

# Faithful Attention Explainer: Verbalizing Decisions Based on Discriminative Features<sup>\*</sup>

Yao Rong<sup>1,\*</sup>, David Scheerer<sup>2</sup> and Enkelejda Kasneci<sup>1</sup>

<sup>1</sup>Technical University of Munich, Arcisstraße 21, 80333 Munich, Germany

<sup>2</sup>University of Tübingen, Sand 14, 72076 Tübingen, Germany

## Abstract

In recent years, model explanation methods have been designed to interpret model decisions faithfully and intuitively so that users can easily understand them. In this paper, we propose a framework, Faithful Attention Explainer (FAE), capable of generating faithful textual explanations regarding the attended-to features. Towards this goal, we deploy an attention module that takes the visual feature maps from the classifier for sentence generation. Furthermore, our method successfully learns the association between features and words, which allows a novel attention enforcement module for attention explanation. Our model achieves promising performance in caption quality metrics and a faithful decision-relevance metric on two datasets (CUB and ACT-X). In addition, we show that FAE can interpret gaze-based human attention, as human gaze indicates the discriminative features that humans use for decision-making, demonstrating the potential of deploying human gaze for advanced human-AI interaction.

## Keywords

Explainable AI (XAI), Saliency Map, Faithfulness, Visual Explanation, Textual Explanations

## 1. Introduction

Explainable AI (XAI) models are being used more, especially in safety-critical applications such as automatic medical diagnosis [1, 2, 3]. An explanation of a decision should be understandable for humans [4], and include objects or features that are responsible for that decision made by a model, i.e., faithful to the model decision [5, 6, 7]. In image-based applications, two modalities are typically used in model explanations: visual and textual explanation [8]. Several related works in this context [9, 10, 11, 12, 13] reveal discriminative (salient) areas for the neural network in decision-making by means of saliency maps. Such saliency maps visualize the post-hoc attention of a deep neural network. However, humans often prefer textual justifications of model decisions since they allow for easier access to the understanding of the causality provided by models [6, 14]. In this work, we introduce a novel method, “Faithful Attention Explainer” (FAE), which generates faithful textual explanations according to the decision made by the classifier.

---


*Late-breaking work, Demos and Doctoral Consortium, colocated with The 2nd World Conference on eXplainable Artificial Intelligence: July 17–19, 2024, Valletta, Malta*

<sup>\*</sup>You can use this document as the template for preparing your publication. We recommend using the latest version of the ceurart style.

<sup>\*</sup>Corresponding author.

✉ yao.rong@tum.de (Y. Rong); david.scheerer@student.uni-tuebingen.de (D. Scheerer); enkelejda.kasneci@tum.de (E. Kasneci)

ORCID 0000-0002-6031-3741 (Y. Rong); 0000-0003-3146-4484 (E. Kasneci)

 © 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

As shown by an example in Figure 1, the explanation of our model includes the object “skateboard,” which is used for the action classification (shown in GradCAM [15]). When we give the GradCAM as the extrinsic attention, the model describes more of the area, such as “standing on a skateboard” and “going down a flight of stairs.” Similarly, human attention also conveys the potential to explain our decisions [16]. It is visualized in the saliency map style and compared to models’ post-hoc attention maps in solving visual question answering and classification tasks [17, 18]. In this context, the language model should also be able to generate a faithful explanation based on human attention. Providing human attention interpretation can help study the human attention mechanism and better integrate it into computer vision applications. To summarize, this work proposes a novel framework, FAE, which generates faithful textual explanations based on attention maps (from models or humans).

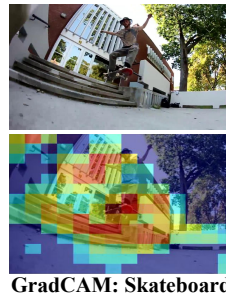


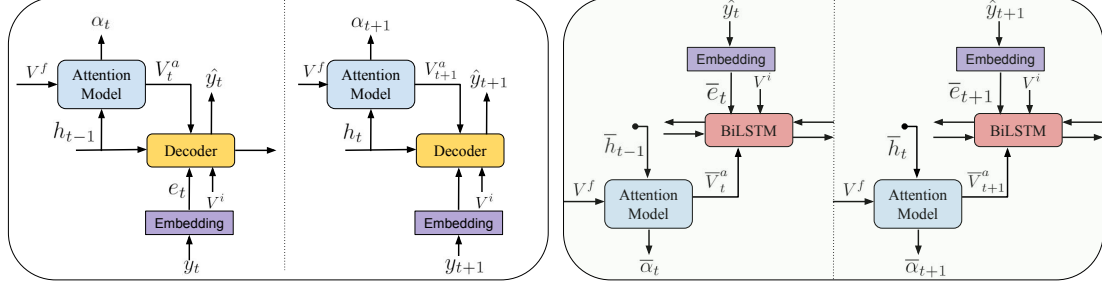
Figure 1: FAE generates faithful explanations (Top). Using attention enforcement, FAE generates a sentence further explaining the attended-to area in GradCAM (Bottom).

