in a discriminative manner by using the next sentence prediction loss (NSP). We randomly select 1 wrong answer from the candidate answers to generate negative samples, together with the ground-truth to form a balanced training dataset. During training, a candidate answer is appended to the dialog sequence, and the model is trained to predict whether or not the appended answer is the correct answer to the current question:

$$L_{NSP} = -E_{(I,w)\sim D} log P(\hat{y} | I, w) \quad (3)$$

where $\hat{y} \in [0, 1]$ is the output probability of the binary classifier at the last layer using the special [CLS] tag representation, which indicates the probability of the appended answer being correct. During inference, we rank the 100 candidate answers using their NSP score, which is the $\hat{y}$ in the above equation. During training, the total loss is the sum of MLM, MIR and NSP losses:

$$L_{total} = L_{MLM} + L_{MIR} + L_{NSP} \quad (4)$$

3.2 Soft Coreference Constraints

As discussed before, the existence of pronouns in language makes the visual dialog a more challenging task than VQA. A naive way to reduce the difficulty would be to use a loss to guide the model to jointly learn to resolve the coreferences in the dialog and to generate an answer to the question. However, the lack of coreference annotations in the visual dialog dataset prevents from using this supervised learning approach. Although it is impossible to resolve coreferences directly in the model, we propose to use linguistic knowledge to improve the model's ability to implicitly resolve the coreferences in an unsupervised way. We do so by exploiting the attention mechanisms of the

transformer architecture where attention weights act as soft constraints to guide the training of the model. The intuition behind it is the following. As the baseline model will output a contextual representation for each input token in the last layer, if a pronoun refers to a noun or noun phrase in the input sequence, the output contextual embedding of this pronoun and its antecedent should be close in the semantic space, which also means that the pronoun should attend most to its antecedent.

Part-Of-Speech Constraint. Our first proposed soft constraint is based on the linguistic knowledge that if the antecedent of a pronoun exists, it can only be a noun or noun phrase. We use the POS tag information and introduce a POS tag prediction loss to help pronouns to find nouns in the dialog. The Stanford CoreNLP POS tagger (Manning et al., 2014) is used to obtain the POS tag of each word in the input dialog, and all sub-word splits from a word share the same POS tag. Similar to the word embedding, we use a learnable embedding for each POS tag, which is further summed with the word piece embedding, position embedding and segment embedding to form the input sequence embedding of the model. In POS prediction loss, similar to the MLM loss, we randomly mask 15% of POS tag of input tokens. Then, the model is trained to predict the ground truth POS tag[2] of non-pronoun masked words, while for masked pronouns we replace the ground truth PRP tag with NN tag forcing the model to learn the contextual pronoun embedding that is close to nouns in the semantic space. The POS prediction loss can be formulated as the equation below:

$$
\begin{aligned}
L_{POS} &= -E_{(w,I)\sim D}(P_{non-pronoun} + P_{pronoun}) \\
P_{non-pronoun} &= -logP(POS(w)|w_{\setminus m}, I) \\
P_{pronoun} &= logP(NN|w_{\setminus m}, I))
\end{aligned} \tag{5}
$$

where $w_{\setminus m}$ denote all unmasked words, and D is the dataset. This soft constraint will make pronouns focus more on nouns instead of other words such as verb, adverb or adjective, etc. As it does not violate any linguistic rules, it will not introduce a bias to the language model.

Nearest Preference Constraint. Our second soft constraint is inspired by the observation that in human dialog a pronoun is more likely to refer to the noun that is close to it. For example, in the visual dialog shown in Figure 1, in round 9 the pronoun "they" refers to the "buildings", in round 6 "they" refers to "bikes". However, it is not always the case that pronouns refer to the noun closest in the previous utterances hence our soft constraint. In visual dialog, some pronouns refer to noun phrases that occur much earlier in the discourse - skipping a few rounds - as utterances are very short. To integrate this preference into the model, we adapt the sinusoidal word position embedding proposed in (Vaswani et al., 2017) and introduce a sentence position embedding:

