

# Visual Dialogue without Vision or Dialogue

Daniela Massiceti\* Puneet K. Dokania\* N. Siddharth\* Philip H.S. Torr

University of Oxford

{daniela, puneet, nsid, phst}@robots.ox.ac.uk

## Abstract

We characterise some of the quirks and shortcomings in the exploration of visual dialogue (VD)—a sequential question-answering task where the questions and corresponding answers are related through given visual stimuli. To do so, we develop an embarrassingly simple method based on canonical correlation analysis (CCA) that, on the standard dataset, achieves near state-of-the-art performance for some standard metric. In direct contrast to current complex and over-parametrised architectures that are both compute and time intensive, our method *ignores the visual stimuli, ignores the sequencing of dialogue, does not need gradients*, uses *off-the-shelf* feature extractors, has at least an *order of magnitude fewer parameters*, and learns in *practically no time*. We argue that these results are indicative of issues in current approaches to visual dialogue relating particularly to implicit dataset biases, under-constrained task objectives, and over-constrained evaluation metrics, and consequently, discuss some avenues to ameliorate these issues.

## 1 Introduction

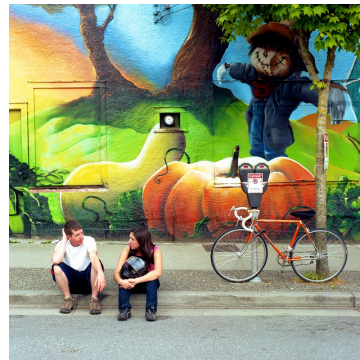
Recent years have seen a great deal of interest in *conversational AI*, enabling natural language interaction between humans and machines, early pioneering efforts for which include ELIZA (Weizenbaum, 1966) and SHRDLU (Winograd, 1971). This resurgence of interest builds on the ubiquitous successes of neural-network-based approaches in the last decade, particularly in the perceptual domains of vision and language.

A particularly thriving sub-area of interest in conversational AI is that of *visually grounded* dialogue, termed visual dialogue (VD), involving an AI agent conversing with a human about visual content (Das et al., 2017a,b; Massiceti et al., 2018). Specifically, it involves answering questions about an image, given some dialogue history—a fragment of previous questions and answers. Typical approaches for learning to do VD, as is standard practice in machine learning (ML), involves defining an objective to achieve, procuring data with which to learn, and establishing a measure of success at the stated objective.

The objective for VD is reasonably clear at first glance—answer in sequence, a set of questions about an image. The primary choice of dataset, *VisDial* (Das et al., 2017a), addresses precisely this criterion, involving a large set of images, each paired with a dialogue—a set of question-answer pairs—collected by pairs of human annotators playing a game to understand an image through dialogue. And finally, evaluation measures on the objective are typically defined through some perceived value of a human-derived “ground-truth” answer in the system.

However, as we will demonstrate, certain quirks in the choices of the above factors, can lead to unintentional behaviour (c.f. Figure 1<sup>2</sup>), which leverages implicit biases in data and methods, to potentially misdirect progress from the desired objectives. Intriguingly, we find that in contrast to

**Caption:** A man and a woman sit on the street in front of a large mural painting.



Question	Answer
How old is the baby?	About 2 years old
What color is the remote?	White
Where is the train?	On the road
How many cows are there?	Three

**Figure 1:** Failures in *visual* dialogue. Visually-unrelated questions, and their visually-unrelated plausible answers<sup>1</sup>.

