Differential Attention for Visual Question Answering

Badri Patro, Vinay P. Namboodiri
IIT Kanpur
{ badri,vinaypn }@iitk.ac.in

Abstract

In the supplementary material, we provide details regarding the experimental setup used while training the proposed method and details about the datasets used. We further provide additional result table that contains all the results for the different variants used. Additionally we provide a number of different attention visualisations that include comparisons with stacked attention, supporting and opposing exemplars and different human attention visualisations.

A. Experimental Setup

A.1. Dataset

We have conducted our experiments on two types of dataset, first one is VQA dataset ,which contains human annotated question and answer based on images on MS-COCO dataset. Second one is HAT dataset based on attention map.

A.1.1 VQA dataset

VQA dataset[1] is one of the largest dataset for VQA benchmark so far. It built on complex images from ms-coco dataset. VQA dataset contains total 204721 images, out of which, 82783 images for training, 40504 images for validation and 81434 images for testing. Each image in the MS-COCO dataset[4] is associated with 3 questions and each question has 10 possible answers. This dataset is annotated by different people. So there are 248349 QA pair for training, 121512 QA pairs for validating and 244302 QA pairs for testing. We use the top 1000 most frequently output as our possible answer set as is commonly used. This covers 82.67% of the train+val answer.

A.1.2 VQA-HAT(Human Attention) dataset

We used VQA-HAT dataset[2], which is developed based on the de-blurring task to answering visual questions. This dataset contains human attention map for training set of 58475 example out of 248349 VQA training set. It contains 1374 validation example out of 121512 examples of question image pair in VQA validation set.

A.2. Evaluation methods

A.2.1 VQA dataset

VQA dataset contain 3 type of answer: yes/no, number and other. The evaluation is carried out using two test splits,i.e test-dev and test-standard. The question in corresponding test split are answered using two ways: Open-Ended[1] and Multiple-choice. Open-Ended task should generate a natural language answer in form of single word or phrase. For each question there are 10 candidate answer provided with their respective confidence level. Our module generates a single word answer on the open ended task. This answer can be evaluated using accuracy metric provide by Antol *et al.*[1] as follows.

$$Acc = \frac{1}{N}\sum_{i=1}^{N} min(\frac{\sum_{t \in T^i} \mathbb{I}[a_i = t]}{3}, 1) \qquad (1)$$

Where a_i the predicted answer and t is the annotated answer in the target answer set T^i of the i^{th} example and $\mathbb{I}[.]$ is the indicator function. The predicted answer a_i is correct if at least 3 annotators agree on the predicted answer. If the predicted answer is not correct then the accuracy score depends on the number of annotator that agree on the answer. Before checking accuracy, we need to convert the predicted answer to lowercase, number to digits and punctuation & article to be removed.

A.2.2 HAT dataset

We used rank correlation technique to evaluate[2] the correlation between human attention map and DAN attention probability. Here we scale down human attention map to 14x14 in order to make same size as DAN attention probability. We then compute rank correlation using the following steps. Rank correlation technique is used to obtain the degree of association between the data. The value of rank correlation[5] lies between +1 to -1. When R_{Cor} is close

Algorithm 1 Rank Correlation Procedure

procedure :(Initialization)
  P_{HAM}: Probability distribution of Human Attention Map
  P_{DAN} : Probability distribution of Differential Attention
Rank:
  Compute Rank of P_{HAM} : R_{HAM}
  Compute Rank of P_{DAN} : R_{DAN}
Rank Difference :
  Compute difference in rank between R_{HAM} & R_{DAN} : $Rank_{Diff}$
  Compute square of rank difference $Rank_{Diff}$:S_{Rank_Diff}
Rank Correlation:
  Compute Dimension of P_{DAN} :N
  Compute Rank Correlation using :

$$R_{Cor} = 1 - \frac{6 * S_{Rank\_Diff}}{N^3 - N}$$

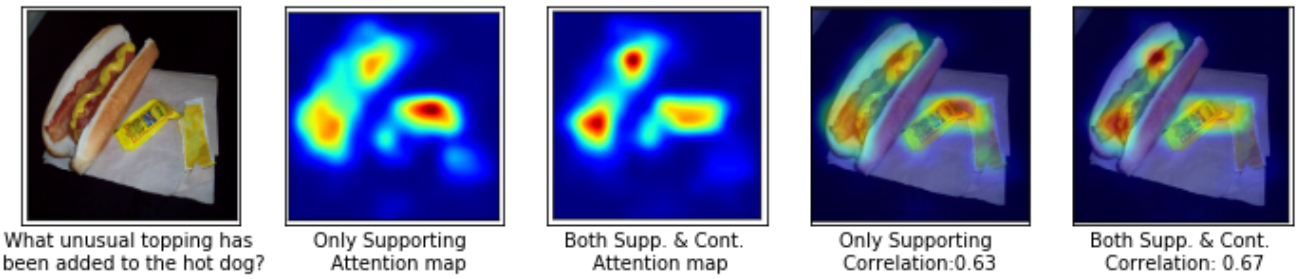

Figure 1. Importance of Supporting exemplar vs both. the first column in the figure indicates about image and corresponding question, the second and third term indicates attention map for supporting exemplar and both supporting and opposing exemplar. The fourth and fifth column gives the value of rank correlation for supporting and both.

to 1, it indicates positive correlation between them, When R_{Cor} is close to -1, it indicates negative correlation between them, and when R_{Cor} is close to 0, it indicates No correlation between them. A higher value of rank correlation is better.