$$PE_{pos,2i} = \frac{1}{k} sin(pos/(M + 10000^{\frac{2i}{d}})) \tag{6}$$

$$PE_{pos,2i+1} = \frac{1}{k} cos(pos/(M + 10000^{\frac{2i}{d}})) \tag{7}$$

where pos is the sentence position in the dialog, d is the hidden state size and the M is the maximum number of sentences, which is 21 in this visual dialog task. k is a scaling factor to control the local interactions brought by sentence position embedding and we use $k = 100$. Compared to the original sinusoidal position embedding proposed in (Vaswani et al., 2017), our sentence position embedding has one more scaling factor and one more element M in the denominator, which aims at restricting the product of the sentence position embedding to be a monotonically decreasing function with respect to $|pos_1 - pos_2|$.

$$PE_{pos_1} \cdot PE_{pos_2} = \frac{1}{k^2} \sum_{i=0}^{\frac{d}{2}-1} cos(w_i \Delta pos) \tag{8}$$

where $w_i = 1/(M + \epsilon^{\frac{2i}{d}})$, and Δpos denotes the distance between two positions. ϵ is a parameter, which makes the wavelength of the sinusoidal function in each dimension to form a geometrical progression.[3] Since $\Delta pos \in [-M, M]$, $w_i \Delta pos \in [-1, 1]$, it follows that $cos(w_i \Delta pos) == cos(w_i |\Delta pos|)$ which is monotonically decreasing in the region of $[0, 1]$.The details of the derivation of equation 8 are presented in appendix A.2. The closer two sentences are, the larger of the product of their sentence position embedding, resulting in stronger local interactions between nearby sentences in the dialog. This soft constraint can be easily integrated into the model by adding the sentence position embedding to the

[2]We use the POS tags used in Penn Treebank (Marcus et al., 1993).

[3]Following Vaswani et al. (2017), we set $\epsilon = 10000$.

input sequence embedding including word piece embedding, position embedding, segment embedding and POS tag embedding.

4 Experiments

4.1 Dataset

We use the real environment VisDial v1.0 dataset in this work. The VisDial v1.0 has 123k, 2k and 8k dialogs for training, validation and test, respectively. Each dialog contains one image with its caption from the MS-COCO dataset (Lin et al., 2014), and ten rounds of question-answer pairs, which were collected from the chatting log of one questioner and one answerer who both were discussing the image. For each question, except for the correct answer, the dataset also provides another 99 candidate answers to form an answer pool from which a model needs to select the relevant answer. Note that, although the VisDial v1.0 dataset also contains a small dense annotation set in which a relevance score is given to each candidate answer in the answer pool, we do not use this small dataset to finetune all our models, as we consider recall at 1 as main evaluation metric, and finetuning on this dense dataset could degrade the model's performance when measured by recall at 1 (Murahari et al., 2019).

4.2 Evaluation Metrics

We evaluate our proposed models using three evaluation metrics: (1) mean reciprocal rank (MRR) (Voorhees, 1999); (2) recall @ k, that is, the existence of the ground truth response in the top-k ranked items of the response list generated by the model with k = 1, 5, or 10; and (3) mean rank (Mean) of the ground truth response, that is, the average rank of the ground truth answer in the model's output ranked list (lower is better).

4.3 Training Details

Inspired by the open source code[4] of Murahari et al. (2019), we have implemented our model using the PyTorch framework (Paszke et al., 2019). Our model has the same configuration as the $\text{BERT}_{\text{BASE}}$ model, which contains 12 transformer layers and each layer has 12 attention heads with a hidden state size of 768. We set the maximum input length to be 256 including 37 image features. All models were trained using the Adam (Kingma and Ba, 2014) algorithm with a base learning rate of $5e^{-5}$. A linear learning rate decay schedule is employed to increase the leaning rate from $5e^{-6}$ to 5^{-5} over $30k$ iterations and decay to $5e^{-6}$ over $40k$ iterations. Together with negative samples, each image in the VisDial v1.0 dataset can generate 20 samples for the NSP task. Since these samples are fairly correlated and following the work of Murahari et al. (2019), we randomly sub-sample 8 out of these 20 during training. We use the validation set to decide when to stop training. The batch size is 32 in all our experiments, and different from the work of Murahari et al. (2019) and Lu et al. (2019), we do not pretrain our model on other vision-language datasets.