## 2. Related Work

Attention models for generating textual descriptions are known to be highly effective [19, 20, 21, 22, 23]. For example, [20] proposes an attention model consisting of linear layers to localize the relevant area in the image for sentence generation. However, the attention model grounds the current word to a wrong region since its current hidden state contains only information of past words [19]. To solve this problem, [24, 25] use extra supervision for correct visual grounding is therefore needed, while [19] proposes the Prophet Attention model which takes both future and past words into account and recreates attention weights and thus does not require extra supervision. Inspired by the PA model, we incorporate future words (generated after the current word) to ground the current word in the image in our attention model. Generating faithful explanations for classifiers is more than image captioning [6, 5] since the generated sentence must rationalize the decision and include discriminative features for the distinctive output class. To generate sentences conditioned on classifiers, previous works [26, 6, 5, 8, 14] use features from the corresponding classifier and feed them into an LSTM layer to generate textual explanations. However, these explanations may not be faithful to each sample since they are trained to be discriminative on class-level and thus can generate features that are not visible in that image [5]. Going beyond previous work, our framework utilizes an attention module for word grounding directly.

## 3. Methodology

Our FAE generates textual explanations for image classifiers, i.e., FAE verbalizes classification decisions by creating sentences containing words related to image regions that have been



**Figure 2:** Overview of Faithful Attention Explainer. The encoder is omitted for simplicity but the output features  $V^f$  and  $V^i$  from the encoder are denoted. The embedding layer is used to transform words into embeddings. **Left:** the attention model and decoder are illustrated. The attention model produces attention  $\alpha$  based on the previous sequence. **Right:** the attention alignment is used to produce  $\hat{\alpha}$  based on the generated sequence  $\hat{y}_{t:T}$ , which tries to align  $\alpha$  with  $\hat{\alpha}$ .

important to the decision of the classifier. In this section, we explain the details of each module in FAE and introduce the Attention Enforcement algorithm in detail. Our network approach follows an Encoder-Decoder framework. The goal of FAE is to take the image  $x \in \mathbb{R}^{H \times W \times C}$  and to predict the class label as well as to create a textual explanation  $\hat{y}$  as a sequence of 1-of-N words:

$$\hat{y} = \{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_T\}, \quad \hat{y}_t \in \mathbb{R}^N \quad (1)$$

where  $T$  denotes the length of the output and  $y_t$  is the predicted word at  $t$  step. FAE exploits the class-discriminative feature vector  $V \in \mathbb{R}^{h \times w \times c}$  from the classifier (also used as the encoder  $\Sigma(\cdot)$ ).  $\Sigma$  is a deep convolutional neural network and can extract several visual feature vectors  $V$  from different layers of the input image  $x$ . Taking ResNet101 as an example,  $V^f$  is the feature map after the last residual block, while  $V^i$  can be a set of feature maps taken from the final layer of the first, second, and third blocks. For each step  $t$ , the attention model  $f_{Att}(\cdot)$  computes attention maps  $\alpha_t$  based on the decoder's (an LSTM model) hidden state  $h_{t-1}$  and the feature vector from the encoder. The output of the attention module  $V_t^a$  is given to the decoder and guides it towards important areas relevant for the explanation: the attention-weighted average of focus features  $V_t^a$  is:

$$V_t^a = \frac{1}{K} \sum_{j=1}^K \alpha_{t,j} V_j^f. \quad (2)$$

Figure 2 (Left) illustrates this architecture that contains the attention module for generating textual explanations. We follow the method proposed in [20] to build and train this model. As the attention model computes weights based on the previous hidden state of the LSTM, which is generated using the previous input word. As a result, the attention weights are also based on the previous word. To tackle this challenge, we introduce a module called attention alignment. Inside the module, we make use of future knowledge (words) to adjust the attention map for the current word. To do so, a Bidirectional LSTM (BiLSTM)[27] is employed to encode the generated sequence. The attention model described in the last section is used to regenerate new attention weights  $\bar{\alpha}_t$  based on the hidden state  $\bar{h}_{t-1}$  from the BiLSTM. Specifically, we get  $\bar{h}_{t-1}$  by concatenating the hidden states from forward and backward paths (and halving the

dimension). Figure 2 (Right) illustrates the attention alignment.