\*Equal Contribution

<sup>2</sup>From online demos of SOTA models—*VisDial* (Das et al., 2017a) and *FlipDial* (Massiceti et al., 2018).

state-of-the-art (SOTA) approaches that employ complex neural-network architectures using complicated training schemes over millions of parameters and taking many hours of time and expensive GPU compute resources, a simple canonical correlation analysis (CCA)-based method only uses standard off-the-shelf feature extractors, avoids computing gradients, involves a few hundred thousand parameters and requires just a few seconds on a CPU to achieve comparable performance—all *without requiring the image or prior dialogue!*

## 2 (Multi-View) CCA for VD

We begin with a brief preliminary for CCA (Hotelling, 1936) and its multi-view extension (Kettenring, 1971). In (standard 2-view) CCA, given access to paired observations  $\{\mathbf{x}_1 \in \mathbb{R}^{n_1 \times 1}, \mathbf{x}_2 \in \mathbb{R}^{n_2 \times 1}\}$ , the objective is to jointly learn projection matrices  $W_1 \in \mathbb{R}^{n_1 \times p}$  and  $W_2 \in \mathbb{R}^{n_2 \times p}$  where  $p \leq \min(n_1, n_2)$ , that maximise the correlation between the projections, formally  $\text{corr}(W_1^\top \mathbf{x}_1, W_2^\top \mathbf{x}_2)$ .

Multi-view CCA, a generalisation of CCA, extends this to associated data across  $m$  domains, learning projections  $W_i \in \mathbb{R}^{n_i \times p}$ ,  $i \in \{1, \dots, m\}$ . Kettenring (1971) shows that  $W_i$  can be learnt by minimising the Frobenius norm between each pair of views, with additional constraints over the projection matrices (Hardoon et al., 2004). Optimising the multi-view CCA objective then reduces to solving a generalized eigenvalue decomposition problem,  $Av = \lambda Bv$ , where  $A$  and  $B$  are derived from the inter- and intra-view correlation matrices (c.f. Appendix A) (Bach and Jordan, 2002).

Projection matrices  $W_i$  are extracted from corresponding rows (for view  $i$ ) and the top  $p$  columns of the (eigenvalue sorted) eigenvector matrix corresponding to this eigen-decomposition. A sample  $\mathbf{x}_i$  from view  $i$  is then embedded as  $\phi_q(\mathbf{x}_i, W_i) = (W_i D_p^q)^\top \mathbf{x}_i$ , where  $D_p^q = \text{diag}(\lambda_1^q, \dots, \lambda_p^q)$  and  $\lambda_1 \geq \dots \geq \lambda_p$  are the eigenvalues. A scaling,  $q \in \mathbb{R}$ , controls the extent of eigenvalue weighting, reducing to the standard objective at  $q = 0^3$ . With this simple objective, one can tackle a variety of tasks at test time—ranking and retrieval across all possible combinations of multiple views—where the cosine similarity between (centred) embedding vectors captures correlation.

For VD, given a dataset of images  $I$  and associated question-answer ( $Q$ - $A$ ) pairs, joint embeddings between question and answer (and optionally, the image) are learnt, with projection matrices  $W_Q$ ,  $W_A$ , (and  $W_I$ ), as appropriate. At test time, correlations can be computed between any, and all, combinations of inputs, helping measure suitability against the desired response.

## 3 Experimental Analyses

In order to employ CCA for VD, we begin by transforming the input images  $I$ , questions  $Q$ , and answers  $A$ , into lower-dimensional feature spaces. For the images, we employ the standard pre-trained ResNet34 (He et al., 2016) architecture, extracting a 512-dimensional feature—the output of the *avg pool* layer after *conv5*. For the questions and answers, we employ the FastText (Bojanowski et al., 2017) network to extract 300-dimensional embeddings for each of the words. We then simply average the embeddings (Arora et al., 2017) for the words, with suitable padding or truncation (up to a maximum of 16 words), to obtain a 300-dimensional embedding for the question or answer.

We then set the hyper-parameters for the CCA objective as  $p = 300$ ,  $q = 1$ , based off of a simple grid search over feasible values, such that we learn a 300-dimensional embedding space that captures the correlations between the relevant domains. It is important to note that the SOTA approaches (Das et al., 2017a,b; Massiceti et al., 2018) also employ pre-trained feature extractors—the crucial difference between approaches is the complexities in modelling and computation *on top of* such feature extraction, as starkly indicated in Table 1.