[4]https://github.com/vmurahari3/visdial-bert

5 Results

5.1 Quantitative Results

We compare the results of our model with the results of the following previously published models obtained on the VisDial v1.0 dataset: LF (Das et al., 2017), HRE (Das et al., 2017), MN (Das et al., 2017), CorefNMN (Kottur et al., 2018), FGA (Schwartz et al., 2019), RVA (Niu et al., 2019), HACAN (Yang et al., 2019), Synergistic (Guo et al., 2019), DAN (Kang et al., 2019), Dual VD (Jiang et al., 2020), and CAG (Guo et al., 2020). To make a fair and transparent comparison, we do not compare our models with models which were pretrained on other vision-language datasets before finetuning them on the VisDial v1.0 dataset, all the more because the vision-language datasets used in the pretraining overlap with the testset of Visdial v1.0. Also for those models, such as FGA, for which the authors also provide results of ensemble models, we only consider the results of their single model.

Results on VisDial v1.0 testset. As presented in Table 1, our best model (baseline model with both soft constraints) significantly outperforms all the previous published models and reach new state-of-the-art performance on MRR, R@1, R@5 and R@10. Specifically, compared to the best performance of previous models, our best model improves around 2.3% on MRR, 1.78% on R@1, 2.13% on R@5 and 2.35% on R@10. The mean rank of our proposed model is also better than all previous models, although the difference is relative small around 0.71. We also tested all our models on the VisDial v1.0 development set, and the results are presented in appendix A.3.

Ablation study. To further study the effectiveness of the two soft constraints, we perform an abla-

Model	MRR ↑	R@1 ↑	R@5 ↑	R@10 ↑	Mean ↓
LF (Das et al., 2017)	55.42	40.95	72.45	82.83	5.95
HRE (Das et al., 2017)	54.16	39.93	70.45	81.50	6.41
MN (Das et al., 2017)	55.49	40.98	72.30	83.30	5.92
CorefNMN (Kottur et al., 2018)	61.50	47.55	78.10	88.80	4.51
FGA (Schwartz et al., 2019)	63.70	49.58	80.98	88.55	4.51
RVA (Niu et al., 2019)	63.03	49.03	80.40	89.83	4.18
HACAN (Yang et al., 2019)	64.22	50.88	80.63	89.45	4.20
Synergistic (Guo et al., 2019)	62.20	47.90	80.43	89.95	4.17
DAN (Kang et al., 2019)	63.20	49.63	79.75	89.35	4.30
Dual VD (Jiang et al., 2020)	63.23	49.25	80.23	89.70	4.11
CAG (Guo et al., 2020)	63.49	49.85	80.63	90.15	4.11
Baseline Model	62.13	47.38	80.40	90.17	4.09
Model + POS Embedding Only	64.20	48.10	81.22	90.20	3.98
Model + POS Loss only	64.78	49.88	82.40	90.85	3.86
Model + C1	65.44	51.20	83.38	92.03	3.64
Model + C2	66.14	51.97	83.63	91.55	3.60
Model + C1 + C2	**66.53**	**52.63**	**84.13**	**92.50**	**3.40**

Table 1: Results of the visual dialog models on the VisDial v1.0 test set. C1 and C2 refer to the POS constraint and nearest preference constraint, respectively.(↑: the higher the better; ↓: the lower the better)

