Improved Fusion of Visual and Language Representations by Dense Symmetric Co-Attention for Visual Question Answering

Duy-Kien Nguyen, Takayuki Okatani

CVPR 2018

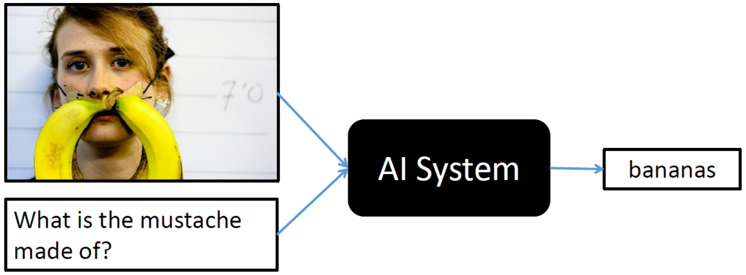
Reading group

Presented by: Sebastian Bujwid

Overview

Overview

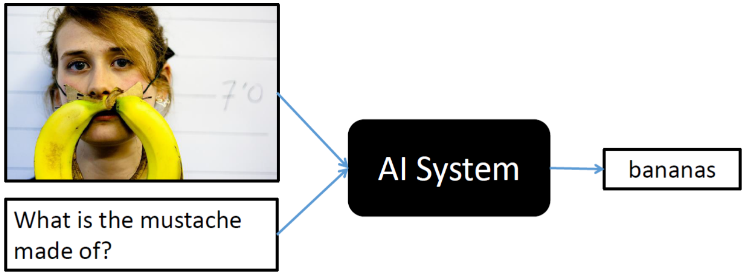
- Authors tackle the problem of Visual Question Answering (VQA)



  - Requires interaction between vision and language
  - Possible to evaluate - small number of possible correct answers

Overview

- Authors tackle the problem of Visual Question Answering (VQA)



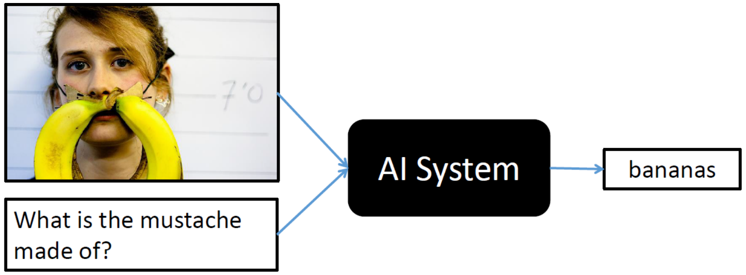
  - Requires interaction between vision and language
  - Possible to evaluate - small number of possible correct answers

- A model for dense, bi-directional interaction between two modalities (text, vision)

  - Model: ***dense co-attention network*** **(DCN)**

Overview

- Authors tackle the problem of Visual Question Answering (VQA)



  - Requires interaction between vision and language
  - Possible to evaluate - small number of possible correct answers

- A model for dense, bi-directional interaction between two modalities (text, vision)

  - Model: ***dense co-attention network*** **(DCN)**

- **State-of-the-art** results on VQA and VQA 2.0 datasets

Background

Soft attention mechanism

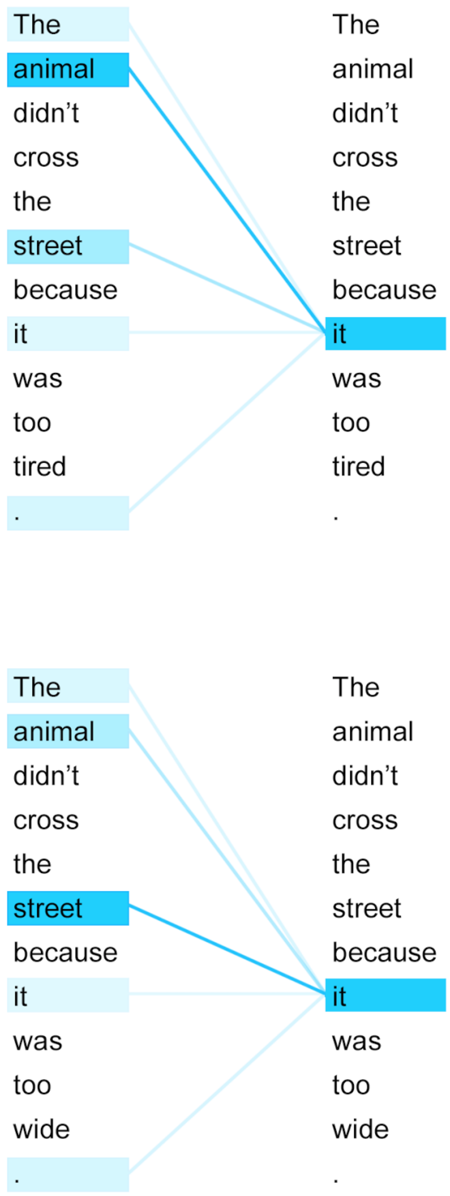
- $q \in \mathbb{R}^{d_q}$ - query
- $K \in \mathbb{R}^{L \times d_k}$ - keys
- $V \in \mathbb{R}^{L \times d_v}$ - values

$$[\alpha_1, \ldots, \alpha_L] = \mathrm{softmax}(f(q, K))$$

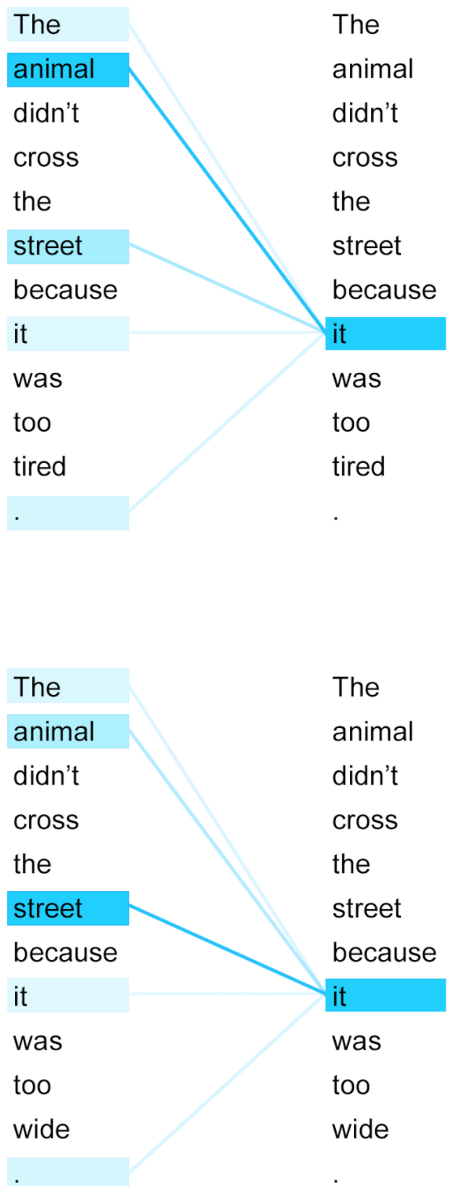
$$\mathrm{attention}(q, K, V) = \sum_{i=1}^{L} \alpha_i v_i$$