A.3. Training and Model Configuration

We trained the differential attention model using joint loss in an end-to-end manner. We have used RMSPROP optimizer to update the model parameter and configured hyper-parameter values to be as follows: learning rate =0.0004 , batch size = 200, alpha = 0.99 and epsilon=1e-8 to train the classification network . In order to train a triplet model, we have used RMSPROP to optimize the triplet model model parameter and configure hyper-parameter values to be: learning rate =0.001 , batch size = 200, alpha = 0.9 and epsilon=1e-8. We have used learning rate decay to to decrease the learning rate on every epoch by a factor given by:

$$Decay\_factor = exp\left(\frac{log(0.1)}{a * b}\right)$$

where value of a=1500 and b=1250 is set empirically. Selection of training controlling factor(ν) has a major role during training. If ν=1 , means updating triplet and classification network parameter at a same rate . If $\nu \gg 1$, means updating triplet net more frequently as compare to classification net. Since, triplet loss decreases much lower then classification loss, we fixed value of $\nu \gg 1$ that is a fixed value of ν=10.

B. Results

In this section we provide additional results that were omitted from the main paper due to space limitation.

B.1. Rank Correlation results for DAN and DCN

In this sub-section we provide a few additional columns that were omitted from the results table in the main paper. Table- 1 provides the statistics of rank correlation on HAT validation dataset for various differential attention networks(DAN) and differential context network(DCN). DAN network is varied by varying the number of nearest supporting and opposing exemplars. We did experiments by considering K=1,2, 3,4,5 and random selections of nearest and

Table 1. Rank Correlation on HAT Validation Dataset for DAN and DCN

Models	Val1	Val2	Val3	Val
DAN (K=1)	0.3147	0.2772	0.2958	0.2959
DAN (K=2)	0.3280	0.2933	0.3057	0.3090
DAN (K=3)	0.3297	0.2947	0.3058	0.3100
DAN (K=4)	**0.3418**	**0.3060**	**0.3133**	**0.3206**
DCN Add_v1(K=1)	0.3172	0.2783	0.2968	0.2974
DCN Add_v2(K=1)	0.3186	0.2812	0.2993	0.2997
DCN Mul_v1(K=1)	0.3205	0.2847	0.3023	0.3025
DCN Mul_v2(K=1)	**0.3227**	**0.2871**	**0.3059**	**0.3052**
DCN Add_v1(K=4)	0.3426	0.3058	0.3123	0.3202
DCN Add_v2(K=4)	0.3459	0.3047	0.3140	0.3215
DCN Mul_v1(K=4)	0.3466	0.3059	0.3163	0.3229
DCN Mul_v2(K=4)	**0.3472**	**0.3068**	**0.3187**	**0.3242**
DAN (K=1,Random)	0.1238	0.1070	0.1163	0.1157
DAN (K=5)	0.2634	0.2412	0.2589	0.2545

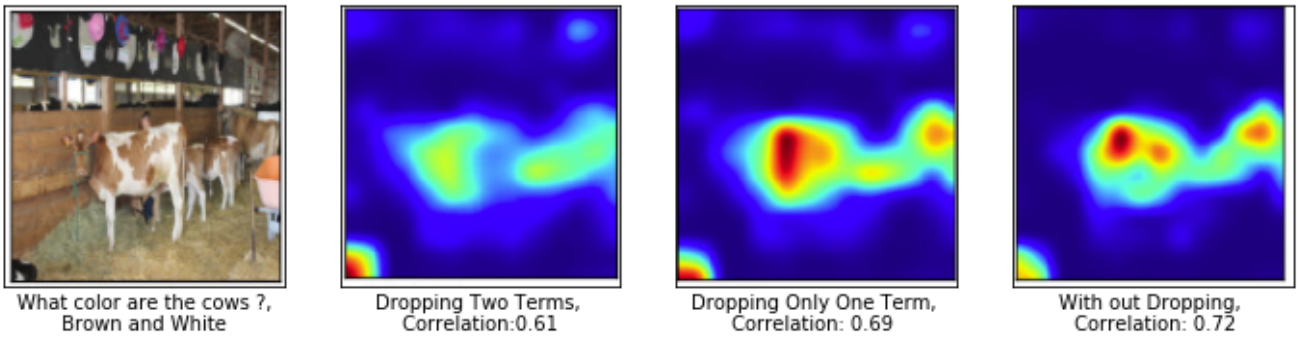

Figure 2. Ablation Results for Dropping terms in equation 3 and 4. The first column indicate the target image and its question, The second column provides the attention map & rank correlation by dropping 2^{nd} in equation 3 & i^{st} term in equation 4. The third column gives the attention map & rank correlation by dropping only i^{st} term in equation 4. Final column provides the attention map & rank correlation by consider every thing in both the equation.

farthest neighbors.

This table also mentions rank correlation for two types of DCN, i.e DCN Add and DCN Mul. Each type of network has two different methods for training , one is fixed scaling weights,i.e DCN Mul and second one is learn-able scaling weights,i.e, DCN Mul_v1. From the statistics of rank correlation in this table indicates that learnable scaling weights performs better than fixed weights. Further, we observed that Multiplication network performs better than addition network in case of differential context. We did experiments for K=1,2,3,4, but this table only shows the results of K=1 and K=4 for number of nearest and farthest neighbors selections.

We have computed rank correlation between attention probability of differential netwrok(DAN or DCN) and Human attention[2] provided by HAT for validation set. This table contains 3 rank correlations for 3 attention maps per image on HAT validation dataset. First attention map gives better accuracy than other two. Finally we take an average of three rank correlation for a particular model. We can observe that, all our model attention maps correlate positively with human attention.

B.2. How important are the supporting and contrasting exemplar?

We carried out an experiment by considering only the supportive exemplar in triplet loss mentioned in equation-2 and obtained consistent result as shown in figure 1. From the rank correlation result, we can conclude that, If we use only the supportive exemplar, we obtain most of the gain in the performance. The quantitative results for this ablation analysis is shown in the table 2, which provides the rank correlation on HAT Validation Dataset.