$$\begin{aligned}\bar{\alpha}_t &= f_{Att}(\bar{h}_{t-1}, V^f) \\ \bar{V}_t^a &= \frac{1}{K} \sum_{j=1}^K \bar{\alpha}_t V_j^f\end{aligned}\quad (3)$$

As a regularization to the training loss, we use the L1 norm between the newly grounded attention weights  $\bar{\alpha}$  and the ones generated by the attention model  $\alpha$ :

$$\mathcal{L}_\alpha(\theta) = \sum_{t=1}^T \|\alpha_t - \bar{\alpha}_t\| \quad (4)$$

Moreover, the learned attention can be given by users, i.e., by replacing attention weights by other attention maps  $\epsilon$ , e.g., GradCAM or human gaze, during inference. We refer this as Attention Enforcement (AE). Concretely, we generate the focus feature  $V^\epsilon$ :

$$V^\epsilon = \frac{1}{K} \sum_{j=1}^K \text{Softmax}(\epsilon) V_j^f \quad (5)$$

## 4. Experiments

**Metrics.** To evaluate and compare our model with other works, we use the following metrics: BLEU-4, ROGUE-L, METEOR, CIDEr. These metrics measure the similarity between generated sentences and their ground-truth. However, they only indicate the sentence quality on a linguistic level but have no insights into the faithfulness of generated explanations. Therefore, we measure the Faithful Explanation Rate (*FER*) in generated explanations compared to ground-truth sentences, inspired by [5]. Specifically, for an image  $x$ , discriminative visual regions used in the model’s decision are found out with the help of GradCAM [15]. Using the part annotations, the decision-related part/object  $y_o$  can be identified (the part that is closest to the maximum value in GradCAM). Noun-phrases of that part in all ground-truth sentences are extracted to form a set  $\{\mathbf{g}_1, \mathbf{g}_2, \dots, \mathbf{g}_M\}$  where  $\mathbf{g}_i$  denotes for a noun-phrase. For the generated sequence  $\hat{\mathbf{y}}$ , we detect whether the  $y_o$  is in  $\hat{\mathbf{y}}$ , if not, the hit rate is 0. If yes, we detect the corresponding noun-phrase  $\hat{\mathbf{g}}$ . Then we compare the word hit rate of  $\hat{\mathbf{g}}$  with all possible  $\mathbf{g}_i$  and use the best one for the FER score.

**Datasets.** We use two datasets for our experiments: the CUB-200-2011 dataset (CUB) and Action Explanation Dataset (ACT-X). CUB contains 11.788 images of birds distributed across 200 species [28]. Each image has ten explanations of the visual appearance collected by [29]. ACT-X [8] has 397 classes of activities and in total 18030 images selected from [30]. For each image, three explanations are provided. We follow the provided train and test splits on both datasets. When evaluating the FER score, we use the part annotations on CUB and object-level annotations on ACT-X. The object-level annotation on ACT-X denoted as MPII-ANO, only contains a few images in ACT-X (150 images with 600 object classes) provided by [5].

Dataset	Method	Backbone	BLEU-4	METEOR	CIDer
CUB	GVE [6]	VGG	-	29.20	56.70
	InterpNET [26]	VGG	<b>62.30</b>	37.90	<b>82.10</b>
	SAT	ResNet-101	57.14	36.71	61.80
	FAE (Ours)	ResNet-50	57.94	36.33	55.98
	FAE (Ours)	ResNet-101	60.19	<b>38.13</b>	66.36
ACT-X	GVE [6]	VGG	12.90	15.90	12.40
	PJ-X [8]	ResNet-152	24.50	21.50	58.70
	SAT [20]	ResNet-101	25.63	24.53	50.39
	FAE (Ours)	ResNet-50	26.66	24.37	57.19
	FAE (Ours)	ResNet-101	<b>27.06</b>	<b>25.33</b>	<b>66.17</b>

Method	CUB	MPII-ANO
SAT [20]	37.43	26.32
FAE (Ours)	39.42	28.40
SAT-AE [20]	38.54	26.84
FAE-AE (Ours)	<b>44.33</b>	<b>29.76</b>