Table 1: CCA vs. SOTA: number of learnable parameters and training time.

Model	#Params	Train time (s)
HCIAE-G-DIS	$2.12 \times 10^7$	–
VisDial	$2.42 \times 10^7$	–
FlipDial	$1.70 \times 10^7$	$2.0 \times 10^5$
CCA (A-Q)	$1.80 \times 10^5$	2.0
<b>Factor (<math>\approx</math>)</b>	<b>90</b>	<b><math>10^5</math></b>

We then learn two joint embeddings—between just the answers and questions, denoted A-Q, and between the answers, questions, and images, denoted A-QI. Note that the answer is always present, since the stipulated task in VD is to answer a given question. The first allows us to explore the utility (or lack thereof) of the image in performing the VD task. The second serves as a useful indicator of how unique any question-image pairing is, in how it affects the ability to answer—performance closer to that of A-Q indicating fewer unique pairings. Also, when embedding all three of A, Q, and I, at test time, we only employ Q to compute a match against a potential answer.

<sup>3</sup>There are cases where values of  $q > 0$  have been shown to give better performance (Gong et al., 2014).

Having now learnt an embedding, we evaluate our performance using the standard ranking measure employed for the *VisDial* dataset. Here, for a given image and an associated question, the dataset provides a set of 100 candidate answers, which includes the human-derived “ground-truth” answer. The task then, is to rank each of the 100 candidates, and observe the rank awarded to the “ground-truth” answer. In our case, we rank on correlation, computed as the cosine distance between centered embeddings between the question and a candidate answer. Then, for all the answers we compute the mean rank (MR), mean reciprocal rank (MRR) (inverse harmonic mean of rank), and recall at top 1, 5, and 10 candidates—measuring how often the “ground-truth” answer ranked within that range.

The results, in Table 2, show that the simple CCA approach achieves comparable performance on the mean rank (MR) metric using the A-Q model that *doesn’t use the image or dialogue sequence!* This solidifies the impression, from Figure 1, that there exist implicit correlations between just the questions and answers in the data, that can be leveraged to perform “well” on a task that simply requires matching “ground-truth” answers. Our experiments indicate that for the given dataset and task, one need not employ anything more complicated than an exceedingly simple method such as CCA on pre-trained feature extractors, to obtain plausible results.

Moreover, another factor that needs to be considered, is that the evaluation metric itself, through the chosen task of candidate-answer ranking, can be insufficient to draw any *actual* conclusions about how well questions were answered. To see this, consider Figure 2, where we deliberately pick examples that rank the “ground-truth” answer poorly despite CCA’s top-ranked answers all being plausible alternatives. This clearly illustrates the limitations imposed by assuming a single “ground-truth” answer in capturing the breadth of correct answers.

To truly judge the validity of the top-ranked answers, regardless of “ground-truth” would require thorough human-subject evaluation. However, as a cheaper, but heuristic alternative, we quantify the validity of the top answers, in relation to the “ground truth”, using the correlations themselves. For any given question and candidate set of answers, we cluster the answers based on an automatic binary thresholding (ISODATA (Ridler and Calvard, 1978), 5 bins) of the correlation with the given question. We then compute the following two statistics based on the threshold i) the average variance of the correlations in the lower-ranked split, and ii) the fraction of questions that have correlation with “ground truth” answer higher than the threshold. The intention being that (i) quantifies how closely clustered the top answers are, and (ii) quantifies how often the “ground-truth” answer is in this cluster. Low values for the former, and high values for the latter would indicate that there exists an equivalence class of answers, all relatively close to the ground-truth answer in terms of their ability to answer the question. Our analysis for the *VisDial* v0.9 dataset reveals values of (i) 0.023 and (ii) 96.9%, supporting our claims that CCA recovers plausible answers.