B.3. Contribution of different term in DCN

We carried out an experiment by dropping the vector projection of s_i^- on s_i term in the supporting context r_i^+ as mentioned in equation-3 and the vector rejection of s_i^+ on

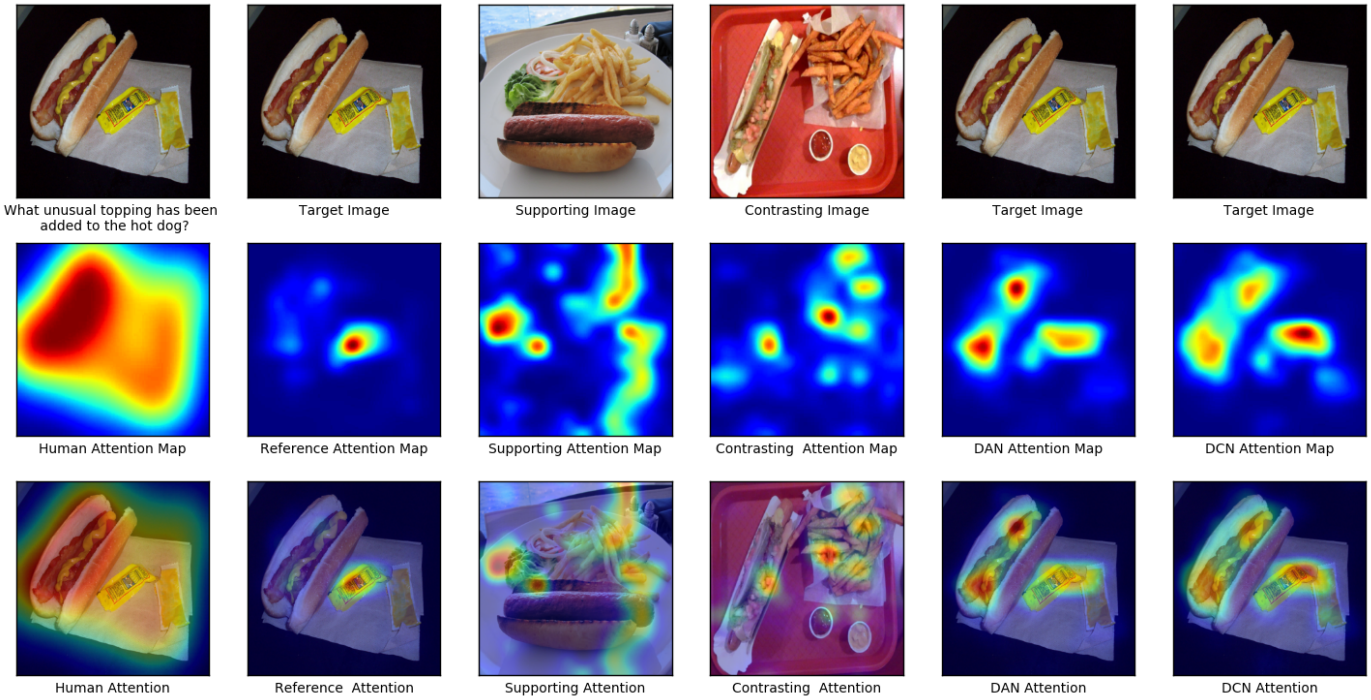

Figure 3. In this figure, the first raw indicates given target image, supporting image and opposing image. second raw indicate the attention map for human[2], reference attention map, supporting attention map , opposing attention map, DAN and DCN attention map respectively. Third raw generate result by applying attention map on corresponding images.

