Learning Conditioned Graph Structures for Interpretable Visual Question Answering

Will Norcliffe-Brown, Stathis Vafeias and Sarah Parisot

NIPS
Neural Information Processing Systems

aimbrain
simply smarter authentication

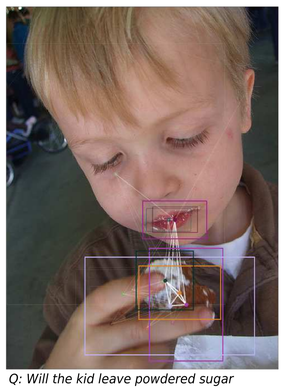
Introduction

Context

- VQA is a challenging topic requiring a combination of natural language and image processing techniques
- An effective VQA system requires modelling the interactions between objects in an image
- Previous approaches have either modelled objects as independent or use engineered graph structures

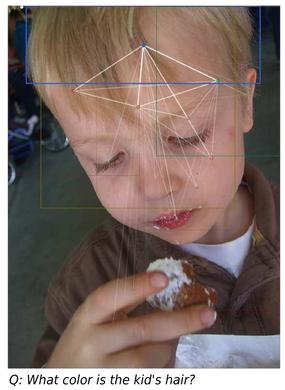
Contributions

- A novel graph-based approach for VQA
- Learn a graph representation of the image, conditioned on the question at hand, that identify the most relevant objects and relationships for answering a question
- Learned graph structures provide interpretability
- Promising results with minimal architecture tuning



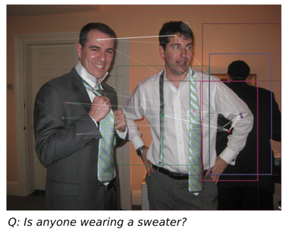
Q: Will the kid leave powdered sugar on his face?

Prediction: Yes



Q: What color is the kid's hair?

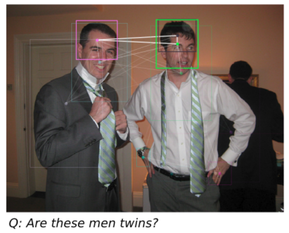
Prediction: Blonde



Q: Is anyone wearing a sweater?

Prediction: Yes

Answer: No

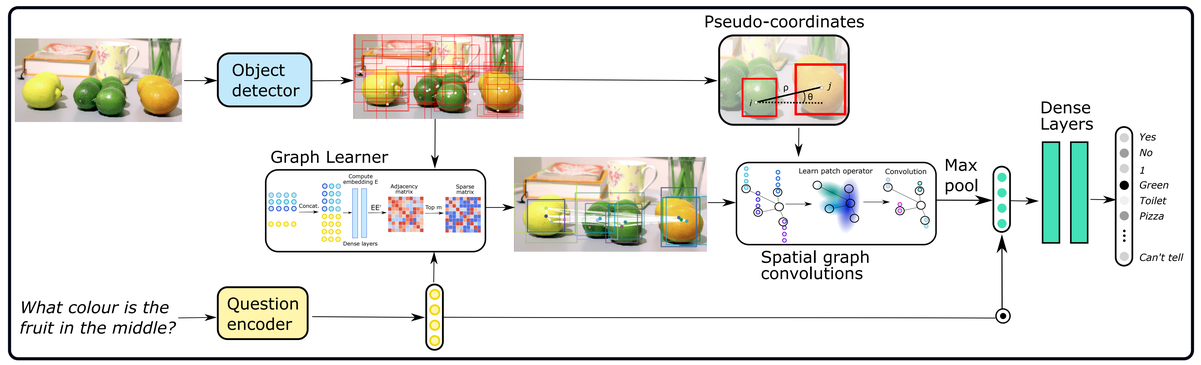


Q: Are these men twins?

Prediction: No

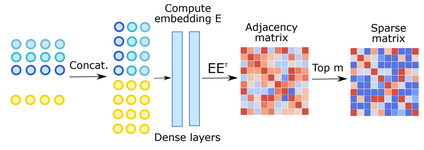
Answer: No

Method Overview



Graph Learner Module

- Define an undirected graph where nodes are object bounding boxes and edges are learned and conditioned on the question
- Compute a joint embedding from concatenated visual and question features: $\mathbf{e}_n = F([\mathbf{v}_n \| \mathbf{q}])$
- Compute an adjacency matrix as $A_{i,j} = \mathbf{e}_i^T \mathbf{e}_j$
- Sparsity imposed by defining each node's neighbourhood system $\mathcal{N}(i)$ as the nodes with the top m strongest edge weights.



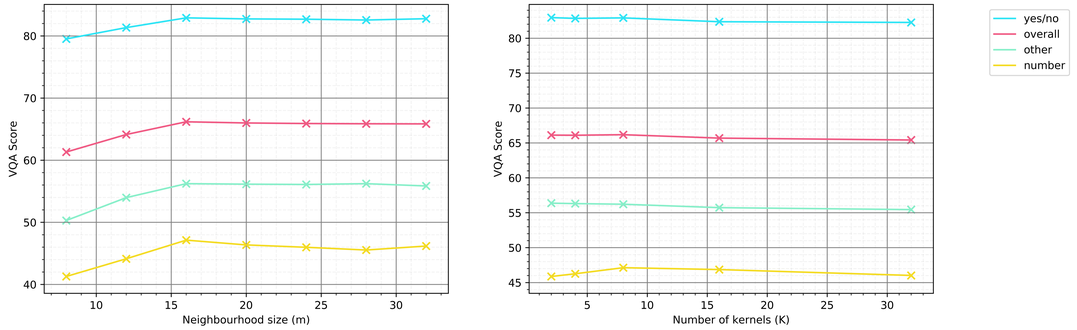
Spatial Graph Convolutions

- Learn object representations informed by graph neighbours
- A patch operator describes the influence of each node in $\mathcal{N}(i)$ as a mixture of Gaussian kernels[2]. Compute updated visual features as

$$\mathbf{f}_k(i) = \sum_{j \in \mathcal{N}(i)} \overbrace{w_k(\underbrace{\mathbf{u}(i,j)}_{\text{Bounding box Pseudo-coordinates}})}^{\text{Gaussian kernel with learnable weights}} \overbrace{\mathbf{v}_j}^{\text{Object/bounding box visual features}} \underbrace{\alpha_{ij}}_{\text{Graph edge weights}}$$

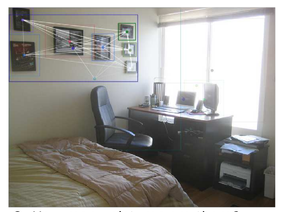
- Output at vertex i: $\mathbf{h}_i = \|_{k=1}^{K} \mathbf{G}_k \mathbf{f}_k(i)$ where $\mathbf{G}_k$ is a matrix of learnable weights.

Parameter Analysis



Results

Answer type	All	Y/N	Num.	Other
ReasonNet	64.61	78.86	41.98	57.39
Bottom-Up [1]	65.67	82.20	43.90	56.26
Counting module [3]	68.41	83.56	51.39	59.11
kNN graph	61.00	79.35	41.63	49.70
Attention	61.90	79.87	42.48	50.95
Ours	66.18	82.91	47.13	56.22



Q: How many pictures are there?

Prediction: 7

Answer: 7

Q: What color is the sleeping woman's bag?

Prediction: Black

Answer: Brown

Q: Are all these animals the same colour?

Prediction: Yes

Answer: Yes

Q: Is this healthy?

Prediction: Yes

Answer: Yes

References

[1] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering (2017).

[2] F. Monti, D. Boscaini, J. Masci, E. Rodolà, J. Svoboda, and M. M. Bronstein. Geometric deep learning on graphs and manifolds using mixture model cnns (2017).

[3] Y. Zhang, J. S. Hare, and A. Prügel-Bennett. Learning to count objects in natural images for visual question (2018).

Links

Code

github.com/aimbrain/vqa-project

Contact

- will.norcliffe@aimbrain.com
- stathis@aimbrain.com
- sarah.parisot@huawei.com