tion study on the VisDial v1.0 dataset with four different models: (1) Baseline model; (2) Model with the POS soft constraint (Model + C1); (3) Model with the nearest preference constraint (Model + C2); (4) Model with both the POS and nearest preference constraints (Model + C1 + C2). Moreover, We conduct an ablation study for the two aspects (POS embedding and POS prediction loss) in POS constraint The results are presented in Table 1. The baseline model obtains the following results in terms of MRR (62.13%), R@1 (47.38%), R@5 (80.40%) and R@10 (90.17%). Models with only POS embedding and only POS prediction loss have better performance than the baseline model. Further combining both leads to our first POS constraint, which improve the performance across all evaluation metrics (3.31% for MRR, 3.82% for R@1, 2.98% for R@5, 1.86% for R@10 and 0.55 for MRR). Similarly, only considering the nearest preference constraint leads to better performance on all evaluation metrics. The last row of Table 1 illustrates that the proposed soft constraints jointly lead to better results. We also study the changes of attention distribution of the model with and without our proposed constraints. Figure 3 shows that the nearest constraint can enhance the local connection in dialog, and the POS constraint is able to make pronouns focus more on nouns. These results indicate the effectiveness of adding linguistic constraints to a neural network. Integrating linguistic knowledge in a transformer neural architecture effectively improves the model's performance in the visual dialog task.

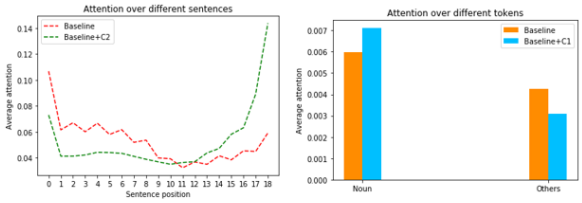

Figure 3: Left: Attention distribution of question over dialog history (Baseline Model+C2). Right: Attention distribution of pronoun over nouns and other words. (Baseline Model+C1

5.2 Coreference Analysis

As we do not have the access to the ground truth of the VisDial v1.0 test set, to further analyze the effectiveness of the proposed models, we create three small datasets, which each consists of samples with 2, 4, and 6 coreferences, respectively, from the Visdial v1.0 validation set. As there is no coreference annotation in the Visdial v1.0 dataset, we assume that each third-person pronoun (he, she, they, him, her, them) and possessive pronouns (its, his, her, their) has one coreference in the dialog.[5]

[5]Note that the dialog is about the objects in an image.

Results tested on validation set with 2 coreferences (237 samples)					
Model	MRR ↑	R@1 ↑	R@5 ↑	R@10 ↑	Mean ↓
Baseline Model	65.06	51.38	82.42	91.79	3.77
Model + C1	67.44	54.04	**84.54**	92.58	3.38
Model + C2	67.19	54.03	84.30	92.08	3.55
Model + C1 + C2	**67.61**	**54.45**	**84.54**	**92.87**	**3.34**
Results tested on validation set with 4 coreferences. (138 samples)					
Model	MRR ↑	R@1 ↑	R@5 ↑	R@10 ↑	Mean ↓
Baseline Model	62.94	48.93	80.78	90.28	3.99
Model + C1	65.77	51.99	83.49	**93.00**	3.48
Model + C2	65.94	52.07	83.36	92.00	3.52
Model + C1 + C2	**66.53**	**52.78**	**84.21**	92.07	**3.46**
Results tested on validation set with 6 coreferences. (58 samples)					
Model	MRR ↑	R@1 ↑	R@5 ↑	R@10 ↑	Mean ↓
Baseline Model	60.02	43.99	80.80	90.40	4.06
Model + C1	62.44	46.40	**83.86**	92.66	3.61
Model + C2	63.53	47.09	83.33	92.28	3.59
Model + C1 + C2	**63.66**	**48.14**	83.33	**92.67**	**3.42**
Results tested on validation set with coreferences. (1227 samples)					
Model	MRR ↑	R@1 ↑	R@5 ↑	R@10 ↑	Mean ↓
Baseline Model	64.62	50.73	82.40	91.42	3.73
Model + C1	**66.29**	52.41	84.17	92.08	3.62
Model + C2	**66.29**	52.54	83.46	91.25	3.59
Model + C1 + C2	65.97	**53.27**	**83.47**	**92.41**	**3.48**

Table 2: Results of the visual dialog models obtained on three small datasets each with a different number of pronoun coreferences that were collected from the VisDial v1.0 validation set. C1 and C2 refer to the POS constraint and nearest preference constraint, respectively.