- $v_i \in V$
- $f : \mathbb{R}^{d_q} \times \mathbb{R}^{L \times d_k} \to \mathbb{R}^L$ - compatibility function

Attention maps - example



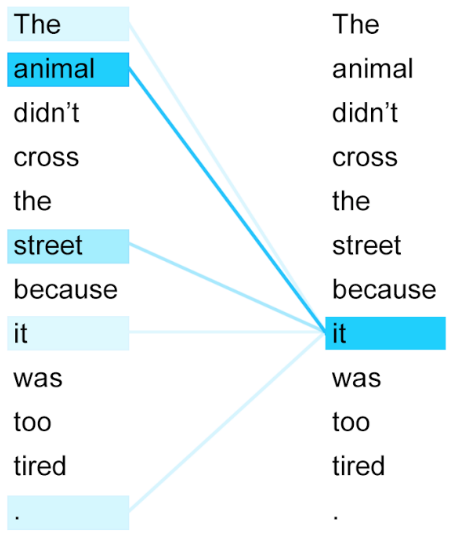
Attention maps - example

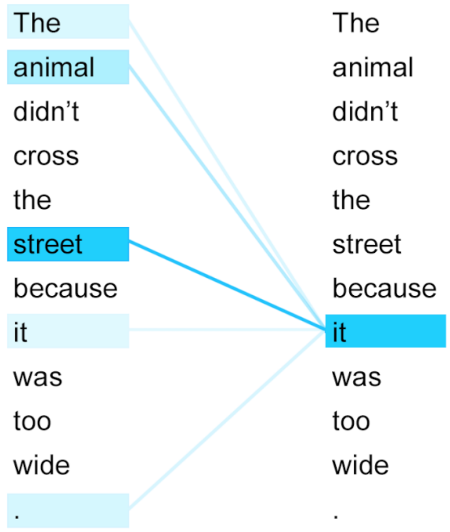


Why attention mechanism?

- Conditional representations
  - Meaning of a word in the context of a sentence
  - Meaning of an object in the context of a question

Attention maps - example

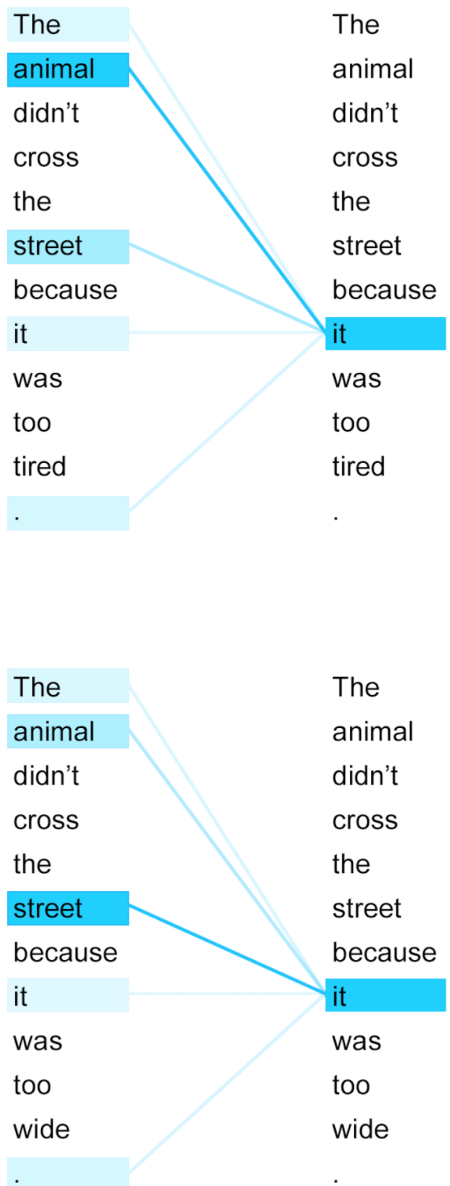




Why attention mechanism?

- Conditional representations
  - Meaning of a word in the context of a sentence
  - Meaning of an object in the context of a question
- Modeling long-term dependencies
  - $O(1)$ vs. $O(N)$ for RNNs

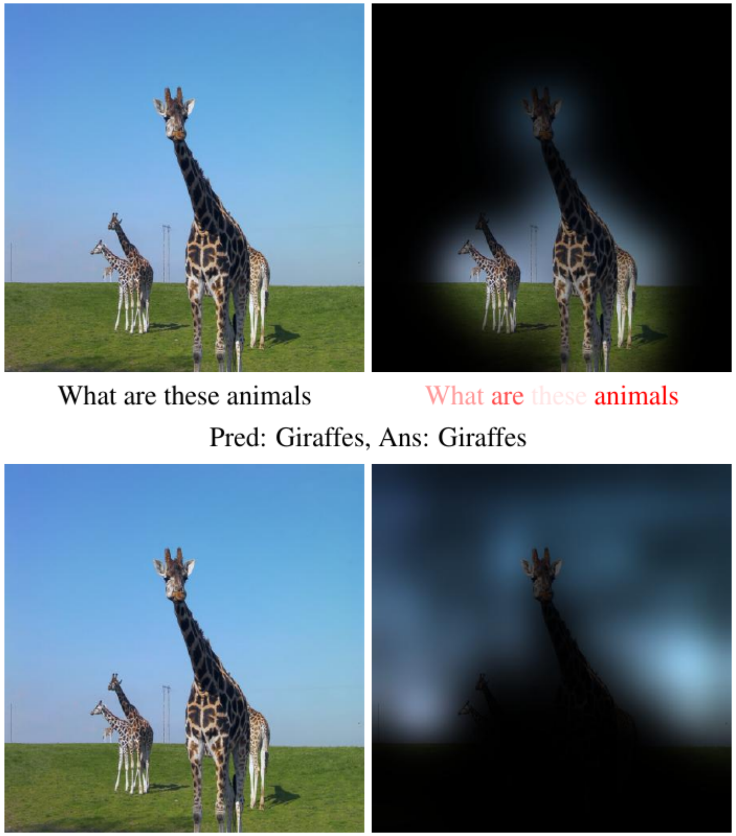
Attention maps - example



Why attention mechanism?