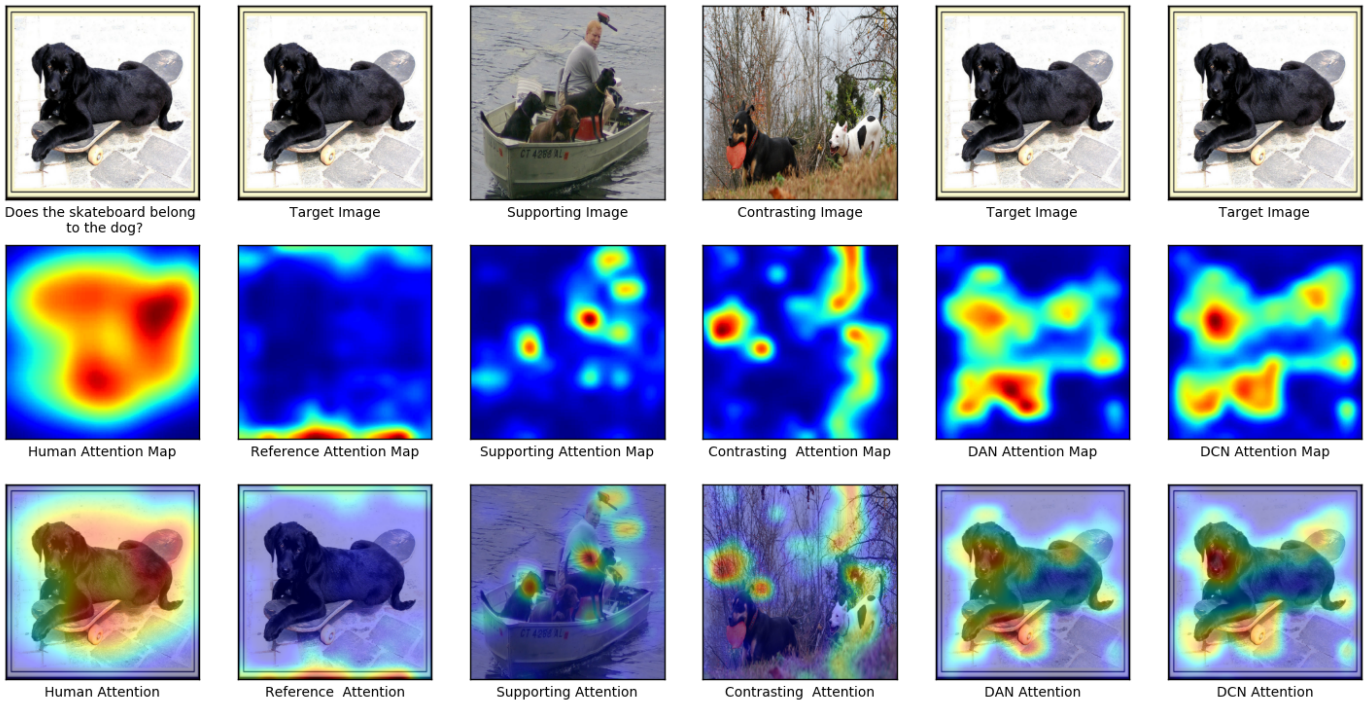

Figure 4. In this figure, the first raw indicates given target image, supporting image and opposing image. second raw indicate the attention map for human[2], reference attention map, supporting attention map , opposing attention map, DAN and DCN attention map respectively. Third raw generate result by applying attention map on corresponding images.

Table 2. Rank Correlation for only Supporting Exemplar

Models	Rank-correlation
DAN (K=4) +LQIA	0.312 ± 0.001
DAN (K=4) +MCB	0.320 ± 0.001

s_i term in opposing context r_i^- as mentioned in equation-4 and obtained consistent result as shown in figure 2. The contribution of these terms in the corresponding equations are very small.The quantitative results for this ablation analysis is shown in the table 3, which provides the rank correlation on HAT Validation Dataset.

Table 3. Rank Correlation by Dropping various terms in DCN

Models	Rank-correlation
DCN Mul_v2(K=4) +LQIA	**0.319± 0.001**
DCN Mul_v2(K=4) +MCB	**0.3287± 0.001**

B.4. Attention Visualization DAN and DCN with Supporting and Opposing Exemplar

The first row of figure- 3 indicates the target image along with a supporting and opposing image. Second row provides human attention map, reference, supporting, opposing, DAN and DCN attention map respectively. Third row gives corresponding attention visualization for all the images. We can observe that from the given the target image and question: "what unusual topping has been added to the hot dog" , the reference model provides attention map(3^{rd} row, 2^{nd} column of figure- 3) somewhere in the yellow part which is different from the ground truth human attention map (3^{rd} row, 1^{st} column of figure- 3). With the help of supporting and contrasting exemplar attention map(3^{rd} row, 3^{rd} & 4^{th} column of figure- 3), the reference model attention is improved, which is shown in DAN and DCN (3^{rd} row, 5^{th} & 6^{th} column of figure- 3). The attention map of DCN model is more correlated with the ground truth human attention map than reference model. Thus we observe that with the help of supporting and contrasting exemplar, VQA accuracy is improving. Also, figure- 4 provides attention visualization for DAN and DCN with the help of supporting and contrasting attention.

B.5. Attention visualization of DCN with various Human Attention Maps

We compute rank correlation for all three ground truth human attention map provide by VQA- HAT[2] val dataset with our DAN and DCN exemplar model and also visualized attention map with all thee ground truth human attention map as shown in figure 5 and 6. We can evaluated our rank correlation for all three human attention map and observed that human attention map one is better than attention map 2 and 3 in term of visualization and rank correlation as mention in figure 5 and 6.

B.6. Attention Visualization of DAN and DCN

We provide the results of the attention visualization in figure 7 and 8. As can be observed in figure 7 and 8, we obtain significant improvement of rank correlation in attention map by using exemplar model(DCN or DAN) as compared to the SAN method [8]. We can observed that DAN and DCN has more correlation with human attention. We observed that DAN and DCN has better rank correlation then SAN attention map.

C. Details of Triplet and Quintuplet Network

C.1. Triplet Model

The concept triplet loss is motivated in the context of larger margin nearest neighbor classification[7], which minimize the distance between target and supporting feature and maximize the distance between target and contrasting feature. $f(x_i)$ is the embedding feature of i^{th} example of training image x_i in n dimensional euclidean space.

- $f(s_i)$: The embedding of target
- $f(s_i^+)$:The embedding of supporting exemplar
- $f(s_i^-)$:The embedding of contrasting exemplar

The objective of triplet loss is to make both supporting features target will have same identity[6] & target and contrasting feature will have differ identity. which means it brings all supporting features more close to target feature than that of contrasting features.

$$\begin{aligned} &D(f(s_i), f(s_i^+)) + \alpha < D(f(s_i), f(s_i^-)) \\ &\forall (f(s_i), f(s_i^+), f(s_i^-)) \in T \end{aligned} \tag{2}$$

where $D(f(s_i), f(s_j)) = ||f(s_i) - f(s_j)||_2^2$ is defined as the euclidean distance between $f(s_i)\&f(s_j)$. α is the margin between supporting and contrasting feature.The default value of α is 0.2. T is training dataset set, which contain all set of possible triplets. The objective function for triplet loss is given by

$$\begin{aligned} &T(s_i, s_i^+, s_i^-) \\ &= \max(0, ||f(s_i) - f(s_i^+)||_2^2 + \alpha - ||f(s_i) - f(s_i^-)||_2^2) \end{aligned} \tag{3}$$

For simplicity ,the notation are replaced like this , $f(s_i) \rightarrow f, f(s_i^+) \rightarrow f^+, f(s_i^-) \rightarrow f^-$.