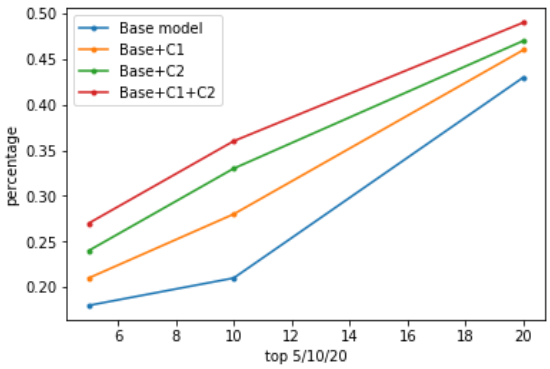

Figure 4: The percentage of the correct antecedent is within top 5/10/20 pronoun's attention distribution.

To create these subsets, we do not take "it" into consideration, as many of the occurrences of "it" [6] do not have a corefering expression, for example, "is it daytime ?". We test our models and the baseline model using these four small datasets, and the results are illustrated in Table 2. Comparing the results in Table 2 with that in Table 1, in some cases the performance in Table 2 is better, which means that the difficulty of these sampled data do not higher than that of the test set. One clear observation is that in almost all cases the model with two soft constraints has the best performance in terms of MRR, R@1, R@5, R@10 and Mean in all three datasets. Another observation is that integrating either the POS constraint or the nearest preference constraint improves the performance across all evaluation metrics in all four datasets, which again shows the effectiveness of our proposed linguistically inspired soft constraints. We can also see the trend that the models' performance is worse when the dialog has more coreferences, which is reasonable and consistent with our previous assumption that the existence of coreferences makes this task more difficult. To further study

[6] "It" has many other functions apart from being a pronoun, such as "empty" subject or object, not referring to anything in particular, to introduce or anticipate the subject or object of a sentence, use in cleft or in passive voice sentences, etc.

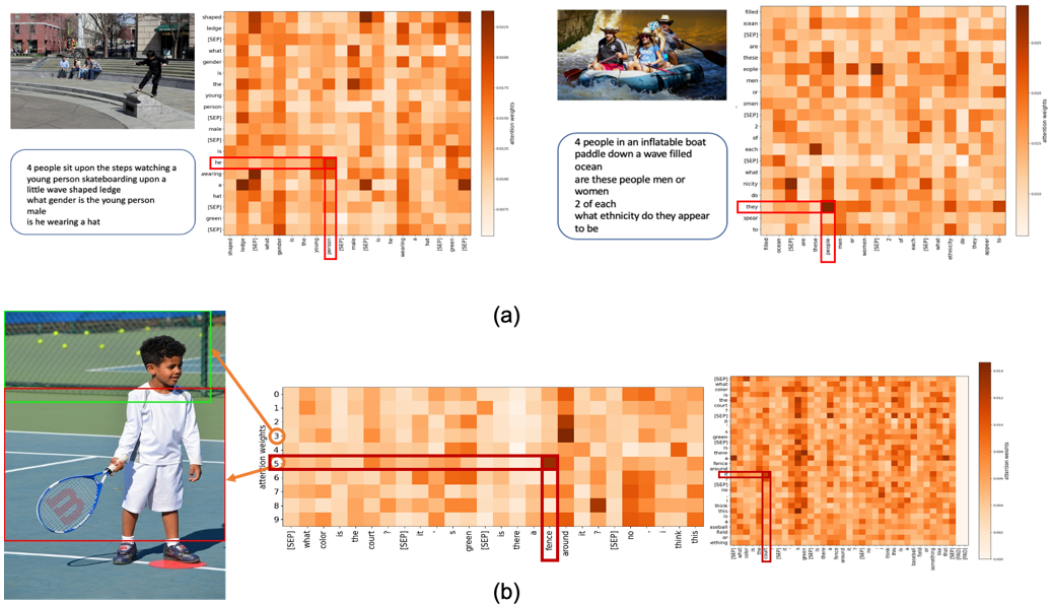

Figure 5: Attention score map of our best model (Model C1+C2). (a) Two correct cases: The attention map shows that the pronoun correctly attends to its antecedent. (b) An incorrect case: Although the coreference is correctly resolved, the word "fence" does not attend to the correct region in the image, which results in selecting the wrong answer.

the effectiveness of our proposed constraints, we manually annotated coreferences in a subset of the validation set and tested all our ablation models on this dataset. The results are presented in Figure 4, which shows that both the proposed constrains can improve the model's ability of resolving coreferences.