- Conditional representations
  - Meaning of a word in the context of a sentence
  - Meaning of an object in the context of a question
- Modeling long-term dependencies
  - $O(1)$ vs. $O(N)$ for RNNs
- Some interpretability

Attention mechanism - VQA

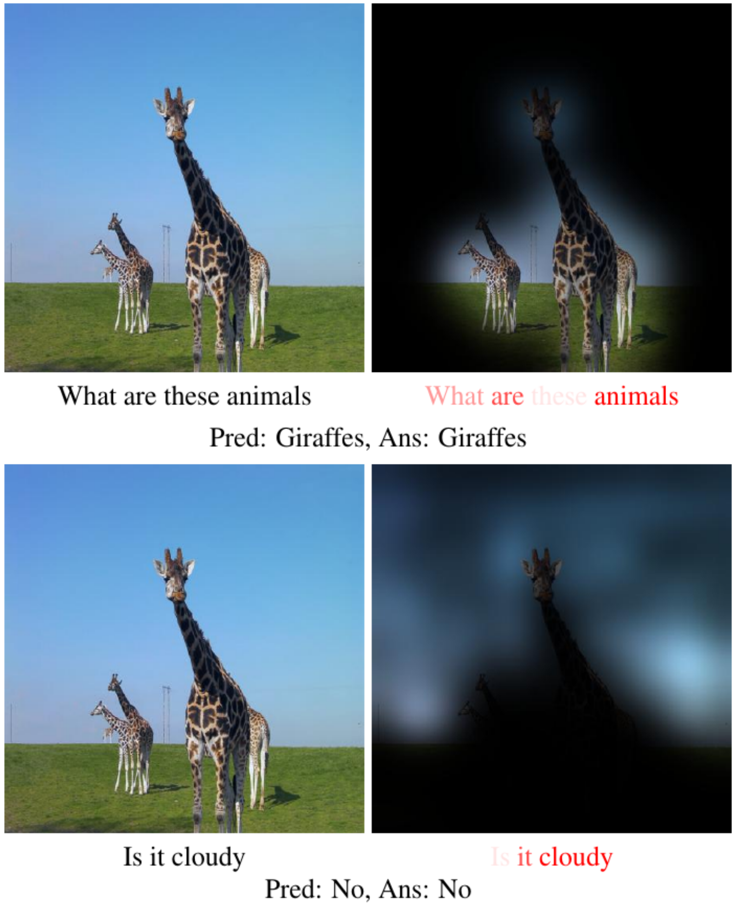


What are these animals What are these animals
Pred: Giraffes, Ans: Giraffes

Is it cloudy Is it cloudy
Pred: No, Ans: No

- Focus at relevant regions or relevant question words

Attention mechanism - VQA



- Focus at relevant regions or relevant question words
- Representations conditioned on the context

What sport is this woman playing

Dot-product attention

- $q \in \mathbb{R}^{d_q}$ - query
- $K \in \mathbb{R}^{L \times d_k}$ - keys
- $V \in \mathbb{R}^{L \times d_v}$ - values

$$\mathrm{attention}(Q, K, V) = \mathrm{softmax}(QK^\top)V$$

- $d_q = d_k$

Dot-product attention

- $q \in \mathbb{R}^{d_q}$ - query
- $K \in \mathbb{R}^{L \times d_k}$ - keys
- $V \in \mathbb{R}^{L \times d_v}$ - values

$$\text{attention}(Q, K, V) = \text{softmax}(QK^\top)V$$

- $d_q = d_k$

Method: DCN

dense, bi-directional interactions between the two modalities

- Each word represented in the context of the image
- Each image region represented in the context of the question

DCN - attention maps

- $Q_l = [q_{l1}, \dots, q_{lN}] \in \mathbb{R}^{d \times N}$ - N question words
- $V_l = [v_{l1}, \dots, v_{lT}] \in \mathbb{R}^{d \times T}$ - T image regions

Compute the affinity matrix:

$$A_l = V_l^\top W_l Q_l$$

- $A_l \in \mathbb{R}^{T \times N}$

DCN - attention maps

- $Q_l = [q_{l1}, \ldots, q_{lN}] \in \mathbb{R}^{d \times N}$ - N question words
- $V_l = [v_{l1}, \ldots, v_{lT}] \in \mathbb{R}^{d \times T}$ - T image regions

Compute the affinity matrix:

$$A_l = V_l^{\top} W_l Q_l$$

- $A_l \in \mathbb{R}^{T \times N}$

Two attention maps:

$$A_{Q_l} = \mathrm{softmax}(A_l)$$

$$A_{V_l} = \mathrm{softmax}\left(A_l^{\top}\right)$$

DCN - attention maps

- $Q_l = [q_{l1}, \ldots, q_{lN}] \in \mathbb{R}^{d \times N}$ - N question words
- $V_l = [v_{l1}, \ldots, v_{lT}] \in \mathbb{R}^{d \times T}$ - T image regions

Compute the affinity matrix:

$$A_l = V_l^\top W_l Q_l$$

- $A_l \in \mathbb{R}^{T \times N}$

Two attention maps:

$$A_{Q_l} = \mathrm{softmax}(A_l)$$

$$A_{V_l} = \mathrm{softmax}\left(A_l^\top\right)$$

This is of course **not** exactly what they do!

DCN - the actual attention maps

- Multiple attention maps: $A_l^{(i)}$ instead of A_l, where i - attention number

DCN - the actual attention maps

- Multiple attention maps: $A_l^{(i)}$ instead of A_l, where i - attention number
- Weight matrix W_l is replaced with two matrices of lower-rank: ${W_{\tilde{V}_l}^{(i)}}^\top W_{\tilde{Q}_l}^{(i)}$ where $W_{\tilde{V}_l}^{(i)} \in \mathbb{R}^{d_h \times d}$, $W_{\tilde{Q}_l}^{(i)} \in \mathbb{R}^{d_h \times d}$

$$A_l^{(i)} = \left(W_{\tilde{V}_l}^{(i)} \tilde{V}_l\right)^\top \left(W_{\tilde{Q}_l}^{(i)} \tilde{Q}_l\right)$$

DCN - the actual attention maps

- Multiple attention maps: $A_l^{(i)}$ instead of A_l, where i - attention number
- Weight matrix W_l is replaced with two matrices of lower-rank: ${W_{\tilde{V}_l}^{(i)}}^\top W_{\tilde{Q}_l}^{(i)}$ where $W_{\tilde{V}_l}^{(i)} \in \mathbb{R}^{d_h \times d}$, $W_{\tilde{Q}_l}^{(i)} \in \mathbb{R}^{d_h \times d}$

$$A_l^{(i)} = \left(W_{\tilde{V}_l}^{(i)} \tilde{V}_l\right)^\top \left(W_{\tilde{Q}_l}^{(i)} \tilde{Q}_l\right)$$

- Alternative low-rank approach:
  Kim, Jin-Hwa, Jaehyun Jun, and Byoung-Tak Zhang. "Bilinear attention networks." *Advances in Neural Information Processing Systems*. 2018.