We note that the *VisDial* dataset was recently updated to version 1.0, where the curators try to ameliorate some of the issues with the single-“ground-truth” answer approach. They incorporate a human-agreement scores for candidate answers, and introduce a modified evaluation which weighs the predicted rankings by these scores. We include our performance on the (held-out) test set for *VisDial* v1.0 in the bottom row of Table 2. However, in making this change, the primary evaluation for this data has now become an explicit classification task on the candidate answers—requiring access, at train time, to all (100) candidates for every question-image pair (see Table 1, pg 8. Das et al., 2017a) and the evaluation results of the Visual Dialog Challenge 2018. For the stated goals of VD, this change can be construed as unsuitable as it falls into the category of redefining the problem to match a potentially unsuitable evaluation measure—how can one get better ranks in the candidate-answer-ranking task. For this reason, although there exist approaches that use the updated data, we do not report comparison to any of them.

Table 2: Results for SOTA vs. CCA on the *VisDial* dataset. CCA achieves comparable performance while ignoring both image and dialogue sequence.

	Model	MR	R@1	R@5	R@10	MRR
SOTA	HCLAE-G-DIS	14.23	44.35	65.28	71.55	0.5467
	CoAtt-GAN	14.43	46.10	65.69	71.74	0.5578
	HREA-QIH-G	16.79	42.28	62.33	68.17	0.5242
CCA	A-Q	16.21	16.77	44.86	58.06	0.3031
	A-QI (Q)	18.29	12.17	35.38	50.57	0.2427
	A-Q	17.08	15.95	40.10	55.10	0.2832
	A-QI (Q)	19.24	12.73	33.05	48.68	0.2393




Image	Question (Rank) GT Answer	CCA Top-3 (Rank) Answer
	What colour is the bear? ④ Floral white	① White and brown ② Brown and white ③ Brown, black & white
	Does she have long hair? ④ No	① No, it is short hair ② Short ③ No it's short
	Can you see any passengers? ④ Not really	① No ② Zero ③ No I can not
	Are there people not on bus? ② Few	① No people ② No, there are no people around ③ I don't see any people

Figure 2: Qualitative results for the A-Q model showing the *top-3* ranked answers for questions where the ground-truth answer is given a low rank—showing them to be perfectly feasible.

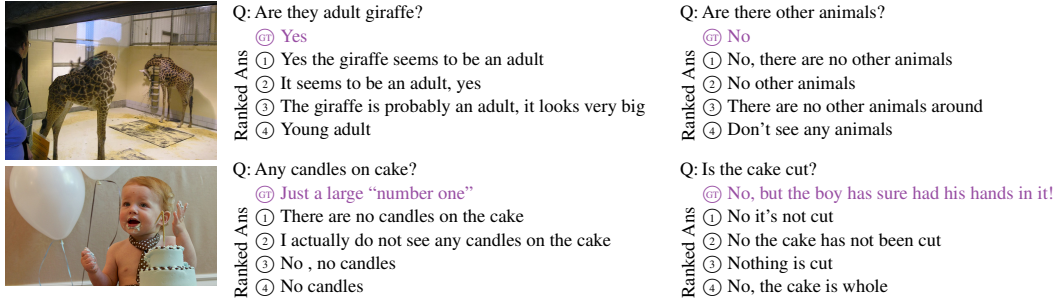


Figure 3: Example answers “generated” using the nearest-neighbours approach. For a given test question, a custom candidate set is constructed by choosing answers corresponding to the 100 closest (by correlation using A-Q) questions from the training data, and the best correlated answers to the given question returned.