5.3 Qualitative Analysis

We visualize the attention scores of the model to investigate whether or not our best model can resolve pronoun coreferences in the dialog. We expect that the pronoun will attend the most to its antecedent, if the model correctly resolves the pronoun referent. Figure 5(a) presents the results of two samples taken from the VisDial v1.0 test set. Note that we only show the attention scores in the window of size 20 around the pronoun and its possible antecedent. The attention score maps illustrate that in these examples the attention weights between pronoun and its antecedent are significantly larger than those between the pronoun and other words, which indicates that the model can implicitly resolve the coreferences correctly.

Error Analysis Figure 5(b) presents a negative example. When facing the question "Is there a fence around it ?", the model answers "no, i think..." instead of the correct answer "yes". As shown in the attention map within language sequences, the pronoun "it" does attend to its antecedent "court", implying that the model successfully resolves the corefering noun. However, looking at the cross-modal attention, the word "fence" incorrectly refers to region 5 (court) in the image, which results in selecting a wrong answer. This negative example indicates that enhancing the model's ability of correct visual grounding is a meaningful future work.

6 Conclusion

In this paper, we have built a multi-layer transformer model for the visual dialog task. Based on linguistic knowledge and human dialog discourse patterns, we have proposed two soft constraints that effectively improve the model's performance by enhancing its ability of implicitly resolving pronoun coreferences. We have used the VisDial v1.0 dataset to evaluate our model. Our model obtains new state-of-the-art performance in correctly answering the dialog questions when compared to existing models without pretraining on other vision-language datasets. Our coreference and qualitative analysis further supports the proposed soft constraints.

Acknowledgement

This research received funding from the Flemish Government (AI Research Program).

References

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

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. 2018b. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3674–3683.

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

Dennis Connolly, John D Burger, and David S Day. 1997. A machine learning approach to anaphoric reference. In *New Methods in Language Processing*, pages 133–144.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technologies, Volume 1 (Long and Short Papers))*.

Quynh Ngoc Thi Do, Steven Bethard, and Marie Francine Moens. 2015. Adapting coreference resolution for narrative processing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2262–2267.

Greg Durrett and Dan Klein. 2013. Easy victories and uphill battles in coreference resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1971–1982.

Zhe Gan, Yu Cheng, Ahmed El Kholy, Linjie Li, Jingjing Liu, and Jianfeng Gao. 2019. Multi-step reasoning via recurrent dual attention for visual dialog. *"Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Dalu Guo, Chang Xu, and Dacheng Tao. 2019. Image-question-answer synergistic network for visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10434–10443.

Dan Guo, Hui Wang, Hanwang Zhang, Zheng-Jun Zha, and Meng Wang. 2020. Iterative context-aware graph inference for visual dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10055–10064.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Xiaoze Jiang, Jing Yu, Zengchang Qin, Yingying Zhuang, Xingxing Zhang, Yue Hu, and Qi Wu. 2020. Dualvd: An adaptive dual encoding model for deep visual understanding in visual dialogue. In *The Association for the Advancement of Artificial Intelligence (AAAI)*, volume 1, page 5.

Mandar Joshi, Omer Levy, Daniel S Weld, and Luke Zettlemoyer. 2019. Bert for coreference resolution: Baselines and analysis. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Gi-Cheon Kang, Jaeseo Lim, and Byoung-Tak Zhang. 2019. Dual attention networks for visual reference resolution in visual dialog. *The 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *ArXiv Preprint ArXiv:1412.6980*.

Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2018. Visual coreference resolution in visual dialog using neural module networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 153–169.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *ArXiv Preprint ArXiv:1908.03557*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 13–23.

Jiasen Lu, Anitha Kannan, Jianwei Yang, Devi Parikh, and Dhruv Batra. 2017. Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model. In *Advances in Neural Information Processing Systems (NIPS)*, pages 314–324.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank.

Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. 2019. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. *ArXiv Preprint ArXiv:1912.02379*.

Yulei Niu, Hanwang Zhang, Manli Zhang, Jianhong Zhang, Zhiwu Lu, and Ji-Rong Wen. 2019. Recursive visual attention in visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6679–6688.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NIPS)*, pages 8026–8037.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 1–40.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 91–99.

Idan Schwartz, Seunghak Yu, Tamir Hazan, and Alexander G Schwing. 2019. Factor graph attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2039–2048.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations. *The International Conference on Learning Representations (ICLR)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, pages 5998–6008.

Aparna Nurani Venkitasubramanian, Tinne Tuytelaars, and Marie-Francine Moens. 2017. Entity linking across vision and language. *Multimedia Tools and Applications*, 76(21):22599–22622.

Ellen M Voorhees. 1999. The trec-8 question answering track report. In *Trec*, volume 99, pages 77–82.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the gap: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics (TACL)*, 6:605–617.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *ArXiv Preprint ArXiv:1609.08144*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning (ICML)*, pages 2048–2057.

Tianhao Yang, Zheng-Jun Zha, and Hanwang Zhang. 2019. Making history matter: History-advantage sequence training for visual dialog. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2561–2569.

A Appendix

A 1

This Appendix shows the computation within a transformer encoder layer. Given an input sequence $H_0 = [e_0, e_1, \cdots, e_n]$, the transformer encodes it into different levels of contextual representations using a multi-head self-attention mechanism:

$$Q = H_{t-1}W_t^Q, K = H_{t-1}W_t^K, V = H_{t-1}W_t^V \tag{1}$$

$$AttentionHead_i(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d}})V \tag{2}$$

$$MultiHeadSelfAttention(H^t) = Concat(head_1, \cdots, head_k)W^O \tag{3}$$

where t is the layer index, $W_t^Q, W_t^K, W_t^V, W_t^O$ are learnable projection matrices, which map the hidden state h_t to the query q vector, key k vector, value v vector and output vector, and H_t denote the learned contextual representations at layer t. d is the size of hidden state.

A 2

The below formulations express the derivation of getting equation 8. Here, we use p to denote the pos in equation 11.

$$\begin{aligned} PE_1 \cdot PE_2 &= \frac{1}{k^2}\sum_{i=0}^{\frac{d}{2}-1} sin(w_i p_1) \cdot sin(w_i p_2) \\ &\quad + cos(w_i p_1) \cdot cos(w_i p_2) \\ &= \frac{1}{k^2}\sum_{i=0}^{\frac{d}{2}-1} cos(w_i(p_1 - p_2)) \\ &= \frac{1}{k^2}\sum_{i=0}^{\frac{d}{2}-1} cos(w_i \Delta p) \end{aligned}$$

A 3

Table 1 in this section shows the results of the four models on the VisDial v1.0 development set. Similar to the results obtained by testing on the test set, the model with both constraints (Model + C1 + C2) has the best performance across all evaluation metrics, and adding any one of the proposed soft constraint improves the performance of the baseline model, which indicates the effectiveness of our approach also during training.

Model	MRR ↑	R@1 ↑	R@5 ↑	R@10 ↑	Mean ↓
Baseline Model	64.80	50.56	82.64	91.02	3.85
Baseline Model + C1	68.59	55.37	84.38	92.29	3.28
Baseline Model + C2	68.32	55.25	84.49	92.25	3.32
Baseline Model + C1 + C2	**69.49**	**56.46**	**85.33**	**93.37**	**3.19**

Table 1: Results of the visual dialog models on the VisDial v1.0 development set. C1 and C2 denote our proposed soft constraints: POS constraint and nearest preference constraint, respectively.