Scaled by $\sqrt{d_h}$ (not justified in the paper)

$$A_{Q_l}^{(i)} = \mathrm{softmax}\left(\frac{A_l^{(i)}}{\sqrt{d_h}}\right)$$

$$A_{V_l}^{(i)} = \mathrm{softmax}\left(\frac{A_l^{(i)\top}}{\sqrt{d_h}}\right)$$

Scaled by $\sqrt{d_h}$ (not justified in the paper)

$$A_{Q_l}^{(i)} = \mathrm{softmax}\left(\frac{A_l^{(i)}}{\sqrt{d_h}}\right)$$

$$A_{V_l}^{(i)} = \mathrm{softmax}\left(\frac{A_l^{(i)\top}}{\sqrt{d_h}}\right)$$

- For high d_h the variance of dot products is high - very small gradients
- The scaling results in smoother distribution

Fusion of attention maps by averaging (usually are concatenated):

$$A_{Q_l} = \frac{1}{h} \sum_{i=1}^{h} A_{Q_l}^{(i)}$$

$$A_{V_l} = \frac{1}{h} \sum_{i=1}^{h} A_{V_l}^{(i)}$$

- $A_{Q_l} \in \mathbb{R}^{\tilde{T} \times \tilde{N}}$ - word probability for each image region
- $A_{V_l} \in \mathbb{R}^{\tilde{N} \times \tilde{T}}$ - image region probability for each word

Fusion of attention maps by averaging (usually are concatenated):

$$A_{Q_l} = \frac{1}{h} \sum_{i=1}^{h} A_{Q_l}^{(i)}$$

$$A_{V_l} = \frac{1}{h} \sum_{i=1}^{h} A_{V_l}^{(i)}$$

- $A_{Q_l} \in \mathbb{R}^{\tilde{T} \times \tilde{N}}$ - word probability for each image region
- $A_{V_l} \in \mathbb{R}^{\tilde{N} \times \tilde{T}}$ - image region probability for each word

Attended feature representations:

$$\hat{Q}_l = \tilde{Q}_l A_{Q_l}[1 : \mathrm{T}, :]^\top$$

- $\hat{Q}_l \in \mathbb{R}^{d \times T}$ - an average of word vectors weighted by their relevance to (compatibility with) the image regions

Fusion of attention maps by averaging (usually are concatenated):

$$A_{Q_l} = \frac{1}{h} \sum_{i=1}^{h} A_{Q_l}^{(i)}$$

$$A_{V_l} = \frac{1}{h} \sum_{i=1}^{h} A_{V_l}^{(i)}$$

- $A_{Q_l} \in \mathbb{R}^{\tilde{T} \times \tilde{N}}$ - word probability for each image region
- $A_{V_l} \in \mathbb{R}^{\tilde{N} \times \tilde{T}}$ - image region probability for each word

Attended feature representations:

$$\hat{Q}_l = \tilde{Q}_l A_{Q_l}[1:\mathrm{T},:]^\top$$

- $\hat{Q}_l \in \mathbb{R}^{d \times T}$ - an average of word vectors weighted by their relevance to (compatibility with) the image regions

$$\hat{V}_l = \tilde{V}_l A_{V_l}[1:\mathrm{N},:]^\top$$

- $\hat{V}_l \in \mathbb{R}^{d \times N}$ - an average of image region vectors weighted by their relevance to (compatibility with) the word

Fusion of attention maps by averaging (usually are concatenated):

$$A_{Q_l} = \frac{1}{h} \sum_{i=1}^{h} A_{Q_l}^{(i)}$$

$$A_{V_l} = \frac{1}{h} \sum_{i=1}^{h} A_{V_l}^{(i)}$$

- $A_{Q_l} \in \mathbb{R}^{\tilde{T} \times \tilde{N}}$ - word probability for each image region
- $A_{V_l} \in \mathbb{R}^{\tilde{N} \times \tilde{T}}$ - image region probability for each word

Attended feature representations:

$$\hat{Q}_l = \tilde{Q}_l A_{Q_l} [1 : \mathrm{T}, :]^\top$$

- $\hat{Q}_l \in \mathbb{R}^{d \times T}$ - an average of word vectors weighted by their relevance to (compatibility with) the image regions

$$\hat{V}_l = \tilde{V}_l A_{V_l} [1 : \mathrm{N}, :]^\top$$

- $\hat{V}_l \in \mathbb{R}^{d \times N}$ - an average of image region vectors weighted by their relevance to (compatibility with) the word

These are **still unimodal representations**, just attended

Fusing representations

Each word is fused with a (**unique**) representation of the image

$$q_{(l+1)n} = \mathrm{ReLU}\left(W_{Q_l}\begin{bmatrix} q_{ln} \\ \hat{v}_{ln} \end{bmatrix} + b_{Q_l}\right) + q_{ln}$$

- $\hat{v}_{ln}$ - meaning of the image in the context of the n-th word

Each image region is fused with a (**unique**) representation of the question

$$v_{(l+1)t} = \mathrm{ReLU}\left(W_{V_l}\begin{bmatrix} v_{lt} \\ \hat{q}_{lt} \end{bmatrix} + b_{V_i}\right) + v_{lt}$$

- $\hat{q}_{lt}$ - meaning of the question in the context of the t-th image region

DCN model

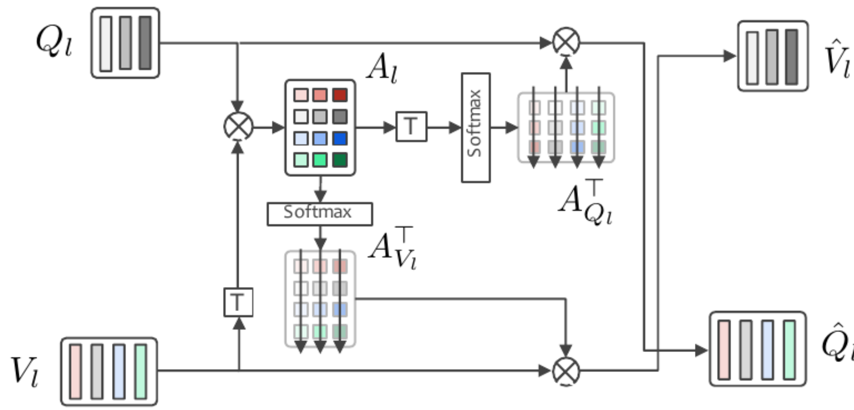


Figure 3: Computation of dense co-attention maps and attended representations of the image and question.

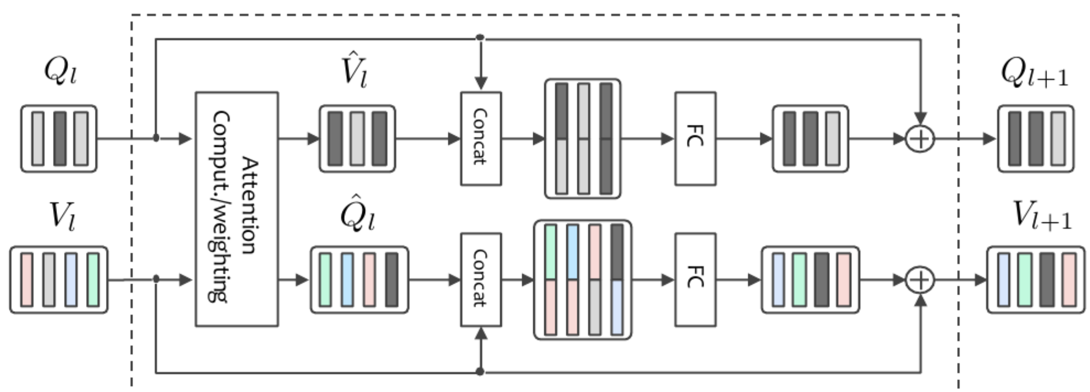


Figure 2: The internal structure of a single dense co-attention layer of layer index $l + 1$.