**Table 1**

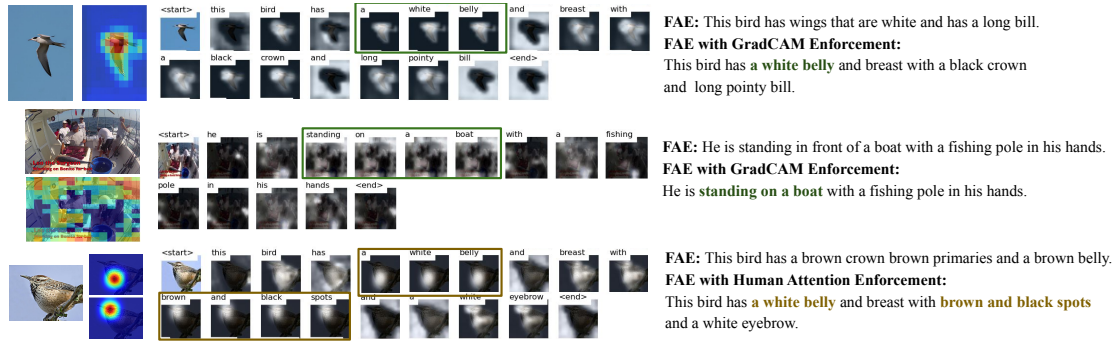
**Left:** Comparison with other methods on CUB and ACT-X in standard sentence quality metrics. **Right:** FER score on CUB and MPII-ANO. The first block contains methods without Attention Enforcement (AE); the second block with using AE. ResNet101 is used as the backbone for all models.

### 4.1. Quantitative Results

We first compare our model with other state-of-the-art approaches in the linguistic quality of generated explanations. In table 1 (Left), we compare our FAE using two backbones with InterpNET [26], Generating Visual Explanations (GVE) [6], and Pointing and Justification Explanation (PJ-X) model [8]. On CUB, our model (using ResNet101 backbone) outperforms GVE, e.g., in the metric CIDer, our model achieves 66.36 while GVE achieves 56.70. Compared to InterpNET, however, our model surpasses only in METEOR. The possible reason is that InterpNET deploys richer features (8192-dim compact bilinear features), two extra hidden layers, and two stacked LSTM layers, which introduces more computational costs and makes the results hard to reproduce. Results on ACT-X are shown in the second block. Our model (ResNet101) achieves higher scores in all three metrics than other methods. Besides the linguistic quality, FER score are shown in table 1 (Right). We compare our framework with SAT since Attention Enforcement (AE) can also be applied to it. For a fair comparison, we evaluate both under the same settings. In the first block, where no AE is used, FAE achieves the best performance: 39.42 on CUB and 28.40 on MPII-ANO, which validates that our FAE is advanced in faithful explanation generation. When using GradCAM attention enforcement, SAT and FAE both improve the FER scores, while FAE surpasses SAT on both datasets. The improvement of using AE in both models validates the generalization of AE.

### 4.2. Qualitative Results

We give GradCAM maps as extrinsic attention maps to guide the model FAE with AE to focus on the area highlighted in the attention map. Two generated sentence examples are illustrated in Figure 3. After applying the enforcement in the first example, the explanation incorporates the part “a white belly”, which is missing before. Nevertheless, when enforcement on the MPII-ANO dataset, the effects are others. Since the GradCAM highlights a lot of area on the boat and in the background (on the sea), the sentence after the enforcement describes the relation between objects correctly: the man is standing “on the boat” instead of “in front of a boat”. The results show that our FAE can provide explanations that are faithful and human-understandable to

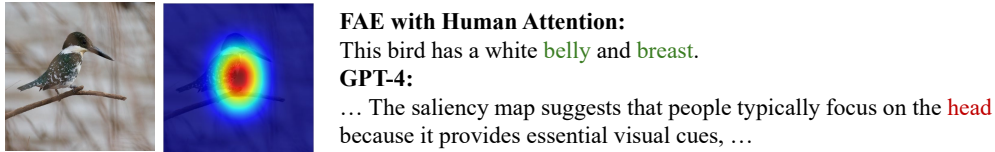


**Figure 3:** Illustration of using attention enforcement on CUB and MPII-ANO. **Left:** Images and extrinsic saliency maps are shown. **Middle:** Frames denote the step where enforcement is activated. **Right:** Sentences generated by FAE with and without attention enforcement. The top two examples use GradCAM from the classifier as extrinsic attention maps, while the bottom one uses human gaze maps.

not only intrinsic but also extrinsic attention maps. Additionally, we try a different source of extrinsic attention for AE: Human Attention (HA). We evaluate our HA-enforcement on the CUB test set and use the HA map provided in CUB Gaze-based Human Attention (CUB-GHA) [18]. This dataset is built by tracking the eye fixations of humans while presenting them of a bird to focus on distinctive features for that species. For each image, there are always multiple attention maps and each attention map represents an eye fixation. In Figure 3, the bottom example shows the HA attention maps. When we deploy our AE on using HA as extrinsic attention information, the sentence describes the two areas: “a white belly” in the first fixation area and “breast with brown and black spots” in the second attention area. This setting confirms that our method can produce accurate textual explanations focusing on user attention, demonstrating the generalizability of our proposed framework.

## 5. Discussion

Large Language Models (LLMs), such as the GPT series, have demonstrated their sophisticated abilities in understanding and generating explanations. Recent advancements enable these models to analyze multimodal data. For example, the GPT-4 model can create textual explanations from an input image. To evaluate its effectiveness, we tested the GPT-4 model with two types of images: an original image and a saliency map highlighting human attention, as illustrated in Figure 4. The GPT-4 successfully generated an analysis of the areas most salient to human gaze. However, we observe the problem in the generated textual explanations: the model fails to correctly identify the area where the user focused. For example, it mistook the belly/breast area as the head. These mistakes rather demonstrate a common weakness in the model: hallucination. To harvest the power of language models, we consider for future work fine-tuning a smaller general language model to generate textual explanations based on the areas of gaze attention of users. This approach can enhance the possibilities for intuitive and direct interaction between humans and AI systems through gaze-based communication.



**Figure 4:** Comparison of our method and GPT-4 in generating textual explanations.

## 6. Conclusion

In this paper, we propose a novel framework FAE that can generate decision explanations faithful to intrinsic attention, i.e., generated by an attention model based on visual features from the classifier. Our results on the CUB and ACT-X datasets validate and confirm the high faithfulness and quality in explanations provided by FAE. Moreover, we extend FAE by using Attention Enforcement and can thus interpret extrinsic attention e.g., human attention. For future work, our method expands opportunities for natural and straightforward communication between humans and AI systems via gaze-driven interactions.

## References

- [1] H. H. Pham, T. T. Le, D. Q. Tran, D. T. Ngo, H. Q. Nguyen, Interpreting chest x-rays via cnns that exploit hierarchical disease dependencies and uncertainty labels, *Neurocomputing* (2021).
- [2] E. Tjoa, C. Guan, A survey on explainable artificial intelligence (XAI): towards medical XAI, *CoRR* (2019). URL: <http://arxiv.org/abs/1907.07374>. arXiv:1907.07374.
- [3] Y. Rong, N. Castner, E. Bozkir, E. Kasneci, User trust on an explainable ai-based medical diagnosis support system, *arXiv preprint arXiv:2204.12230* (2022).
- [4] Y. Rong, T. Leemann, T.-T. Nguyen, L. Fiedler, P. Qian, V. Unhelkar, T. Seidel, G. Kasneci, E. Kasneci, Towards human-centered explainable ai: A survey of user studies for model explanations, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).
- [5] S. Wickramanayake, W. Hsu, M. Lee, Flex: Faithful linguistic explanations for neural net based model decisions, in: *AAAI*, 2019.
- [6] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, T. Darrell, Generating visual explanations, *CoRR* (2016). URL: <http://arxiv.org/abs/1603.08507>. arXiv:1603.08507.
- [7] Y. Rong, T. Leemann, V. Borisov, G. Kasneci, E. Kasneci, A consistent and efficient evaluation strategy for attribution methods, in: *International Conference on Machine Learning*, PMLR, 2022, pp. 18770–18795.
- [8] D. H. Park, L. A. Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, M. Rohrbach, Multimodal explanations: Justifying decisions and pointing to the evidence, *CoRR* (2018). URL: <http://arxiv.org/abs/1802.08129>. arXiv:1802.08129.
- [9] V. Petsiuk, A. Das, K. Saenko, Rise: Randomized input sampling for explanation of black-box models, *BMVC* (2018).
- [10] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: *CVPR*, 2016.

- [11] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: ICCV, 2017.
- [12] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: ICML, 2017.
- [13] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, in: ICML, 2017.
- [14] J. Kim, A. Rohrbach, T. Darrell, J. F. Canny, Z. Akata, Textual explanations for self-driving vehicles, CoRR (2018). URL: <http://arxiv.org/abs/1807.11546>. arXiv:1807.11546.
- [15] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, D. Batra, