Visual Dialogue without Vision or Dialogue

Daniela Massiceti*, Puneet K. Dokania*, N. Siddharth*, Philip H.S. Torr
Department of Engineering Science, University of Oxford
{daniela, puneet, nsid, phst}@robots.ox.ac.uk

CCA for Visual Dialogue (VD)

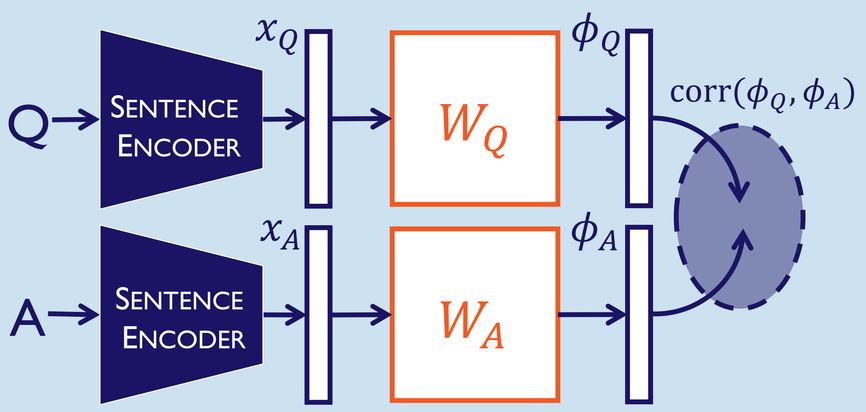
- **Visual Dialogue (VD)** involves answering a sequence of questions about an image
- We apply Canonical Correlation Analysis (CCA) to *just* questions and answers
- Our method:
  - Ignores visual stimulus & dialogue sequence,
  - Does not need gradients,
  - Uses off-the-shelf feature extractors,
  - Uses ~0.009% parameters of state-of-the-art models, and
  - Learns in a few (CPU) seconds.
- Surprisingly good performance highlights implicit dataset biases & quirks of evaluation metrics
- Need for better balanced visuo-linguistic datasets and evaluation protocols

What's ailing visual dialogue?

Plausible answers to visually-unrelated questions

** Results from online demos of SOTA models*

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



CCA vs. SOTA in ranking performance on VisDial

		Model	#params	Time (s)	MR	R@1	R@5	R@10	MRR
SOTA	v0.9	HCIAE-G-DIS [3]	2.12×10^7	-	14.23	44.35	65.28	71.55	0.5467
		CoAtt-GAN [4]	-	-	14.43	46.10	65.69	71.74	0.5578
		HREA-QIH-G [5]	2.42×10^7	-	16.69	42.28	62.33	68.17	0.5242
CCA	v0.9	A-Q	$\mathbf{1.80 \times 10^5}$	**2.0**	16.21	16.77	44.86	58.06	0.3031
		A-QI (Q)	3.33×10^5	3.0	18.29	12.17	35.38	50.57	0.2427
	v1.0	A-Q	$\mathbf{1.80 \times 10^5}$	**2.0**	17.08	15.95	40.10	55.10	0.2832
		A-QI (Q)	3.33×10^5	3.0	19.24	12.73	33.05	48.68	0.2393

Multi-view CCA [1,2]

- Given question $x_Q \in \mathbb{R}^{n_Q \times 1}$ and answer $x_A \in \mathbb{R}^{n_A \times 1}$, learn projections $W_Q \in \mathbb{R}^{n_Q \times p}$, $W_A \in \mathbb{R}^{n_A \times p}$ where $p \leq \min(n_Q, n_A)$ such that $\mathrm{corr}\left(W_Q^T x_Q, W_A^T x_A\right)$, is maximised

$$\left[\lambda_1 \cdots \lambda_p \cdots \lambda_{n_Q+n_A}\right], \begin{bmatrix} v_{1,1} & \cdots & v_{1,p} & \cdots & v_{1,n_Q+n_A} \\ \vdots & & \vdots & & \vdots \\ v_{n_Q,1} & \cdots & v_{n_Q,p} & \cdots & v_{n_Q,n_Q+n_A} \\ v_{n_Q+1,1} & \cdots & v_{n_Q+1,p} & \cdots & v_{n_Q+1,n_Q+n_A} \\ \vdots & & \vdots & & \vdots \\ v_{n_Q+n_A,1} & \cdots & v_{n_Q+n_A,p} & \cdots & v_{n_Q+n_A,n_Q+n_A} \end{bmatrix} = EVD\left[\begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}\begin{bmatrix} x \end{bmatrix} = \lambda \begin{bmatrix} C_{11} & 0 \\ 0 & C_{22} \end{bmatrix}\begin{bmatrix} x \end{bmatrix}\right]$$

(W_Q: rows $1 \dots p$; W_A: rows $n_Q+1 \dots n_Q+n_A$)

C_{11}, C_{22} and C_{12}, C_{21} are intra- and inter-view correlation matrices

- Projection $\phi\,(x_i, W_i) = \left(W_i D_p^k\right)^T x_i$ where $D_p^k = diag(\lambda_1^k, \dots, \lambda_p^k)$ and $\lambda_1 \geq \cdots \geq \lambda_p$ are eigenvalues and $\mathrm{k} \in \mathbb{R}$ is a scaling factor
- 2-view approach can be generalised to views $x_i \in \mathbb{R}^{n_i}$ and $W_i \in \mathbb{R}^{n_i \times p}, i \in \{1, \dots, m\}$

Experimental Analyses

- Represent answers & questions by average *FastText* vectors (300D), images by *ResNet* features (512D)
- **Train**: CCA to learn joint embeddings between:
  - Answers & questions (A-Q, 2-view CCA)
  - Answers, questions & images (A-QI, 3-view CCA)
- **Evaluate**: rank candidate answer set per question using embedding
- Near SOTA MR with ~0.009% parameters & seconds on CPU
- For given question & corresponding candidate answers
  - 96.9% ground-truth answer ranks $< T_c(\sigma^2 = 0.023)$
  - 87.2% ground-truth answer rank $< T_g(\sigma^2 = 0.018)$

  where T_c and T_G are ISODATA thresholds computed on *VisDial* candidates & candidates "generated" by CCA A-Q

Generating plausible answers with CCA

Recalling top-k answers to nearest-neighbour questions in train set

(Q) Are they adult giraffe? Yes (GT)	(Q) Are there other animals? No (GT)
① Yes the giraffe seem to be adult ② It seems to be adult, yes ③ The giraffe is probably an adults, it looks very big ④ Young adult	① No, there are no other animals ② No other animals ③ There are no other animals around ④ Don't see any animals

(Q) Any candles on the cake? Just a large "number one" (GT)	(Q) Is the cake cut? No, but the boy sure has had his hands in it! (GT)
① There are no candles on the cake ② I actually do not see any candles on the cake ③ No, no candles ④ No candles	① No it's not cut ② No the cake has not been cut ③ Nothing is cut ④ No, the cake is whole

Bad mean rank (MR) doesn't always mean bad answers

Top-ranked candidates are plausible, but rank assigned to ground-truth answer is high

Question Rank + GT Answer	CCA Top-3 Rank + Answer
What colour is the bear? ⑤⓪ Floral white	① White and brown ② Brown and white ③ Brown, black and white
Does she have long hair? ④① No	① No, it is short hair ② Short ③ No it's short
Can you see any passengers? ④⑧ Not really	① No ② Zero ③ No I cannot
Are there people not on bus? ②② Few	① No people ② No, there are no people around ③ I don't see any people

Conclusions

- It is possible to perform "well" without a visual stimulus
- Poor ranking performance doesn't always correspond to poor answers
- Assigning a single ground-truth answer is restrictive – *VisDial v1.0* ameliorates this with similarity scores for candidate answers
- Embedding space learned by CCA is useful for answer "generation"
- Simple methods like CCA should be used alongside deep approaches

[1] H. Hotelling. Relations between two sets of variates. *Biometrika*, 1936.
[2] J. R. Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 1971
[3] J. Lu, A. Kannan, J. Yang, D. Parikh, and D. Batra. Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model. In *NIPS*, 2017.
[4] Q. Wu, P. Wang, C. Shen, I. Reid, and A. van den Hengel. Are you talking to me? Reasoned visual dialog generation through adversarial learning. *arXiv*, 2017.
[5] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J.M.F. Moura, D. Parikh, and D. Batra. Visual Dialog. In *CVPR*, 2017.