Gradient computation of L2 norm is given by

$$\frac{\partial}{\partial x}||f(x)||_2^2 = 2 * f(x)\frac{\partial}{\partial x}f(x) \tag{4}$$

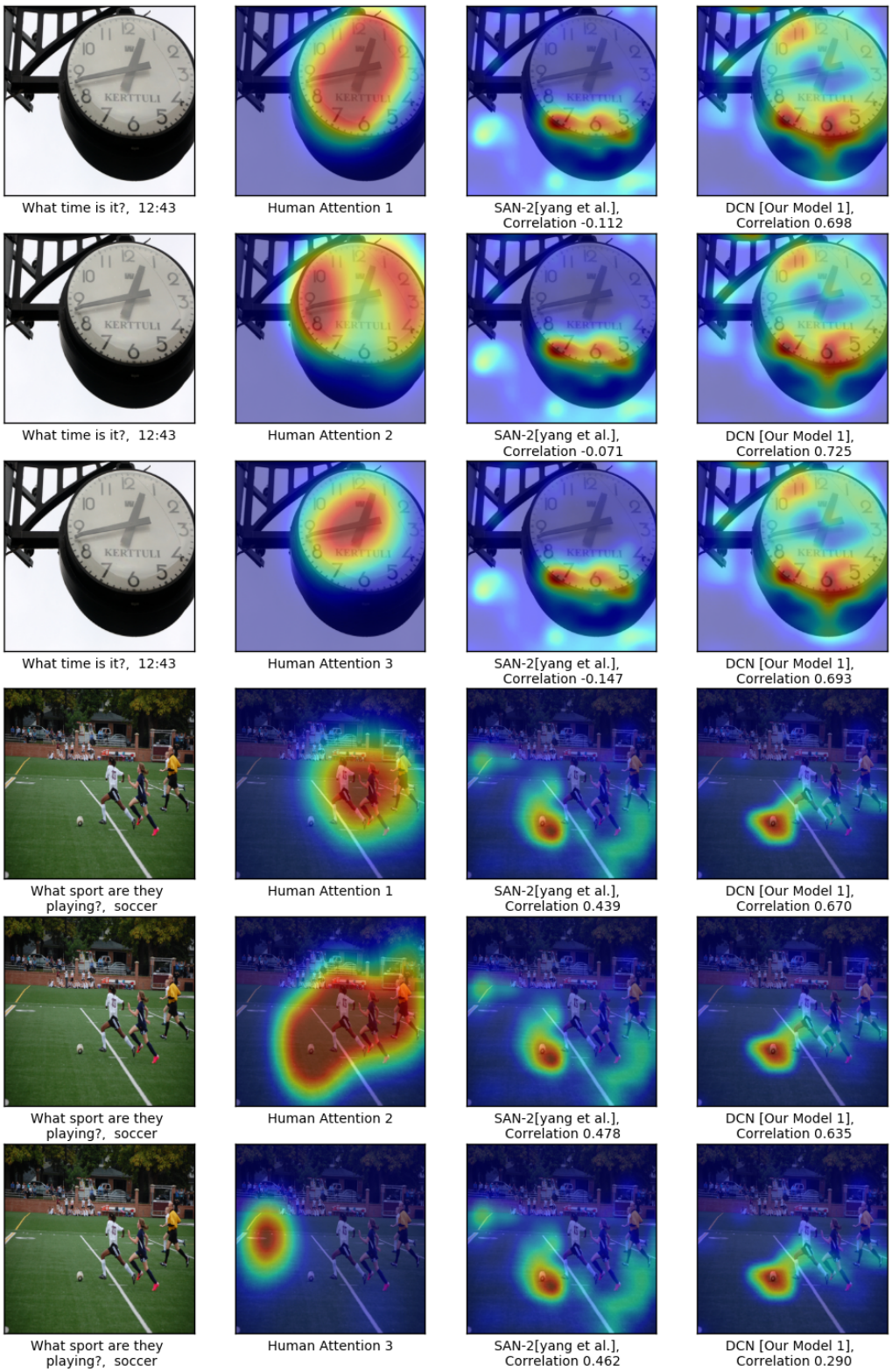

Figure 5. DCN Attention Map with all ground truth three human attentions. The first row provides results for the human attention map-1 for HAT dataset[2]. The second row and third row provides results for the human attention map-2 and human attention map-3. Similar results for another example.

The gradient of loss w.r.t the "Supporting" input f^+:

$$\frac{\partial L}{\partial f^+} = \begin{cases} \Delta L^+, & \text{if } (\alpha + ||f - f^+||_2^2 - ||f - f^-||_2^2) \geq 0 \\ 0, & \text{otherwise.} \end{cases}$$

where $\Delta L^+ = 2 * (f - f^+) \frac{\partial (f - f^+)}{\partial f^+}$

$$\frac{\partial L}{\partial f^+} = \begin{cases} -2(f - f^+), & \text{if } (\alpha + ||f - f^+||_2^2 - ||f - f^-||_2^2) \geq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

The gradient of loss w.r.t the "Opposing" input f^-:

$$\frac{\partial L}{\partial f^-} = \begin{cases} \Delta L^-, & \text{if } (\alpha + ||f - f^+||_2^2 - ||f - f^-||_2^2) \geq 0 \\ 0, & \text{otherwise.} \end{cases}$$

where $\Delta L^- = -2 * (f - f^-) \frac{\partial (f - f^-)}{\partial f^-}$

$$\frac{\partial L}{\partial f^-} = \begin{cases} 2(f - f^-), & \text{if } (\alpha + ||f - f^+||_2^2 - ||f - f^-||_2^2) \geq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

The gradient of loss w.r.t the "Target" input f:

$$\frac{\partial L}{\partial f} = \begin{cases} \Delta L, & \text{if } (\alpha + ||f - f^+||_2^2 - ||f - f^-||_2^2) \geq 0 \\ 0, & \text{otherwise.} \end{cases}$$

where $\Delta L = 2 * (f - f^+) \frac{\partial (f - f^-)}{\partial f} - 2 * (f - f^-) \frac{\partial (f - f^-)}{\partial f}$

$$\frac{\partial L}{\partial f} = \begin{cases} 2(f^- - f^+), & \text{if } (\alpha + ||f - f^+||_2^2 - ||f - f^-||_2^2) \geq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

C.2. Quintuplet Model

Unlike triplet model, In this model we considered two supporting and two opposing image along with target image. we have selected supporting and opposing image by clustering. i.e, The 2000th nearest neighbor is divided into 20 cluster based on the distance from the target image. That is first cluster mean distance is minimum cluster distance from target and 20th cluster mean distance is the maximum cluster distance from the target.

- $a_i = f(s_i)$: The embedding of Target
- $p_i^+ = f(s_i^+)$:The embedding of supporting exemplar from cluster 1
- $n_i^- = f(s_i^-)$:The embedding of opposing exemplar from cluster 20
- $p_i^{++} = f(s_i^{++})$:The embedding of supporting exemplar from cluster 2
- $n_i^{--} = f(s_i^{--})$:The embedding of opposing exemplar from cluster 19

The objective of quintuplet is to bring p_i^+(cluster 1) supporting feature more close to target feature than that of p_i^{++} (cluster 2) supporting feature than that of n_i^{--} (cluster 19) opposing feature than that of n_i^-(cluster 20) opposing feature.

$$\begin{aligned} &D(a_i, p_i^+) + \alpha_1 < D(a_i, p_i^{++}) + \alpha_2 < \\ &D(a_i, n_i^{--}) + \alpha_3 < D(a_i, n_i^-), \\ &\forall (a_i, p_i^+, p_i^{++}, n_i^{--}, n_i^-) \in T \end{aligned} \tag{8}$$

where α_1,α_2,α_3 are the margin between $p_i^+ \& p_i^{++}$, $p_i^{++} \& n_i^{--}$, $n_i^{--} \& n_i^-$ respectively. T is training dataset set, which contain all set of possible quintuplet set.

Objective function for Quintuplet loss[3] is defined as :

$$\min \sum_{i=1}^{N} (\varepsilon_i + \chi_i + \phi_i) + \lambda ||\theta||_2^2 \tag{9}$$

subjected to :
$max(0, \alpha_1 + D(a_i, p_i^+) - D(a_i, p_i^{++})) \leq \varepsilon_i$
$max(0, \alpha_2 + D(a_i, p_i^{++}) - D(a_i, n_i^{--})) \leq \chi_i$
$max(0, \alpha_3 + D(a_i, n_i^{--}) - D(a_i, n_i^-)) \leq \phi_i$
$\forall i, \varepsilon_i \geq 0, \chi_i \geq 0, \phi_i \geq 0$

where $\varepsilon_i, \phi_i, \chi_i$ are the slack variable and θ is the parameter of attention network and λ is a regularizing control parameter.The value of α_1, α_2, α_3 are 0.006, 0.2,0.006 set experimentally.

D. Algorithm for Differential Attention

The algorithm 2 for differential attention illustrates the dimensions of inputs and outputs.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015. 1

[2] A. Das, H. Agrawal, C. L. Zitnick, D. Parikh, and D. Batra. Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions? In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016. 1, 3, 4, 5, 6, 9

[3] C. Huang, Y. Li, C. Change Loy, and X. Tang. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5375–5384, 2016. 7

[4] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 1

Algorithm 2 Attention Mechanism

procedure :MODEL(g_i,g_s,g_c,f_i)
Compute Attention Maps:
    $s_i = ATTENTION\_MAP(g_i, f_i)$, //dimension:1x196
    $s_i^+ = ATTENTION\_MAP(g_s, f_i)$, //dimension:1x196
    $s_i^- = ATTENTION\_MAP(g_c, f_i)$, //dimension:1x196
If DAN Model:
    $Loss\_Triplet = triplet\_loss(s_i, s_i^+, s_i^-)$
    $P_{att} = s_i$, //dimension:1x196
If DCN Model:
    Compute Context: r_i^+, r_i^- as in eq-3,4, //dimension:1x196
    $d_i = s_i \circledast tanh(W_1 r_i^+ - W_2 r_i^-)$, //dimension:1x196
    $P_{att} = d_i$, //dimension:1x196
Compute Img & Ques Attention :
    $V_{att} = \sum_i P_{att}(i) G_{imgfeat}(i)$, //dimension:1x512
    $A_{att} = V_{att} + f_i$, //dimension:1x512
    $Ans = softmax(W_A A_{att} + b_A)$, //dimension:1x1000

---

procedure :ATTENTION_MAP(g_i,f_i)
    g_i:Image feature, //dimension:14x14x512
    f_i: Question feature, //dimension:1x512
Match dimension:
    $G_{imgfeat}$:Reshape g_i to 196x512 :$reshape(g_i)$
    $F_{quesfeat}$: Replicate f_i to 196 times: $clone(f_i)$
Compute Attention Distribution:
    $h_{att} = tanh(W_I G_{imgfeat} \bigoplus (W_Q F_{quesfeat} + b_q))$
    $P_{vec} = softmax(W_P h_{att} + b_P)$, //dimension:1x196
    Return P_{vec}

[5] J. H. McDonald. *Handbook of biological statistics*, volume 2. 2009. 1

[6] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015. 5

[7] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(Feb):207–244, 2009. 5

[8] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 21–29, 2016. 5

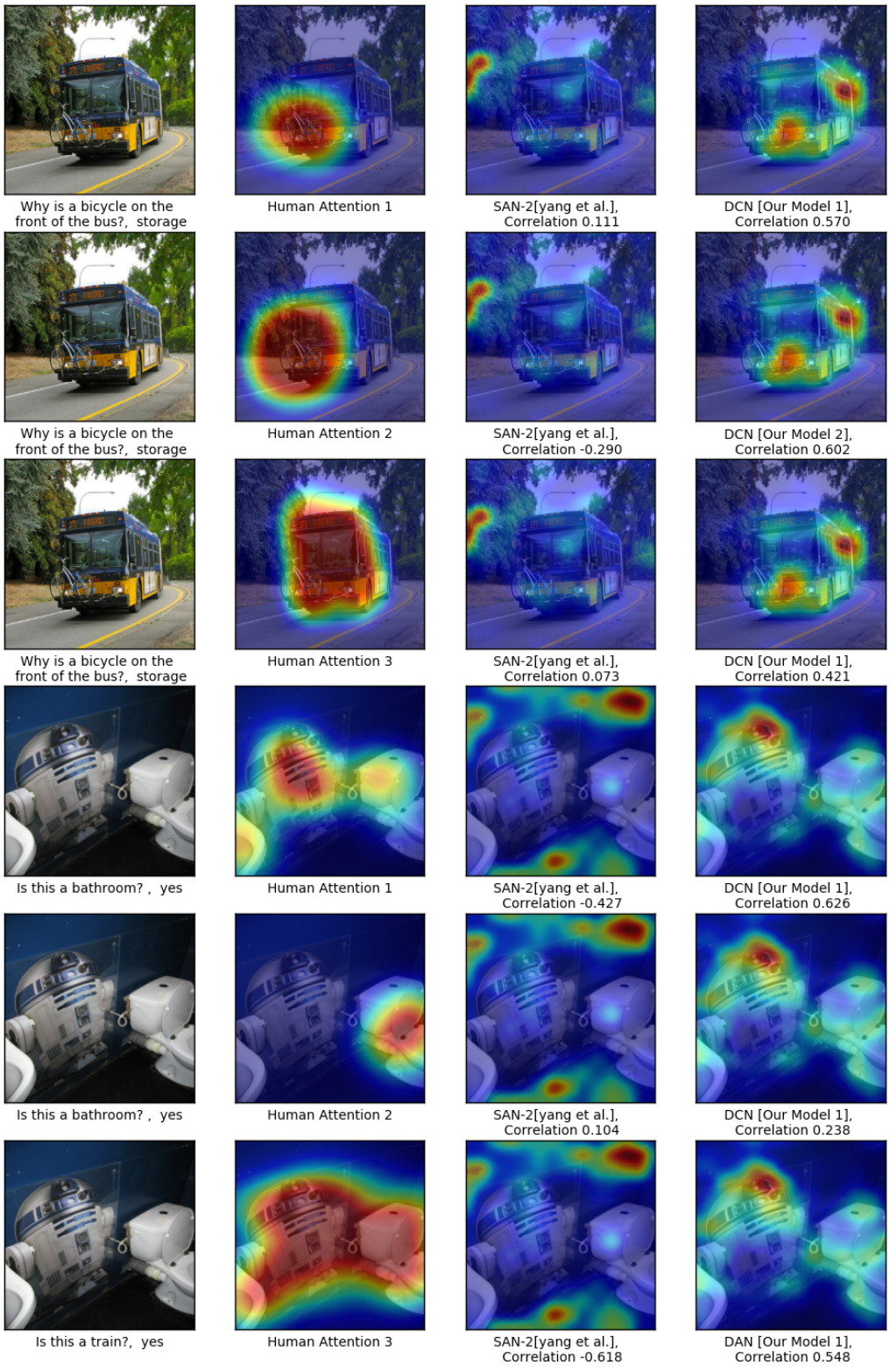

Figure 6. DCN Attention Map with all ground truth three human attentions. The first row provides results for the human attention map-1 for HAT dataset[2]. The second row and third row provides results for the human attention map-2 and human attention map-3. Similar results for another example.

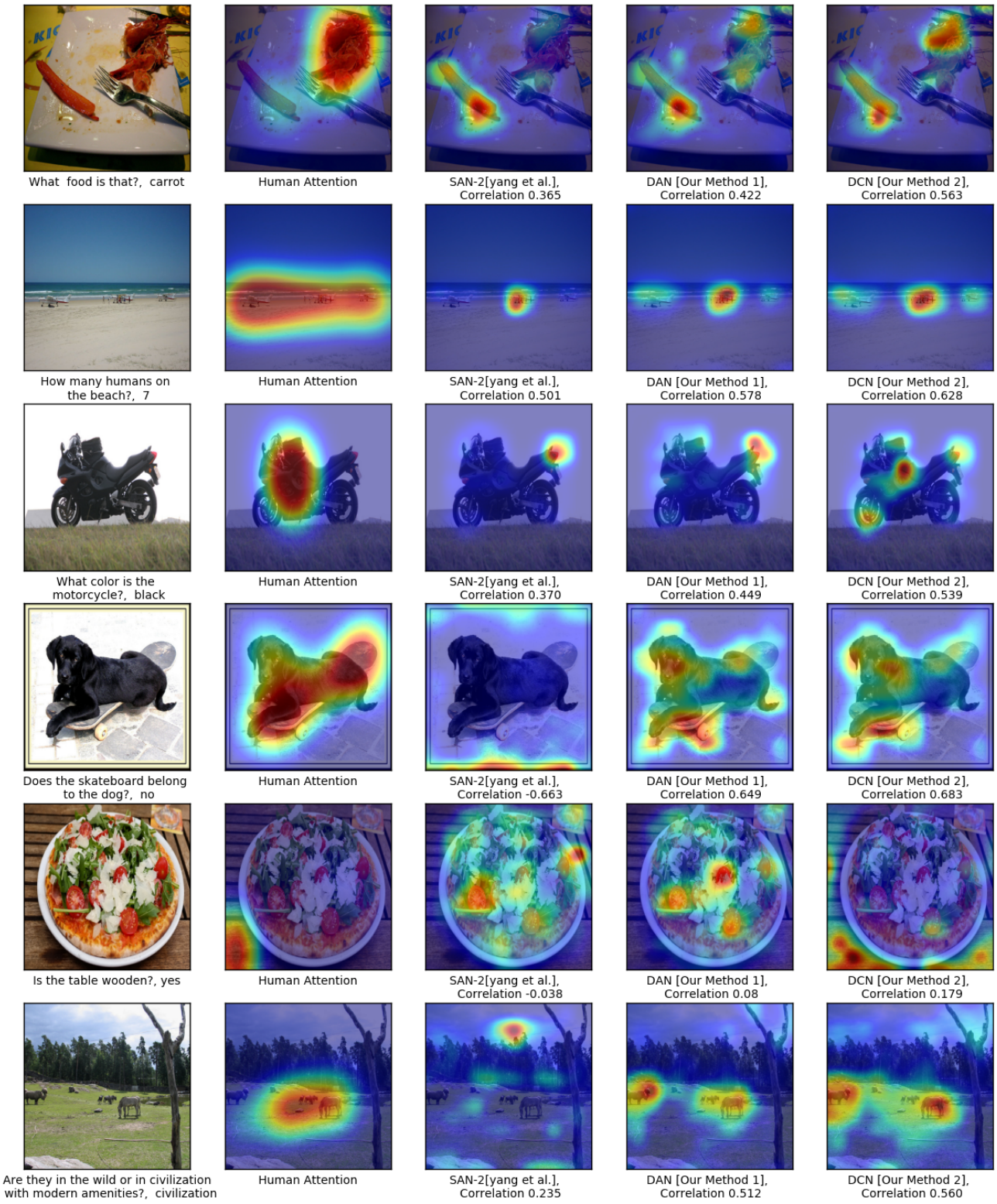

Figure 7. Attention Result for DAN and DCN. In this figure, the first column indicates target question and corresponding image, second column indicates reference human attention map in HAT dataset, third column refer to generated attention map for SAN, fourth column refers to rank correlation of our DAN model and final column refers to rank correlation for our DCN model.

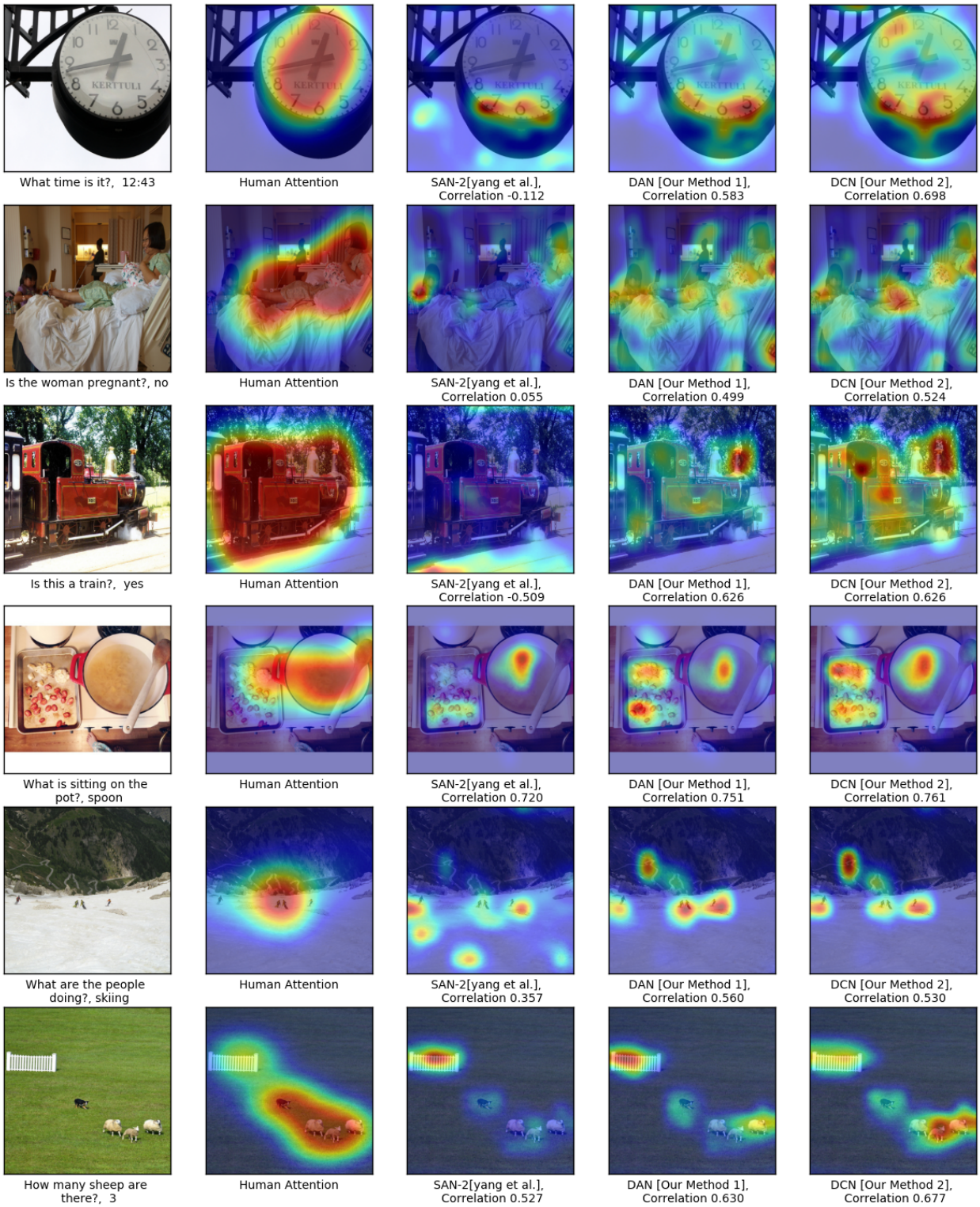

Figure 8. Attention Result for DAN and DCN,In this figure, the first column indicates target question and corresponding image, second column indicates reference human attention map in HAT dataset, third column refer to generated attention map for SAN, fourth column refers to rank correlation of our DAN model and final column refers to rank correlation for our DCN model.