Although standard evaluation for VD involves ranking the given candidate answers, there remains an issue of whether, given a question (relating to an image), the CCA approach really “answers” it. From one perspective, simply choosing from a given candidate set can seem a poor substitute for the ability to *generate* answers, in the vein of [Das et al. \(2017a\)](#); [Massiceti et al. \(2018\)](#). To address this, we construct a simple “generative” model using our learned projections between questions and answers (A-Q model, c.f. [Figure 3](#)). For a given question, we select the corresponding answers to the 100 nearest-neighbour questions using solely the train set and construct a custom candidate-answer set. We then compute their correlations with the given question, and sample the top-correlated answers as “generated” answers<sup>4</sup>. We also compute our heuristic for the validity of the top-ranked answers in relation to the “ground-truth” as before, with average variance (i) 0.018 and fraction correct (ii) 87.2%, indicating reliable performance.

## 4 Discussion

We use the surprising equivalence from § 3 as evidence of several issues with current approaches to VD. The biggest concern our evaluation, and a similar by [Anand et al. \(2018\)](#), reveals is that, for standard datasets in the community, *visually grounded* questions can be answered “well”, without referring to the visual stimuli. This reveals an unwanted bias in the data, whereby correlations between question-answer pairs can be exploited to provide reasonable answers to visually-grounded questions. Moreover, the dataset also includes an implicit bias that any given question *must* necessarily relate to a given image—as evidenced by visually-unrelated questions getting visually-unrelated, but plausible answers ([Figure 1](#)). A particularly concerning implication of this is that current approaches to visual dialogue ([Das et al., 2017a,b](#); [Massiceti et al., 2018](#)) may not actually be targeting the *intended* task.

Our simple CCA method also illustrates, that the standard evaluation used for VD has certain shortcomings. Principally, the use of “candidate” answers for each question, with a particular subset of them (1 in *VisDial* v0.9, and K-human-derived weighted choices in v1.0) are deemed to be the “ground-truth” answers. However, as we show in [Figure 2](#), such an evaluation can still be insufficient to capture the range of all plausible answers. The task of designing evaluations on the “match” of expected answers in for natural language, though, is fraught with difficulty, as one needs to account for a high degree of syntactic variability, with perhaps little semantic difference.

Responses to addressing the issues observed here, can take a variety of forms. For the objective itself, one could alternately evaluate the effectiveness with which the dialogue enables a downstream task, as explored by some ([Das et al., 2017b](#); [De Vries et al., 2017](#); [Khani et al., 2018](#); [Lazaridou et al., 2016](#)). Also, to address implicit biases in the dataset, one could adopt synthetic, or simulated, approaches, such as [Hermann et al. \(2017\)](#), to help control for undesirable factors. Fundamentally, the important concern here is to evaluate visual dialogue on it’s actual utility—conveying information *about the visual stimuli*—as opposed to surface-level measures of suitability.

And finally, we believe an important takeaway from our analyses is that it is highly effective to begin exploration with the simplest possible tools one has at one’s disposal. This is particularly apposite in the era of deep neural networks, where the prevailing attitude appears to be that it is preferable to start exploration with complicated methods that aren’t well understood, as opposed to older, perhaps even *less fashionable* methods that have the benefit of being rigorously understood. Also, as shown in [Table 1](#), choosing simpler methods can help minimise human effort and cost in terms of both compute and time, and crucially provide the means for cleaner insights into the problems being tackled.

<sup>4</sup>This only additionally requires disk-space to persist the training data—costing roughly \$0.25/GB



## References

- Ankesh Anand, Eugene Belilovsky, Kyle Kastner, Hugo Larochelle, and Aaron Courville. Blindfold baselines for embodied qa. *arXiv preprint arXiv:1811.05013*, 2018.
- Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.
- F. R Bach and M. I. Jordan. Kernel independent component analysis. *JMLR*, 2002.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 2017.
- Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. Visual Dialog. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2017a.
- Abhishek Das, Satwik Kottur, José M.F. Moura, Stefan Lee, and Dhruv Batra. Learning cooperative visual dialog agents with deep reinforcement learning. In *ICCV*, 2017b.
- Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron C Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *IJCV*, 2014.
- David R. Hardoon, Sandor R. Szedmak, and John R. Shawe-taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 2004.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- Karl Moritz Hermann, Felix Hill, Simon Green, Fumin Wang, Ryan Faulkner, Hubert Soyer, David Szepesvari, Wojciech Marian Czarnecki, Max Jaderberg, Denis Teplyashin, Marcus Wainwright, Chris Apps, Demis Hassabis, and Phil Blunsom. Grounded language learning in a simulated 3d world. *CoRR*, abs/1706.06551, 2017. URL <http://arxiv.org/abs/1706.06551>.
- Harold Hotelling. Relations between two sets of variates. *Biometrika*, 1936.
- J. R. Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 1971.
- Fereshte Khani, Noah D. Goodman, and Percy Liang. Planning, inference and pragmatics in sequential language games. *CoRR*, abs/1805.11774, 2018. URL <http://arxiv.org/abs/1805.11774>.
- Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. Towards multi-agent communication-based language learning. *CoRR*, abs/1605.07133, 2016. URL <http://arxiv.org/abs/1605.07133>.
- Jiasen Lu, Anitha Kannan, Jianwei Yang, Devi Parikh, and Dhruv Batra. Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model. In *Advances in Neural Information Processing Systems*, pages 314–324, 2017.
- Daniela Massiceti, N. Siddharth, Puneet K. Dokania, and Philip H.S. Torr. Flipdial: A generative model for two-way visual dialogue. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- T. W. Ridler and S. Calvard. Picture thresholding using an iterative selection method. *IEEE Transactions on Systems, Man, and Cybernetics*, 8(8):630–632, Aug 1978.
- Joseph Weizenbaum. ELIZA—A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 1966.
- Terry Winograd. Procedures as a representation for data in a computer program for understanding natural language. Technical report, Massachusetts Institute of Technology, Cambridge, MA, 1971.
- Qi Wu, Peng Wang, Chunhua Shen, Ian Reid, and Anton van den Hengel. Are you talking to me? reasoned visual dialog generation through adversarial learning. *arXiv preprint arXiv:1711.07613*, 2017.

## A Multi-view Canonical Correlation Analysis

Among several possible ways to formulate the CCA objective for multiple variables (Kettenring, 1971), we choose the Forbenius-norm-based objective as it provides better insights. Let us assume that there are  $m$  views and  $\mathbf{x}_i \in \mathbb{R}^{n_i}$  represents the observation from the  $i$ -th view. Then, the objective is to jointly learn projection matrices  $W_i \in \mathbb{R}^{n_i \times p}$  for all the  $m$  views such that the embeddings in the  $p(\leq n_i \forall i)$  dimensional space are maximally correlated. This is achieved by optimizing the following problem:

$$\begin{aligned} \min_{W_1, \dots, W_m} \sum_{i,j=1, i \neq j}^m \|W_i^\top \mathbf{x}_i - W_j^\top \mathbf{x}_j\|_F \\ \text{s.t. } W_i^\top C_{ii} W_i = I, w_i^k{}^\top C_{ij} w_j^l = 0, k, l = 1, \dots, p, k \neq l, \end{aligned} \quad (1)$$

where,  $w_i^k$  is the  $k$ -th column of  $W_i$  projection matrix. It turns out that optimizing (1) reduces to solving a generalized eigenvalue decomposition problem (Bach and Jordan, 2002):

$$Av = \lambda Bv, \quad \text{where } A = \begin{pmatrix} C_{11} & C_{12} & \cdots & C_{1m} \\ C_{21} & C_{22} & \cdots & C_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ C_{m1} & C_{m2} & \cdots & C_{mm} \end{pmatrix}, B = \begin{pmatrix} C_{11} & 0 & \cdots & 0 \\ 0 & C_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & C_{mm} \end{pmatrix}. \quad (2)$$

where,  $C_{ij}$  is the correlation matrix obtained using the observations from  $i$ -th and  $j$ -th views.