DCN model

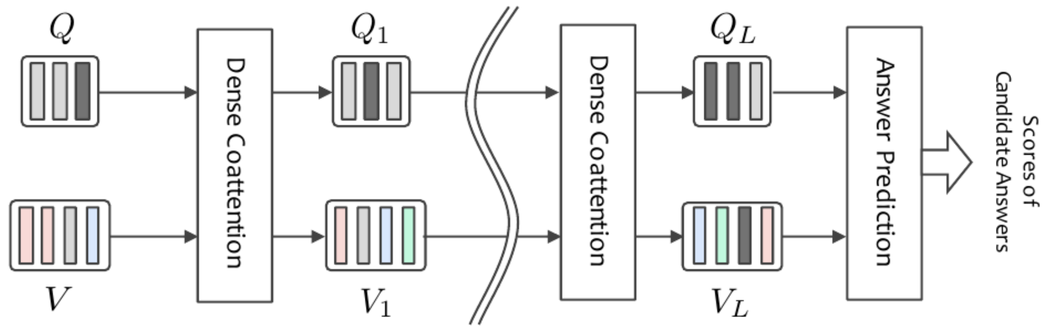


Figure 1: The global structure of the dense co-attention network (DCN).

Question representation

$$\overrightarrow{q_n} = \text{Bi-LSTM}\left(\overrightarrow{q_{n-1}}, e_n^Q\right)$$

$$\overleftarrow{q_n} = \text{Bi-LSTM}\left(\overleftarrow{q_{n+1}}, e_n^Q\right)$$

- e_n^Q - GloVe embedding of the n-th word

$$q_n = \left[\overrightarrow{q_n}^\top, \overleftarrow{q_n}^\top\right]^\top$$

$$Q = [q_1, \ldots, q_N] \in \mathbb{R}^{d \times N}$$

Image representation

4 layers from ResNet-152

- Each layer of different depth
- Different shapes → max pooling and 1 x 1 convolution
  → 4 layers, each of shape $d \times 14 \times 14$

Image representation

4 layers from ResNet-152

- Each layer of different depth
- Different shapes → max pooling and 1 x 1 convolution
  → 4 layers, each of shape $d \times 14 \times 14$

The relative importance of features corresponding to each depth level depends on the given question:

$$[\alpha_1, \alpha_2, \alpha_3, \alpha_4] = \mathrm{softmax}(\mathrm{MLP}\,(s_Q))$$

- Features weighted by alphas are summed together
- $V = [v_1, \ldots, v_T] \in \mathbb{R}^{d \times T}$
- $T = 14 \times 14$

DCN - predicting answers

$$s_{Q_L} = \sum_{n=1}^{N} \alpha_n^Q q_{Ln}$$

$$s_{V_L} = \sum_{n=1}^{N} \alpha_n^V v_{Ln}$$

Different methods for predicting answers:

$$(\text{score of answers encoded as } s_A) = \sigma\left(s_A^\top W\left(s_{Q_L} + s_{V_L}\right)\right), \quad (16)$$

$$(\text{ score of answers }) = \sigma\left(\text{MLP}\left(s_{Q_L} + s_{V_L}\right)\right), \quad (17)$$

$$(\text{ score of answers }) = \sigma\left(\text{MLP}\left(\begin{bmatrix} s_{Q_L} \\ s_{V_L} \end{bmatrix}\right)\right), \quad (18)$$

DCN - predicting answers

$$s_{Q_L} = \sum_{n=1}^{N} \alpha_n^Q q_{Ln}$$

$$s_{V_L} = \sum_{n=1}^{N} \alpha_n^V v_{Ln}$$

Different methods for predicting answers:

$$\text{(score of answers encoded as } s_A) = \sigma\left(s_A^\top W\left(s_{Q_L} + s_{V_L}\right)\right), (16)$$

$$(\text{ score of answers }) = \sigma\left(\text{MLP}\left(s_{Q_L} + s_{V_L}\right)\right), (17)$$

$$(\text{ score of answers }) = \sigma\left(\text{MLP}\left(\begin{bmatrix} s_{Q_L} \\ s_{V_L} \end{bmatrix}\right)\right), (18)$$

- (17) and (18) can produce **only answers that are considered when training**

Experiments

Datasets

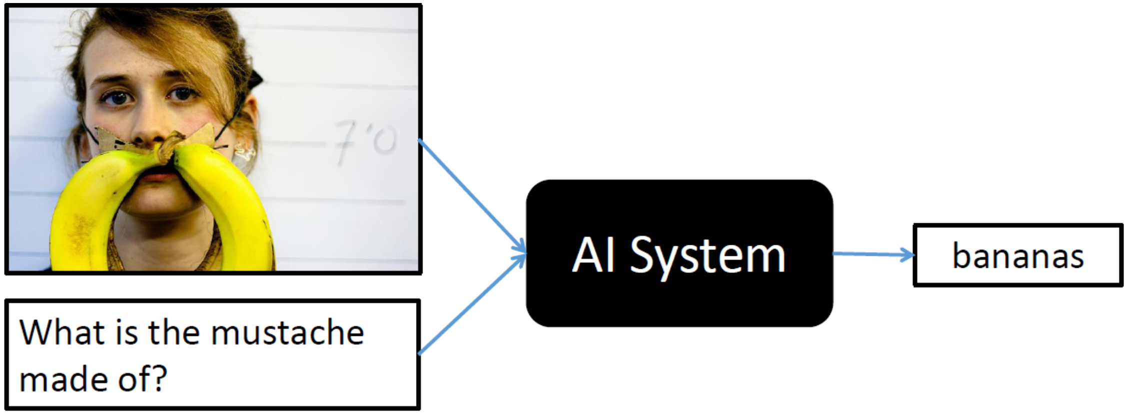
- Images from MS-COCO (200k+ images)

VQA (1.0):

- 240k+ train, 120k+ val, 240k+ test questions

VQA 2.0:

- The largest VQA dataset
- 440k+ train, 210k+ val, 440k+ test questions
- Reduced language bias



Results VQA 1.0

Table 2: Results of the proposed method along with published results of others on VQA 1.0 in similar conditions (i.e., a single model; trained without an external dataset).

