Data Science and Engineering (2020) 5:193–204
https://doi.org/10.1007/s41019-020-00119-z

Meta-supervision for Attention Using Counterfactual Estimation

Seungtaek Choi[1] · Haeju Park[1] · Seung-won Hwang[1]

Received: 16 March 2020 / Accepted: 20 April 2020 / Published online: 13 May 2020

Abstract

Neural attention mechanism has been used as a form of explanation for model behavior. Users can either passively consume explanation or actively disagree with explanation and then supervise attention into more proper values (attention supervision). Though attention supervision was shown to be effective in some tasks, we find the existing attention supervision is biased, for which we propose to augment counterfactual observations to debias and contribute to accuracy gains. To this end, we propose a counterfactual method to estimate such missing observations and debias the existing supervisions. We validate the effectiveness of our counterfactual supervision on widely adopted image benchmark datasets: CUFED and PEC.

Keywords Counterfactual · Attention supervision · Meta-supervision · Event-specific ranking

1 Introduction

Neural attention mechanism has gained interests, due to its contribution toward enhancing both accuracy and explainability. By generating a heatmap over attended regions [1] or highlighting a word of importance [2], the decision of the underlying model can be explained in a human interpretable manner. However, such work treats attention, only as a by-product of prediction or latent variables for explanation [3, 4], while attention coefficients can also be considered as output variables, which can be human supervised.

We study the latter problem of **attention supervision** (**AS**). The existing work suggests that, when such explanation coincides with human perception, accuracy also improves [5–10]. We illustrate our problem with an image attention supervision scenario for event-type annotation [11].

Specifically, given a folder of unannotated personal images, our task is to predict its event type out of E types. Given the first row of images in Fig. 1, the model is tasked to predict its event type ThemePark of the given album. For this prediction, neural attention [4] may identify that the images of a Ferris wheel and an animal highly contribute to the machine prediction. In **AS** problem, human can supervise attention, by giving a scalar importance score for each image in contexts of ThemePark type. CUFED [12] is a dataset annotating such human supervisions, where human annotators are asked to give a scalar importance score for each image for the given event type: For the first row, the image of Ferris wheel and zebra were annotated to be important for detecting ThemePark event, with a high scalar score 1.5, shown as a bar and a number in Fig. 1. Such score is low for the image of sky.

Our key claim is that: CUFED **attention supervision** S of image I is a **biased observation** toward the given event type y. This observation can be debiased if we can observe (or estimate) its counterfactuals: The **unobserved supervision** for image in event $\tilde{y} \neq y$. A closely related problem is obtaining an unbiased relevance estimation [13], from biased click observations to the ranking provided to the user.

One way to debias is to **collect** the counterfactual observations, or online A/B testing. In our problem setting, we may ask annotations of the same image for all E event types, which multiplies annotation overhead E-fold.

In contrast, we propose to **estimate** counterfactual observations to keep human annotation cost as low as **AS**, or offline A/B testing. That is, in Fig. 1, we estimated the dotted distribution of considering the attention **distribution** for all types, where only a value shown in the bar is observed. This distributional attention view (which we name **DistAS**)

✉ Seung-won Hwang
seungwonh@yonsei.ac.kr

Seungtaek Choi
hist0613@yonsei.ac.kr

Haeju Park
phj0225@yonsei.ac.kr

[1] Yonsei University, 50 Yonsei-ro, Seoul, Korea

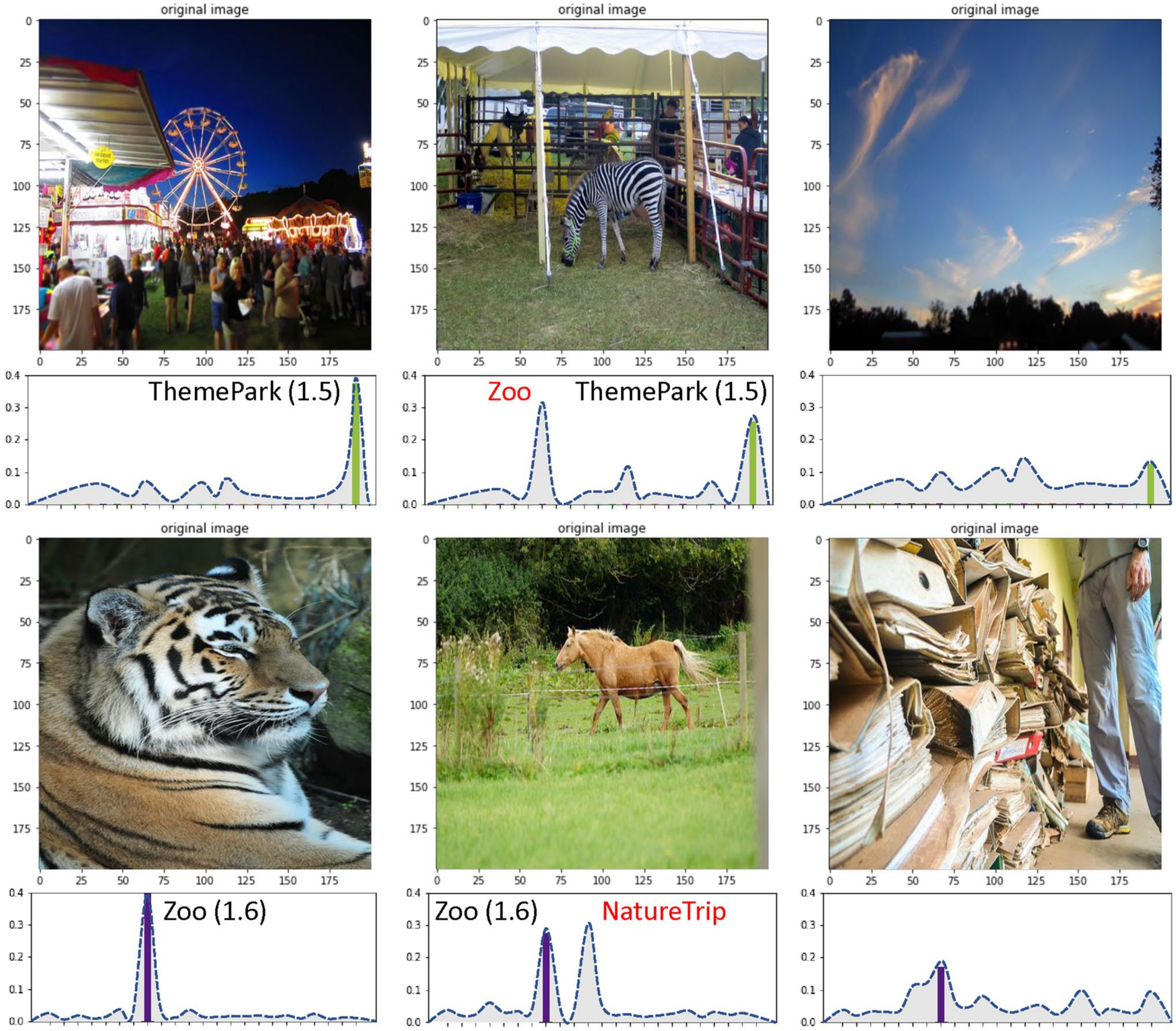


Fig. 1 Ground-truth importance score in the CUFED dataset [12] (shown as a bar) and the estimated counterfactual supervisions (shown as a distribution). The first row is from ThemePark album, and the second is from Zoo album

allows us to realize a bias: Zebra image seems critical for detecting Zoo event, given a high observed value shown in the bar. However, the estimated annotation suggests that this image is relevant to many other types as well and cannot contribute much to conclude the event type. In other word, its importance needs to be debiased into a lower value. This is similar in spirit with propensity weighting [14] that has been a standard approach to correct for item selection bias: interactions are biased to the documents presented at the annotation time.

Our key contribution is to leverage image semantics for propensity weighting: For example, in Fig. 1, to estimate the importance of zebra in Zoo event, we can consider observed annotations for a similar image (such as a horse in the second row) with high weights. If two given images are similar, we force the two images to have similar distribution of supervisions across event types.

We validate the effectiveness of our estimation, by *directly* comparing with human annotation on image importance or *indirectly* by the accuracy of event-type prediction. In both tasks, our proposed model, purposely built upon simple RNN and CNN models, outperforms more complex state of the arts [4, 11, 12], leveraging counterfactual supervisions. Specifically, our proposed models outperform the existing methods by up to 10.6%

point on two personal image benchmark datasets: CUFED and PEC.[1]

This work builds on and extends [15] in the following three ways:

- *Extensive Evaluation* To further isolate the effect of our proposed similarity-based counterfactual estimation, we report the ablation study on CUFED dataset. Our study confirms that all components consistently contribute to the performance improvement (Sect. 5.1).
- *Extensive Survey* We conduct an additional survey over multi-label instances, in comparison with manually identified cases. This also suggests the counterfactual supervisions highly correlate with the human perception about the personal events (Sect. 5.2).
- *Qualitative Examples* Toward better understanding, we add some qualitative examples of our augmented supervisions (Sect. 5.3).

2 Problem Formulation

We aim to solve the task of event-type recognition for a set of unannotated images (album) $A = \{I_1, I_2, \dots, I_T\}$, where T is the number of images. Let $X = \{x_1, x_2, \dots, x_T\}$ denotes the CNN features for each image of the album A, where $x_i \in \mathbb{R}^d$ of feature dimensionality d. For event recognition, we are tasked to train a recognition function $f : \mathbb{R}^{T\times d} \rightarrow \mathbb{R}^E$, which predicts a correct event type $y \in \{y_1, y_2, \dots, y_E\}$ among E event types.

Our goal is to improve the neural attention mechanism by learning attention $\alpha \in \mathbb{R}^T$ to follow the gold importance $S = \{S_1, S_2, \dots, S_T\}$ as closely as possible (we call **attention supervision**), which can be evaluated in the following two ways, by comparing **event-type prediction** and **event-specific image ranking** with the human annotations. For the second evaluation, we regard the attentions as an alternative of importance scoring function $g : \mathbb{R}^{T\times d} \rightarrow \mathbb{R}^T$, employed in the recognition function for weighting purpose $f : \mathbb{R}^{T\times d} \xrightarrow{\alpha} \mathbb{R}^E$.

For the sake of this discussion and without loss of generality, we will consider a decomposition of the recognition network into two functional components—an album feature extractor $X \xrightarrow{\alpha} z$ and a decision network $z \rightarrow \hat{y}$. The former combines the image features X into an album representation z by weighting the image features with attention α. In the latter, the album representation z is used to make event-type prediction $\hat{y}$. Our intention is to keep this decision network as simple as possible to make the point that, with advanced attention supervision, simple models can beat more complex state of the arts. We thus consider simple CNN- and RNN-based models below.

2.1 CNN-Att

Dependent on the event type, the importance of images does vary and more important images should contribute more to the album representation, which can be modeled as neural attention [2, 4]. Specifically, attentions, $\alpha = \{\alpha_1, \alpha_2, \dots, \alpha_T\}$, are computed with a feed-forward network. We model the attentions as a probability distribution over all the images via a softmax layer. Then, z is defined by a weighted sum of image features according to their attentions. We denote this model variant as CNN-Att. Specifically,

$$u_i = \tanh(W[x_i;x_{\text{avg}}] + b), \tag{1}$$

$$\alpha_i = \frac{\exp(u_i^\top e)}{\sum_{j=1}^{T} \exp(u_j^\top e)}, \tag{2}$$

$$z = \sum_{i=1}^{T} \alpha_i \cdot x_i, \tag{3}$$

where x_{avg} denotes the average of all the image features in the given album and [;] means concatenation of the features. W, b and e are learnable parameters. Intuitively, attention will measure the relative importance of an image with regard to the whole album. The context vector e represents a latent query asking for image importance for the given event.

2.2 RNN-Att

Alternatively, some event type has a strong temporal dependence, such that input features are better represented as recurrent models, such as LSTM [16] and GRU [17]. We thus employ bidirectional GRU network into our attention architecture, named RNN-Att. Specifically, input images are first sorted in chronological order and fed into the BiGRU network, and hidden states of the recurrent network are used as input for the attention computation.

$$h_i = \text{BiGRU}(x_i), \tag{4}$$

$$u_i = \tanh(W[h_i; h_{\text{avg}}] + b), \tag{5}$$

$$\alpha_i = \frac{\exp(u_i^\top e)}{\sum_{j=1}^{T} \exp(u_j^\top e)}, \tag{6}$$

[1] Our code is available at https://github.com/hist0613/DistAS.

$$z = \sum_{i=1}^{T} \alpha_i \cdot h_i. \tag{7}$$

The decision network takes the album representation z and predicts log-probabilities over output classes (E event types).

3 Approach

Our next task is to supervise such attentions for accurate prediction of both event type and importance, using public annotations, known as CUFED [12], of gold event label y and event-specific importance S for each album.