Model	Test-dev				Test-standard			
	Overall	Other	Number	Yes/No	Overall	Other	Number	Yes/No
VQA team [2]	57.75	43.08	36.77	80.50	58.16	43.73	36.53	80.569
SMem [31]	57.99	43.12	37.32	80.87	58.24	43.48	37.53	80.80
SAN [32]	58.70	46.10	36.60	79.30	58.90	-	-	-
FDA [12]	59.24	45.77	36.16	81.14	59.54	-	-	-
DNMN [1]	59.40	45.50	38.60	81.10	59.40	-	-	-
HieCoAtt [21]	61.00	51.70	38.70	79.70	62.10	-	-	-
RAU [24]	63.30	53.00	39.00	81.90	63.20	52.80	38.20	81.70
DAN [23]	64.30	53.90	39.10	83.00	64.20	54.00	38.10	82.80
Strong Baseline [14]	64.50	55.20	39.10	82.20	64.60	55.20	39.10	82.00
MCB [6]	64.70	55.60	37.60	82.50	-	-	-	-
N2NMNs [11]	64.90	-	-	-	-	-	-	-
MLAN [35]	64.60	53.70	40.20	83.80	64.80	53.70	40.90	83.70
MLB [16]	65.08	54.87	38.21	84.14	65.07	54.77	37.90	84.02
MFB [36]	65.90	56.20	39.80	84.00	65.80	56.30	38.90	83.80
MF-SIG-T3 [5]	66.00	56.37	39.34	84.33	65.88	55.89	38.94	84.42
DCN (16)	66.43	56.23	42.37	84.75	66.39	56.23	41.81	84.53
DCN (17)	**66.89**	57.31	**42.35**	**84.61**	**67.02**	**56.98**	**42.34**	**85.04**
DCN (18)	66.83	**57.44**	41.66	84.48	66.66	56.83	41.27	84.61

Results VQA 2.0

Table 3: Results of the proposed method along with published results of others on VQA 2.0 in similar conditions (i.e., a single model; trained without an external dataset). DCN(number) indicates the DCN equipped with the prediction layer that uses equation (number) for score computation. *: trained with external datasets. ‡: the winner of VQA challenge 2017, unpublished.

Model	Test-dev				Test-standard			
	Overall	Other	Number	Yes/No	Overall	Other	Number	Yes/No
VQA team-Prior [8]	-	-	-	-	25.98	01.17	00.36	61.20
VQA team-Language only [8]	-	-	-	-	44.26	27.37	31.55	67.01
VQA team-LSTM+CNN [8]	-	-	-	-	54.22	41.83	35.18	73.46
MCB [6] reported in [8]	-	-	-	-	62.27	53.36	38.28	78.82
MF-SIG-T3 * [5]	64.73	55.55	42.99	81.29	-	-	-	-
Adelaide Model * ‡ [28]	62.07	52.62	39.46	79.20	62.27	52.59	39.77	79.32
Adelaide + Detector * ‡ [28]	65.32	56.05	44.21	81.82	65.67	56.26	43.90	82.20
DCN (16)	**66.87**	**57.26**	46.61	83.51	66.97	**57.09**	46.98	83.59
DCN (17)	66.72	56.77	**46.65**	**83.70**	**67.04**	56.95	**47.19**	83.85
DCN (18)	66.60	56.72	46.60	83.50	67.00	56.90	46.93	**83.89**

Ablation study

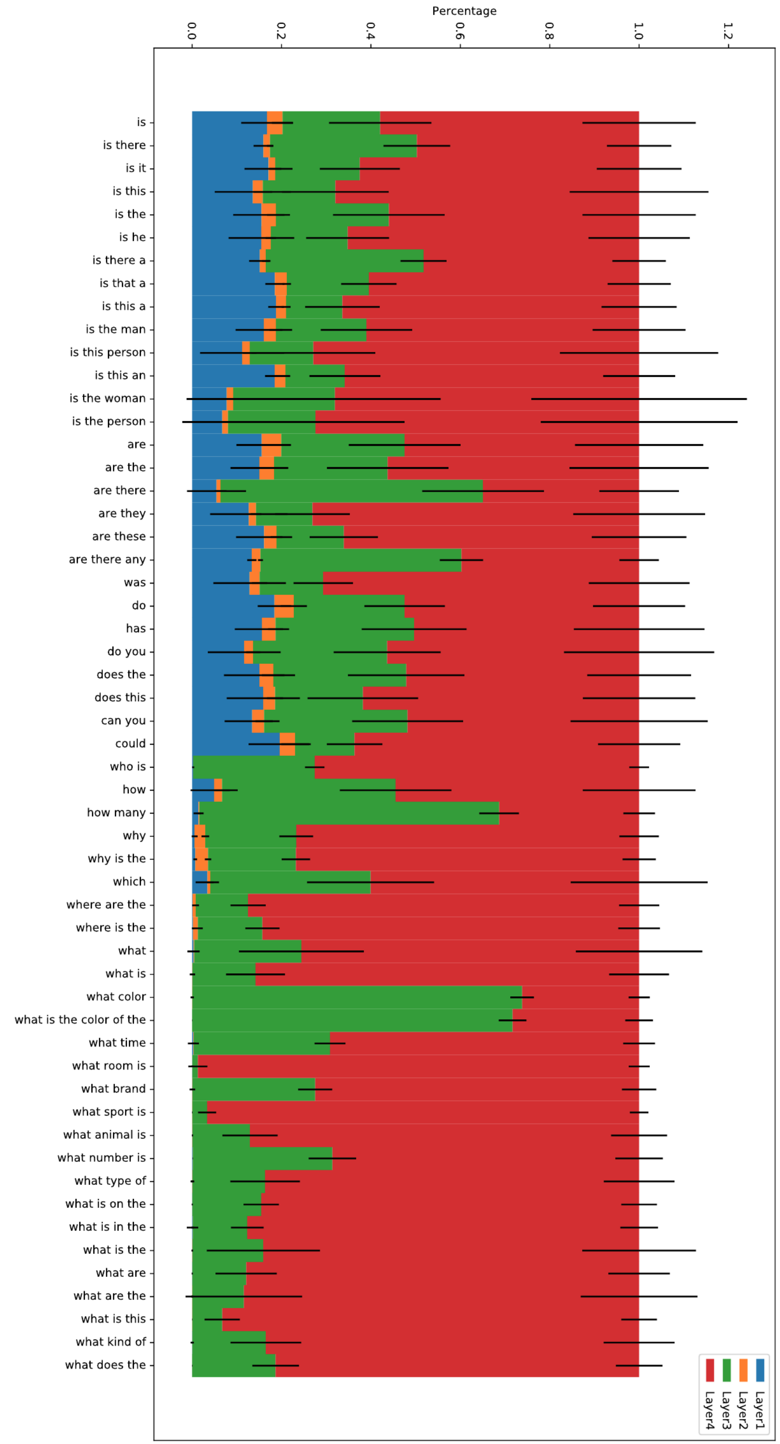
Table 1: Ablation study on each module of DCNs using the validation set of the Open-Ended task (VQA 2.0). * indicates modules employed in the final model.

Category	Detail	Accuracy
Attention direction	$I \leftarrow Q$	60.95
	$I \rightarrow Q$	62.63
	$I \leftrightarrow Q$*	62.94
Memory size (K)	1	62.53
	3*	62.94
	5	62.83
Number (h) of parallel attention maps	2	62.82
	4*	62.94
	8	62.81
Number (L) of stacked layers	1	62.43
	2	62.82
	3*	62.94
	4	62.67
Attention in answer prediction layer	Attention used*	62.94
	Avg of features	61.63
Attention in image extraction layer	Attention used*	62.94
	Only last conv layer	62.39

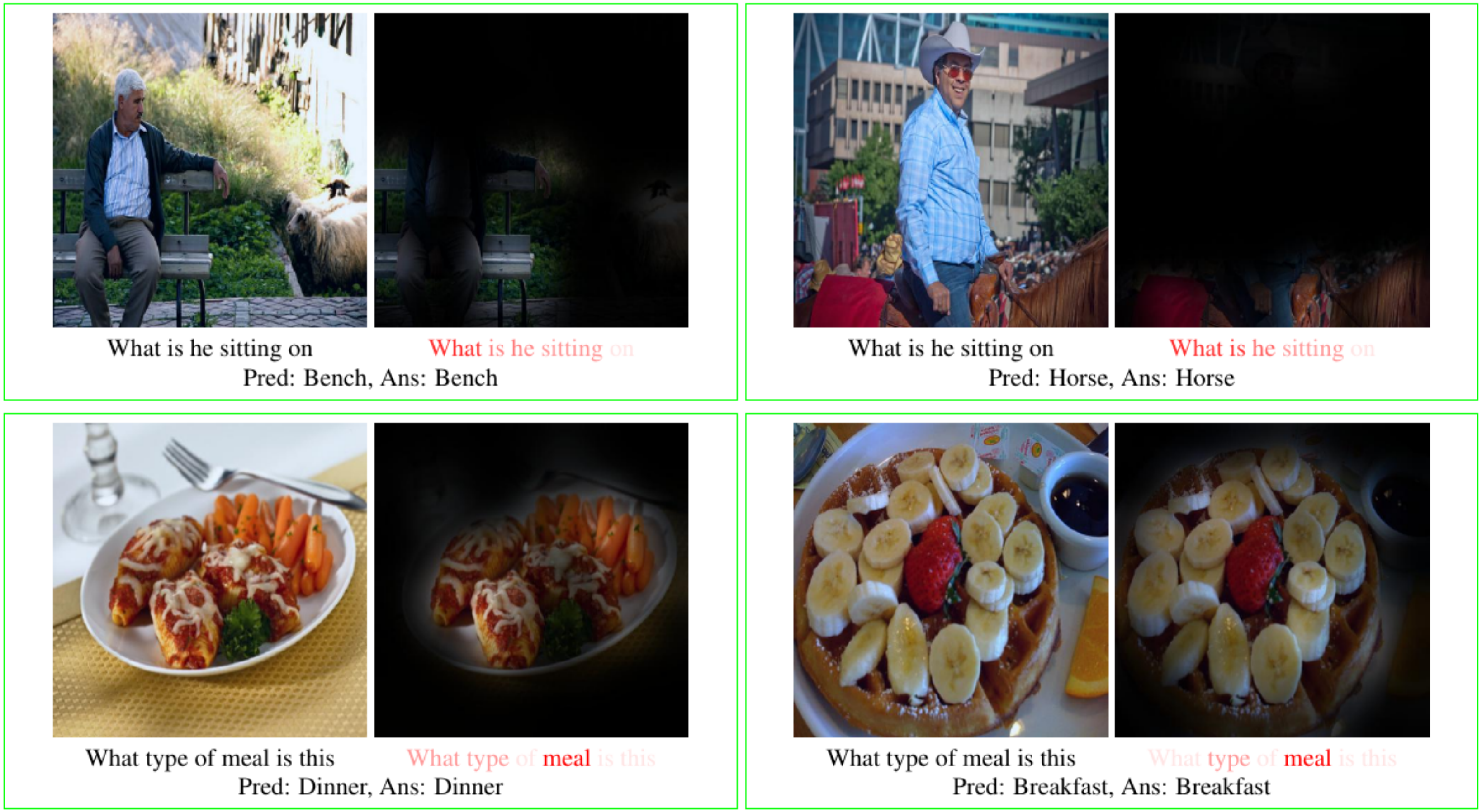
- $(I \leftarrow Q)$ - question-guided attention on image region
- $(I \rightarrow Q)$ - image-guided attention on question words
- $(I \leftrightarrow Q)$ - DCN co-attention: attention in both directions

How deep features?

- Layer 1:
  - *Yes/No* questions
  - *is/are/does/can/could*
- Layer 3:
  - High importance on questions about colors
- Layer 4:
  - Highest importance in general
  - **semantics**: *what*



Qualitative evaluation



How many vases are in the photo

How many vases are in the photo

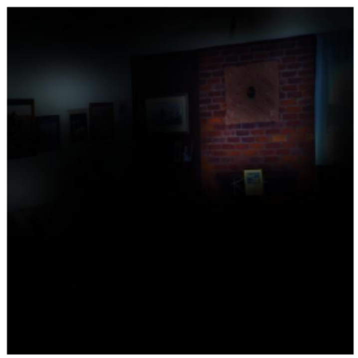
Pred: 2, Ans: 2

How many vases are in the photo

How many vases are in the photo

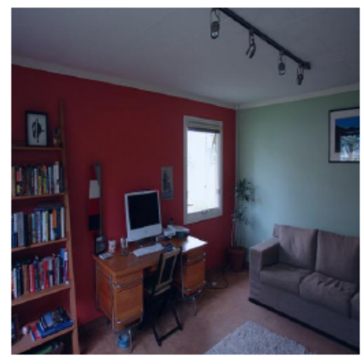
Pred: 1, Ans: 1

What is the darker wall made of

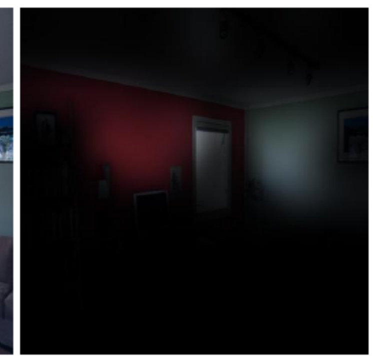


What is the darker wall made of

Pred: Brick, Ans: Brick



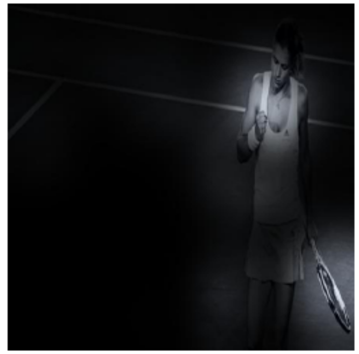
What is the darker wall made of



What is the darker wall made of

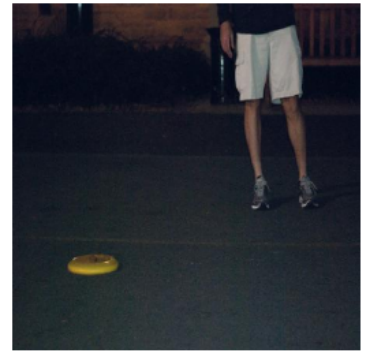
Pred: Drywall, Ans: Drywall

What sport is this woman playing

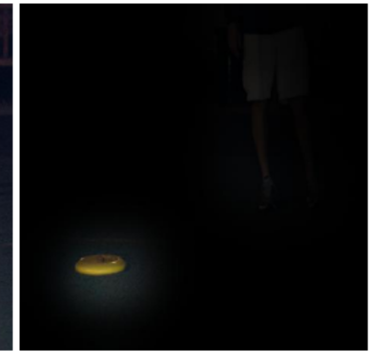


What sport is this woman playing

Pred: Tennis, Ans: Tennis



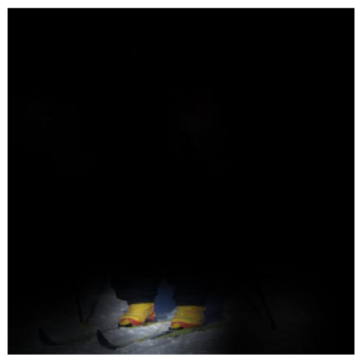
What sport is this woman playing



What sport is this woman playing

Pred: Frisbee, Ans: Frisbee

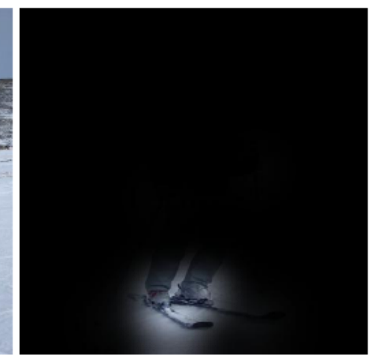
What color are the skiers shoes



What color are the skiers shoes

Pred: Yellow, Ans: Yellow

What color are the skiers shoes



What color are the skiers shoes

Pred: White, Ans: White

What breed of dog is this | What breed of dog is this

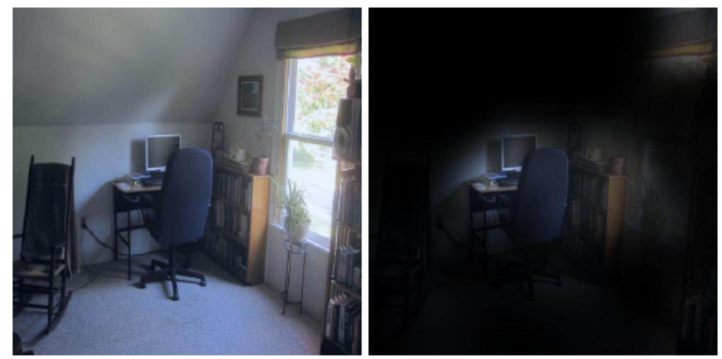
Pred: Mutt, Ans: Lab *(Error type: 1)*

What breed of dog is this | What breed of dog is this

Pred: Terrier, Ans: Terrier

What room is this | What room is this

Pred: Bedroom, Ans: Bedroom



What room is this | What room is this

Pred: Living room, Ans: Office *(Error type: 1)*

What is the name of the utensil What is the name of the utensil
Pred: Fork, Ans: Fork

What is the name of the utensil What is the name of the utensil
Pred: Fork, Ans: Spoon *(Error type: 1)*

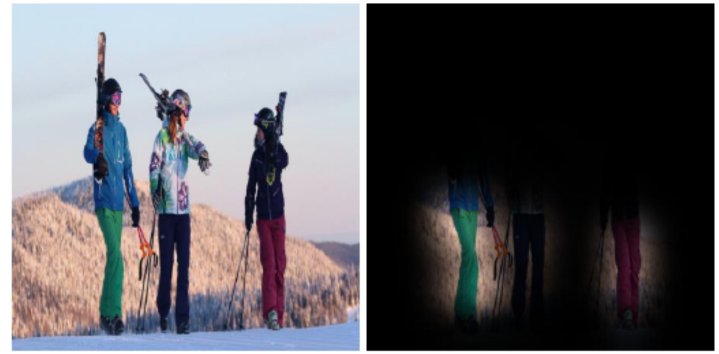
How tall is he How tall is he
Pred: 5 feet, Ans: Tall *(Error type: 1)*

How tall is he How tall is he
Pred: 5 feet, Ans: 6 feet *(Error type: 2)*

What is the color of pants the woman is wearing

What is the color of pants the woman is wearing

Pred: Plaid, Ans: Red and White *(Error type: 4)*



What is the color of pants the woman is wearing

What is the color of pants the woman is wearing

Pred: Green, Ans: Black *(Error type: 4)*

What color is lit up on the street lights

What color is lit up on the street lights

Pred: Yellow, Ans: Green *(Error type: 3)*

What color is lit up on the street lights

What color is lit up on the street lights

Pred: White, Ans: None *(Error type: 1)*

Thoughts

- matter the Does order?

Thoughts

- matter the Does order?
  - Permuation of the order of the features (but not the inputs) has no effect
  - Global or relative positions

Thoughts

- matter the Does order?
  - Permuation of the order of the features (but not the inputs) has no effect
  - Global or relative positions
- Counting from weighted averages

Thoughts

- matter the Does order?
  - Permuation of the order of the features (but not the inputs) has no effect
  - Global or relative positions
- Counting from weighted averages
- Not sure how conclusive ablation study is

Thoughts

- matter the Does order?
  - Permuation of the order of the features (but not the inputs) has no effect
  - Global or relative positions
- Counting from weighted averages
- Not sure how conclusive ablation study is
- Other paper with dense interactions:
  - Kim, Jin-Hwa, Jaehyun Jun, and Byoung-Tak Zhang. "Bilinear attention networks." *Advances in Neural Information Processing Systems*. 2018.

Thoughts

- matter the Does order?
  - Permuation of the order of the features (but not the inputs) has no effect
  - Global or relative positions
- Counting from weighted averages
- Not sure how conclusive ablation study is
- Other paper with dense interactions:
  - Kim, Jin-Hwa, Jaehyun Jun, and Byoung-Tak Zhang. "Bilinear attention networks." *Advances in Neural Information Processing Systems*. 2018.
- Often high dataset biases in VQA problems
- Do attention maps look at the same regions as humans?
  - Das, Abhishek, et al. "Human attention in visual question answering: Do humans and deep networks look at the same regions?." *Computer Vision and Image Understanding 163 (2017): 90-100.*

The End