3.1 Baseline: ScalarAS

Formally, we design a model to predict event type with minimal error (represented by objective function $\mathcal{L}_{\text{cls}}$), but also event-specific importance (represented as $\mathcal{L}_{\text{ScalarAS}}$).

First, for $\mathcal{L}_{\text{cls}}$, all models are trained with the classification objective, to minimize the categorical crossentropy loss $\mathcal{L}_{\text{cls}}$ between the ground-truth y and predicted event-type label $\hat{y}$.

$$\mathcal{L}_{\text{cls}} = \sum_{\mathcal{A}} -y \ln \hat{y}, \tag{8}$$

where $\mathcal{A}$ denotes the entire albums in the training set.

Second, for $\mathcal{L}_{\text{ScalarAS}}$, the objective is to ensure the distribution of attention α is closer to the target distribution β:

$$\mathcal{L}_{\text{ScalarAS}} = \sum_{\mathcal{A}} \sum_{i=1}^{T} -\beta_i \log \alpha_i \tag{9}$$

Following [12], we focus on the relative importance of each image in the given album, rather than directly predicting the exact importance scores, due to the hardness of learning a reliable absolute importance. We turn the importance scores (i.e., supervisions) into a probability distribution of $\sum_{i=1}^{T} \beta_i = 1$ as follows:

$$\beta_i = \frac{\exp(\lambda S_i)}{\sum_{j=1}^{T} \exp(\lambda S_j)}, \tag{10}$$

where λ is a positive hyper-parameter that controls a score contrast: When the λ increases, the distribution of target attention β becomes more skewed, guiding to attend a few of more important images.

We then set the total loss is the weighted sum of the two loss terms: $\mathcal{L} = \mu_{\text{cls}} \cdot \mathcal{L}_{\text{cls}} + \mu_{\text{AS}} \cdot \mathcal{L}_{\text{ScalarAS}}$, where μ_{cls} and μ_{AS} denote the balancing coefficients between the two terms. We apply this loss function on CNN-Att and RNN-Att, respectively, and denote these variants as CNN-ScalarAS and RNN-ScalarAS.

3.2 Distributional Attention Supervision (DistAS)

This section questions whether CUFED annotation S is an optimal supervision for the attention α. Rather, we propose the supervision vector $S \in \mathbb{R}^T$ should be expanded into a matrix $S^* \in \mathbb{R}^{T \times E}$, to annotate unobserved image importance for other event types as well. That is to say, CUFED annotation can only sparsely supervise for such matrix, by annotating S^*_{iy} for the importance for each image I_i and gold event y, namely biased toward the prediction. The same image is not considered for other types, such that $S^*_{ik} = 0$ where $k \neq y$.

Now, the question is, can we augment zero entries $S^*_{ik} = 0$ for $k \neq y$, with better estimates? Existing frameworks leverage the labeled data from other event types, by inventing Siamese structure looking into multiple types [12], or iterative convergence [11], as implicit data augmentation. Instead, we keep structures simple and augment annotations into S^*_{ik} (replacing 0 with a counterfactual estimation), which we discuss later.

Given the expanded target supervision matrix S^*, our attention supervision goal is formally stated as follows:

$$\mathcal{L}_{\text{DistAS}} = \sum_{\mathcal{A}} \sum_{k=1}^{E} \sum_{i=1}^{T} -\beta^*_{ik} \log \alpha^*_{ik}, \tag{11}$$

where β^* is initialized with $\beta^*_{ik} = \frac{\exp(\lambda S^*_{ik})}{\sum_{j=1}^{T} \exp(\lambda S^*_{jk})}$. Note that we apply a softmax function across the images for each event type, which aims to preserve the observed ranking information within the event type. Because S^*_{ik} is zero-initialized, the softmax yields uniform distribution of $\beta^*_{ik} = \frac{1}{T}$ for $k \neq y$.

To accept the expanded supervisions S^*, our attention architecture needs to be expanded to have multiple context vectors e_k as many as E, intuitively querying “important images for k-th event”. This modification yields event-wise attention weights α^* as follows:

$$\alpha^*_{ik} = \frac{\exp(u_i^\top e_k)}{\sum_{j=1}^{T} \exp(u_j^\top e_k)} \quad . \tag{12}$$

The overall architecture of our proposed model is presented in Fig. 2.

3.3 Counterfactual Supervision Estimation

From the observed importance S^*_{iy}, our goal is to estimate the unobserved importance S^*_{ik} for $k \neq y$ at training time. The zero entries, $S^*_{ik} = 0$, may mean either the image is absolutely unimportant in the given event or important yet unobserved. In contrast to ScalarAS built on only the former assumption, we take the latter assumption by taking the

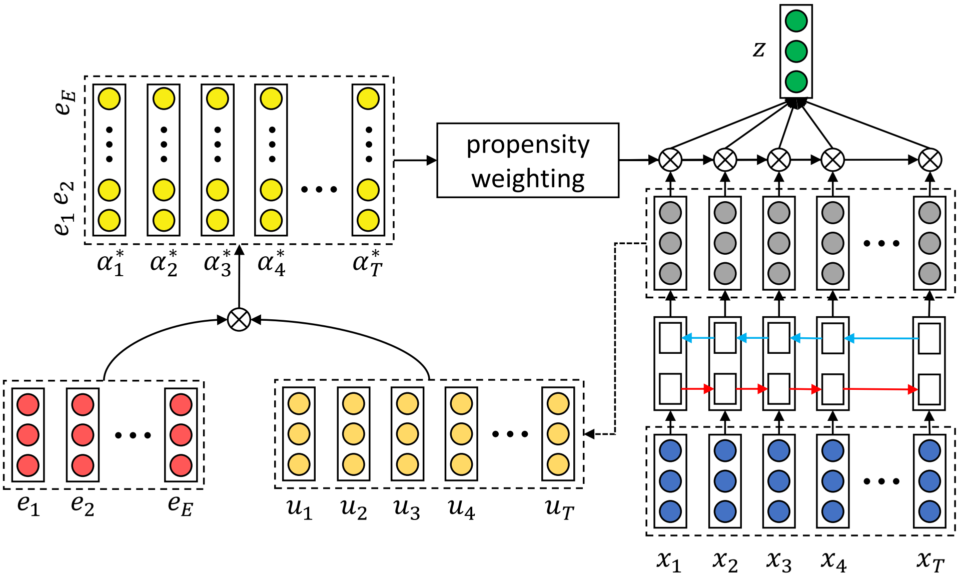

z
propensity weighting
e_E
e_1 e_2
α_1^* α_2^* α_3^* α_4^* α_T^*
e_1 e_2 e_E
u_1 u_2 u_3 u_4 u_T
x_1 x_2 x_3 x_4 x_T

Fig. 2 The overall architecture of RNN-DistAS

missing supervisions S^*_{ik} as optimization variables, which can be estimated by the observed importance in other events.

Inspired by propensity weighting [14], we propose a (propensity-)weighted aggregation of observed importance for debiasing, based on the following intuition: if the two images in different events have similar image features (or, propensity), they have similar importance distributions across multiple events $S^*_i \in \mathbb{R}^E$ (a row vector of matrix S^*). In other words, human annotations on the given image for an unobserved event type $S^*_{i\tilde{y}}$ is close to their annotation on other similar image presented for $S^*_{j\tilde{y}}$.

Formally, we set our goal as to minimize the difference between two different image similarity metrics obtained from image features and importance distributions as follows: $\text{sim}(x_i, x_j) - \text{sim}(\beta^*_i, \beta^*_j)$. In order to efficiently introduce such objective into the existing training process, we additionally sample an album $\tilde{A}$ whose gold event is $\tilde{y}(\neq y)$ and build two matrices of image similarities $M^{\text{feat}} \in \mathbb{R}^{T\times T}$ and $M^{\text{imp}} \in \mathbb{R}^{T\times T}$ by comparing the two albums A and $\tilde{A}$. The $(i,\ j)$-th entry is calculated as $M^{\text{feat}}_{ij} = \text{sim}(x_i, x_j)$ and $M^{\text{imp}}_{ij} = \text{sim}(\beta^*_i, \beta^*_j)$, where j denotes the index of an image in album $\tilde{A}$. In this work, we use cosine similarity as similarity measure, i.e., $\text{sim}(a, b) = \cos(a, b)$.

Meanwhile, the above estimation provides small, yet nonzero scores for dissimilar pairs such as $(x_{\text{elephant}}, x_{\text{mountain}})$, generating noisy supervision. We thus redefine M^{feat} with an introduction of threshold δ, where an entry with value smaller than δ becomes 0:

$$M^{\text{feat}}_{ij} = \begin{cases} \text{sim}(x_i, x_j), & \text{if } \text{sim}(x_i, x_j) > \delta, \\ 0, & \text{otherwise,} \end{cases} \tag{13}$$

where we empirically set the threshold δ to 0.8. Such thresholding allows our model to deal with a poor estimation of feature-based image similarity.

From the two similarity matrices, we define new estimation loss $\mathcal{L}_{\text{sim}}$ as the Frobenius norm of the error matrix $M^{\text{feat}} - M^{\text{imp}}$:

$$\mathcal{L}_{\text{sim}} = ||M^{\text{feat}} - M^{\text{imp}}||_F, \tag{14}$$

where

$$||M||_F = \sqrt{\sum_{i\in[1,T]} \sum_{j\in[1,T]} |M_{ij}|^2}. \tag{15}$$

In summary, our attention module will be jointly trained with the following two objectives:

- $\mathcal{L}_{\text{sim}}$, estimating S^*, which is initially a sparse matrix with many unobserved importances $S^*_{i\tilde{y}}$.
- $\mathcal{L}_{\text{DistAS}}$, supervising α^*_{ik}, attached to CNN or RNN models, to follow the estimated supervision S^*_{ik}.

The entire model is trained with a new loss function: $\mathcal{L} = \mu_{\text{cls}} \cdot \mathcal{L}_{\text{cls}} + \mu_{\text{AS}} \cdot \mathcal{L}_{\text{DistAS}} + \mu_{\text{sim}} \cdot \mathcal{L}_{\text{sim}}$. We introduce an additional coefficient μ_{sim} for $\mathcal{L}_{\text{sim}}$ for balancing β^* estimation in the loss function. Training the expanded attention view α^* with the counterfactual supervisions S^*, we name these models as CNN-DistAS and RNN-DistAS in the later experiments.

3.4 Debiased Ranking from Attention Distribution

In this section, we discuss about how we generate the debiased ranking from our attention distribution. A naive prediction of treating the maximum of the attention distribution as importance score could be inherently biased toward the observed event. We argue that better debiased ranking could be achieved by learning to discount the images, which are important in many event types, but not showing the discriminative parts, like the *zebra* image in Fig. 1.

For evaluation of debiased ranking, we employ the concept of **Inverse Propensity Scoring (IPS)** [18] by giving penalty to the high frequent images across multiple event types (multiple documents). Specifically, event-specific importance, namely **relevance** R_i of the given image I_i in the album A, should be the probability of the image being relevant in *gold* event y, normalized by it being relevant in *other* events $\tilde{y}$, which we define as the *propensity* of the image. It can be estimated with α^* as follows:

$$R_i = \frac{P(R=1|I_i, y)}{P(R=1|I_i, \tilde{y})} \approx \frac{S^*_{iy}}{\sum_j \mathrm{sim}(x_i, x_j) \cdot S^*_{j\tilde{y}}} \approx \frac{\max \alpha^*_i}{1 - \max \alpha^*_i}. \tag{16}$$

In this work, we treat the similarity between the two images (x_i, x_j) as a propensity score of x_i over different events. This architecture design is targeted to inference time, when we are not aware of what the gold event type is. It is not yet guaranteed the maximum attention is of gold event y. However, by maximizing the attention score of gold event in training time, where the target supervision S^*_{iy} is initialized only at gold event, such metric could achieve correct guidance.

Finally, we obtain the album representation z according to the normalized coefficients r_i via a softmax layer:

$$r_i = \frac{\exp(R_i)}{\sum_{j=1}^{T} \exp(R_j)}, \tag{17}$$

$$z = \sum_{i=1}^{T} r_i \cdot h_i. \tag{18}$$

For event-specific ranking, we use the debiased relevance score r_i as the sorting criteria for the image I_i.

4 Experiments

4.1 Dataset

To evaluate the effectiveness of DistAS, we conduct experiments on two public benchmark datasets: CUration of Flickr Events Dataset (CUFED) [12] and Personal Events Collection (PEC) [19], for event recognition and event-specific ranking. Due to no available ranking annotations in PEC, we only report the result for event recognition to show the effectiveness of our counterfactual approach and debiased ranking. PEC dataset could be regarded as an extreme scenario of no human annotation.

4.2 Baselines

We compare the proposed approach **DistAS** with the current state-of-the-art baselines.

- *Siamese-CNN* [12] is trained to predict the difference of importance scores between a pair of images with the piece-wise ranking loss. When evaluation, the output of CNN is used as sorting criteria.
- *Iterative-CNN-LSTM* [11] consists of three different modules: (1) CNN for image-level event recognition, (2) LSTM for album-level event recognition, and (3) Siamese networks for importance prediction from Wang et al. [12]. The same ResNet architecture is used as the base network in module 1 and 3. The prediction is iteratively improved by updating the output of module 1 and 2 with the importance predicted by module 3.

4.3 Model Configuration

Due to the page limitation, we report the hyper-parameter settings in CUFED dataset only. The details in PEC dataset is available with our experiment codes. We use ResNet50 [20] features as the image feature x_i of dimension size 2048. The size of context vector (e and e_k) is set to 128. For recurrent models, the size of hidden states is fixed to 512, yielding 1024 in bidirectional model. The decision network, i.e., the last fully connected layers, contains two feed-forward layers of 300-dimension with 0.2 dropout rate.

Regarding the other hyper-parameters: λ is empirically set to 3.0, making more clear contrast between important and unimportant images. We observe μ_{AS} works differently in two different AS approaches: 0.2 for <u>ScalarAS</u> and 0.8 for <u>DistAS</u>. We posit that such difference stems from that the target attention β^* used in <u>DistAS</u> already contain rich information about gold event label y. For the counterfactual estimation, we observe that 0.01 for μ_{sim} works well. There was unstable training problem when we use larger value for μ_{sim}, such as 0.1 and 1. One possible reason is that the randomly sampled $\tilde{A}$ introduces unnecessary training signals at the beginning of training, before learning useful ranking information.

4.4 Training Details

Following [21], the training is done following the same protocol of extracting multiple subsets from an album, where we extract 16 images (i.e., $T = 16$) over 20 times. To diminish the side effect of such sampling, we report the average performance over 5 runs. We use Adam [22] optimizer with

Table 1 Results of event-specific ranking on CUFED

Model	Precision@K%		
	10	20	30
Random	9.0	19.3	29.8
CNN-Att	15.1	28.4	39.9
RNN-Att	24.4	39.0	50.3
Siamese-CNN	28.1	40.4	49.7
Iterative-CNN-LSTM	30.0	41.3	50.7
CNN-ScalarAS	30.9	48.9	61.0
RNN-ScalarAS	34.4	50.1	63.5
CNN-DistAS	36.6	53.0	63.7
RNN-DistAS	**40.6**	**57.5**	**70.1**

learning rate of 0.001. Models are trained over 50 epochs to ensure convergence of training loss with batch size of 64. All models are evaluated when showing their best ranking performance at validation set.

4.5 Direct Evaluation: Event-Specific Ranking

We begin the assessment of our model with a *direct* evaluation to show the superiority of our model. For this evaluation, we follow the protocol of Wang et al. [11], reporting precision@K% metric, which tells how many images of the highest predicted importance score α_i are ranked in top K% images ordered by the ground-truth importance.

The experimental results are shown in Table 1. Our finding could be summarized as twofold: First, as expected, our attention supervision approaches are better able to rank images than the state-of-the-art baselines. In particular, RNN-DistAS achieves 40.6% at P@10% metric, outperforming the previous state-of-the-art Iterative-CNN-LSTM model by 10.6% point. Notably, we could observe substantial improvement even in the weakest model CNN-ScalarAS among our proposed models, achieving 7.6% at P@20% and 10.3% at P@30%, compared to Iterative-CNN-LSTM. It demonstrates the effectiveness of our problem formulation, employing the supervised attention as internal ranking function.

Second, we manifest the effectiveness of our counterfactual supervisions, particularly at P@10%. CNN-DistAS achieves the 5.7% improvement compared to CNN-ScalarAS, and RNN-DistAS achieves 6.2% point gain over RNN-ScalarAS. It demonstrates that debiasing the importance of images, which are important at multiple events, is essential for selecting the most representative image in the given album.

For further analysis, we show qualitative examples in Fig. 3. We present the top-8 ranked images by each model. As discussed, we can observe that RNN-ScalarAS incorrectly gives high scores for irrelevant images, such as the *flower* image in ARCHITECTURE (more important in NATURETRIP album). Meanwhile, our approach better highlights more discriminative image. Even when the ranked images are not optimally correlated with human-ordered images, RNN-DistAS consistently shows a reasonable ordering, such as the *tiger* image at top-1 in Zoo event, compared to *building* image of RNN-ScalarAS.

4.6 Indirect Evaluation: Album Event Recognition

The main objective of our work is to investigate the impact of counterfactual supervisions. Following the direct evaluation, here we evaluate our model in terms of their contribution to event recognition task. The results of album event recognition on the two datasets are provided in Table 2. From the table, we can observe similar trends with direct evaluation, showing the strength of the debiased ranking in attention mechanism.

Our best performing model RNN-DistAS, reaching an accuracy of 75.7%, shows a 3.4% improvement over the state-of-the-art baseline Iterative-CNN-LSTM in CUFED dataset. At the same time, RNN-DistAS achieves better performance 91.1% than Hierarchical-CNN-Att 90.1%, showing the strength of debiased ranking even in the extreme scenario of no human annotation.

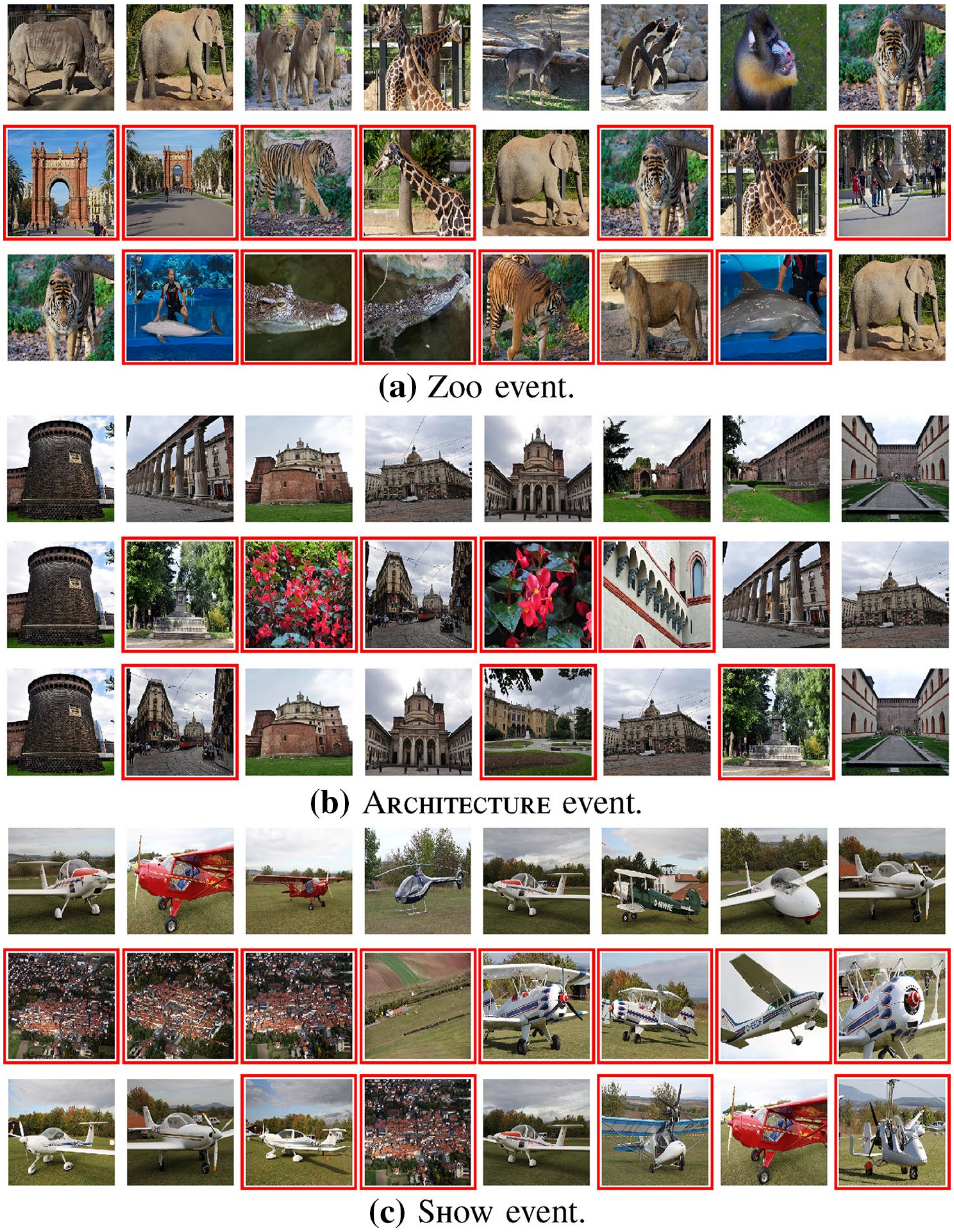
5 Analysis and Discussion

5.1 Ablation Study (CUFED)

To show the effectiveness of each component in DistAS, we conduct an ablation study on CUFED dataset (Table 3). First, the effectiveness of counterfactual estimation $\mathcal{L}_{\text{sim}}$ is tested in model ①. When we removed the $\mathcal{L}_{\text{sim}}$ by setting $\mu_{\text{sim}} = 0$, the performance significantly drops to 36.3 at P@10%. Second, in ②, we simply replace R_i with the maximum value of α_i^*, rather than IDF relevance (Eq. 16). The lower performance of model ② demonstrates that the relevance modeling works in a meaningful way. Lastly, we discard the filter of unnecessary training signals in Eq. 13 by setting the threshold δ to 0. The result of ③ shows that the feature-based image similarity $\mathrm{sim}(x_i, x_j)$ is not the optimal measure, such that simple thresholding could significantly contribute to the performance.

From these results, we have the following observations: (1) all components including counterfactual estimation $\mathcal{L}_{\text{sim}}$, IDF relevance, and threshold consistently contribute to the performance improvement; (2) the counterfactual augmentation is the most important component which leads to more substantial improvement compared to other components.

Fig. 3 Qualitative examples of three different events. For comparison, we present the ranked list of images by (1) ground-truth importance, (2) RNN-ScalarAS, and (3) RNN-DistAS for each event. Red boxes represent the false positive images not included in the top-8 ground-truth images



(a) Zoo event.

(b) Architecture event.

(c) Show event.

Table 2 Results of album event recognition

Model	Accuracy (%)	
	CUFED	PEC
CNN-Att	71.9	86.6
RNN-Att	72.2	87.1
Hierarchical-CNN-Att [4]	–	90.1
Iterative-CNN-LSTM [11]	72.3	–
CNN-ScalarAS	73.3	–
RNN-ScalarAS	73.7	–
CNN-DistAS	75.1	*90.5*
RNN-DistAS	**75.7**	***91.1***

Table 3 Ablation study of event-specific ranking on CUFED dataset

Model		Precision@K%		
		10	20	30
★	RNN-DistAS	**40.6**	**57.5**	**70.1**
①	$\mathcal{L}_{sim}$	36.3	52.2	64.9
②	IDF relevance	37.4	53.9	62.8
③	Threshold $\delta = 0$	38.8	58.6	69.7

Table 4 Comparison between human annotated two-label examples and our counterfactually founded two-label examples

Categories	Event types
Top 10 event types of two-label albums by human [11]	**(PersonalSports, Sports)**, **(UrbanTrip, Architecture)**, (Zoo, NatureTrip), **(Show, PersonalMusicActivity)**, (CasualFamilyGather, GroupActivity), **(Birthday, CasualFamilyGather)**, (Halloween, GroupActivity), **(BeachTrip, Cruise)**, (Show, GroupActivity)
Top 20 event types of two-label images by $\mathcal{L}_{sim}$	**(UrbanTrip, Architecture)**, (Architecture, Museum), (Birthday, Wedding), (ReligiousActivity, Wedding), (Birthday, GroupActivity), **(Birthday, CasualFamilyGather)**, (Birthday, Halloween), (Graduation, ReligiousActivity), (BusinessActivity, Graduation), **(PersonalSports, Sports)**, (Birthday, Christmas), (ThemePark, Zoo), (Christmas, Halloween), (Architecture, ReligiousActivity), (Show, ThemePark), **(Show, PersonalMusicActivity)**, **(BeachTrip, Cruise)**, (Christmas, ThemePark), (ThemePark, UrbanTrip), (PersonalSports, Show)

Each tuple (e_i, e_j) denotes the event-type pair. The overlapping pairs are in **bold**, which represents that our counterfactual estimation is highly correlated with human perception

5.2 Multi-label Analysis on β^*

In this section, we conduct further analysis on the counterfactual estimation, specifically focusing on multi-label cases. Recently, Wang et al. [11] studied the problem of ambiguous album in CUFED dataset, having multiple event types as label (multi-label), for which they manually disambiguate the albums by re-annotating the dataset. Table 4 shows examples of the most frequently appeared event-type pairs of two-label albums, e.g., (Birthday, CasualFamilyGather). For comparison, we find such frequent event-type pairs from two-label images in CUFED training set, by selecting the images of higher gold importance (included in top 10%), but discounted by our IDF relevance ranking (excluded from top 10%), obtained from the counterfactual supervisions.

In Table 4, we can observe that our counterfactually identified event-type pairs are comparable with manually identified pairs (overlapping pairs are marked as **bold**), even though we did not use any human annotations in training. These results show that the counterfactual supervisions highly correlate with the human perception about the personal events. And, we stress that the multi-label characteristics of CUFED dataset could be found automatically by our approach, while that required human efforts in [11].

5.3 Qualitative Analysis on β^*

Figure 4 shows the real examples of counterfactual supervisions, extracted from RNN-DistAS. As discussed above, there are closely related event pairs (e.g., BeachTrip and Cruise), and some images are visually similar, and important on both events (e.g., *swimming* with 1.6 importance). Their visual similarity makes it possible to augment the CUFED annotations in a counterfactual way, represented as the distribution. And, at the same time, it makes IDF relevance effectively decrease the importance of multi-label images, which do not show the discriminative parts of specific event. We found a total of 1088 multi-label images from CUFED dataset, and the significant amount of such images emphasizes the necessity of our counterfactual approach, not requiring human efforts.

6 Related Work

6.1 Attention Supervision

6.1.1 Vision Tasks

This paper raises a bias problem in the existing attention supervision, while previous literature assumes no such bias: Das et al. [23] is the pioneering work that introduces the inconsistency between the attentions of human and machine in Visual Question Answering (VQA) task. Toward plausible (to human insights) attentions, Gan et al. [8] and Yu et al. [10] use human attention annotations, i.e., human gaze, to supervise the attention of neural architecture in vision tasks. However, they incur expensive overheads of human annotations, such that methods for replacing human annotations are explored [7, 9, 24, 25].

Although several work propose to supervise the neural attention for each specific task, to the best of our knowledge, our work is the first to study the augmentation of counterfactual supervisions for providing improved attentions, without increasing annotation overheads on human side.

6.1.2 Language Tasks

This paper studies how to machine-enhance the quality and quantity of human attention supervision. Related concept in language tasks is faithfulness [26, 27], stating that attention weights of unsupervised attention are too poorly correlated with the contribution of each word for machine decision (or, unfaithful). Our work can be considered as a means of enhancing faithfulness with machine self-supervision.

Another related concept is plausibility [28], requiring more expensive human annotations, namely rationale, for

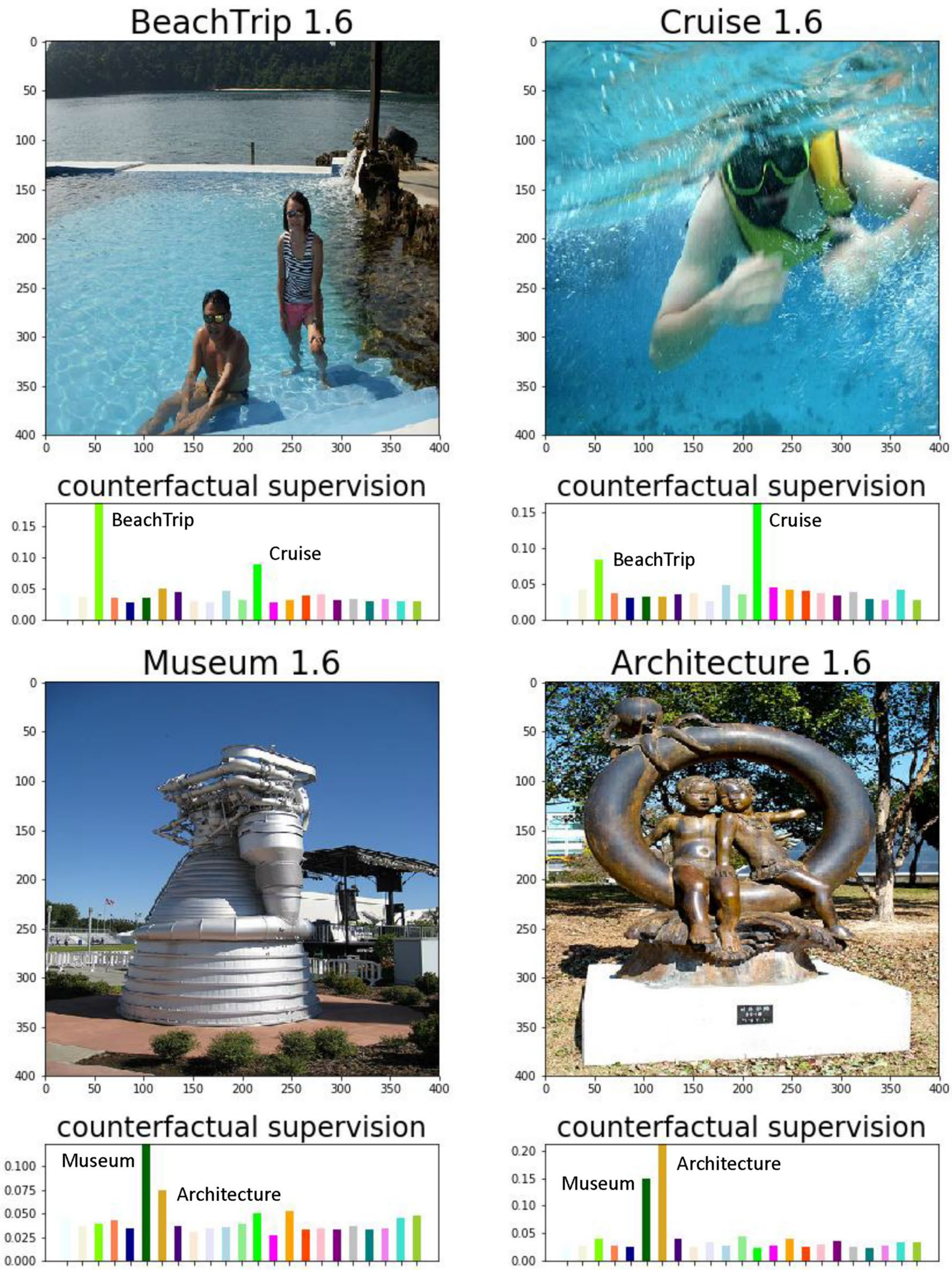


Fig. 4 Qualitative examples of multi-label images and their counterfactual supervisions β^*. Counterpart images are presented in the same row, e.g., BeachTrip and Cruise. Each counterpart event is identified at the second highest score in the counterfactual supervision. The multi-label images across the closely related event pairs have high visual similarity, such as *statue* in Museum and Architecture

sample-specific annotation. Our work can be viewed as an alternative direction of leveraging machine self-supervision and keeping human annotation to vocabularies: Such human annotation overhead can even be replaced by the existing pre-annotated resources: Zou et al. [29] consider sentiment lexicon dictionary such as SentiWordNet, for a related task. Our contribution is to show that simple human annotations (often replaced by public resources), with machine augmentation, can contribute toward improving the accuracy and robustness of model.

There have been several works of using attention supervision for different language tasks. Mi et al. [6] and Liu et al. [5] employ an explicit aligner as an attention prior in machine translation and [30] leverages user authenticated domains to narrow down the scope of attentions. Strubell et al. [31] injects word dependency relations to recognize the semantic roles in text. These supervision mechanisms mainly focused on injecting task-specific knowledge. In contrast, our distinction lies in improving the given attention supervision with sample-specific adaptations.

6.2 Event-Specific Ranking and Recognition

The goal of event recognition is to assign labels (e.g., CasualFamilyGather and Birthday) to the given image or album. With the recent advances for image understanding [20, 32], many event recognition approaches use deep learning models, such as CNN, to capture the semantic of single image (or, multiple images in the album). For example, for representing an album, to effectively combine single image features, a neural attention is introduced by Guo et al. [4], which we adopt as a baseline. A key distinction of our work is, we study the task of supervising such attentions, which would contribute to boosting representation quality.

7 Conclusion

In this paper, we study the problem of counterfactual attention supervision in the personal album recognition and ranking tasks. We propose to augment attention supervision by estimating the missing image importance in the counterfactual events, without additional annotation overheads. This augmented supervision can combine with simple models, improving the event-specific relevance modeling, and outperforms more sophisticated state of the arts.

Acknowledgements This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2020-2016-0-00464) supervised by the IITP (Institute of Information and Communications Technology Planning and Evaluation). Hwang is a corresponding author.

References

1. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y (2015) Show, attend and tell: neural image caption generation with visual attention. In: ICML. pp 2048–2057
2. Yang Z, Yang D, Dyer C, He X, Smola AJ, Hovy EH (2016) Hierarchical attention networks for document classification. In: HLT-NAACL. pp 1480–1489
3. Yu L, Bansal M, Berg T (2017) Hierarchically-attentive RNN for album summarization and storytelling. In: EMNLP. pp 977–982
4. Guo C, Tian X, Mei T (2018) Multigranular event recognition of personal photo albums. IEEE Trans Multimed 20(7):1837–1847
5. Liu L, Utiyama M, Finch A, Sumita E (2016) Neural machine translation with supervised attention. In: COLING. pp 3093–3102
6. Mi H, Wang Z, Ittycheriah A (2016) Supervised attentions for neural machine translation. In: EMNLP. pp 2283–2288
7. Liu C, Mao J, Sha F, Yuille AL (2017) Attention correctness in neural image captioning. In: AAAI. pp 4176–4182
8. Gan C, Li Y, Li H, Sun C, Gong B (2017) Vqs: Linking segmentations to questions and answers for supervised attention in VQA and question-focused semantic segmentation. In: Proceedings of the IEEE international conference on computer vision. vol 3
9. Qiao T, Dong J, Xu D (2017) Exploring human-like attention supervision in visual question answering. arXiv preprint arXiv:1709.06308
10. Yu Y, Choi J, Kim Y, Yoo K, Lee S-H, Kim G (2017) Supervising neural attention models for video captioning by human gaze data. In: CVPR. pp 2680–29
11. Wang Y, Lin Z, Shen X, Mech R, Miller G, Cottrell GW (2017) Recognizing and curating photo albums via event-specific image importance. In: BMVC
12. Wang Y, Lin Z, Shen X, Mech R, Miller G, Cottrell GW (2017) Event-specific image importance. In: CVPR. pp 4810–4819
13. Agarwal A, Takatsu K, Zaitsev I, Joachims T (2019) A general framework for counterfactual learning-to-rank. In: ACM conference on research and development in information retrieval (SIGIR)
14. Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. Biometrika 70(1):41–55
15. Choi S, Park H, Hwang S (2019) Counterfactual attention supervision. In: 2019 IEEE international conference on data mining (ICDM). IEEE, pp 1006–1011
16. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
17. Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. arXiv
18. Jagerman R, Oosterhuis H, de Rijke M (2019) To model or to intervene: a comparison of counterfactual and online learning to rank from user interactions

19. Bossard L, Guillaumin M, Van Gool L (2013) Event recognition in photo collections with a stopwatch HMM. In: ICCV
20. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: CVPR. pp 770–778
21. Wu Z, Huang Y, Wang L (2015) Learning representative deep features for image set analysis. IEEE Trans Multimed 17(11):1960–1968
22. Kingma DP, Ba JL (2014) Adam: A method for stochastic optimization. In: Proceedings of the 3rd international conference on learning representations
23. Das A, Agrawal H, Zitnick L, Parikh D, Batra D (2017) Human attention in visual question answering: Do humans and deep networks look at the same regions? Comput Vis Image Underst 163:90–100
24. Zhang Y, Niebles JC, Soto A (2018) Interpretable visual question answering by visual grounding from attention supervision mining. arXiv preprint arXiv:1808.00265
25. Wang Z, Liu X, Chen L, Wang L, Qiao Y, Xie X, Fowlkes C (2018) Structured triplet learning with pos-tag guided attention for visual question answering. arXiv preprint arXiv:1801.07853
26. Jain S, Wallace BC (2019) Attention is not explanation. arXiv preprint
27. Serrano S, Smith NA (2019) Is attention interpretable? arXiv preprint
28. Zhong R, Shao S, McKeown K (2019) Fine-grained sentiment analysis with faithful attention. arXiv preprint
29. Zou Y, Gui T, Zhang Q, Huang X (2018) A lexicon-based supervised attention model for neural sentiment analysis. In: Proceedings of the 27th international conference on computational linguistics. pp 868–877
30. Kim J-K, Kim Y-B (2018) Supervised domain enablement attention for personalized domain classification. In: EMNLP
31. Strubell E, Verga P, Andor D, Weiss D, McCallum A (2018) Linguistically-informed self-attention for semantic role labeling. In: Proceedings of the 2018 conference on empirical methods in natural language processing. pp 5027–5038
32. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556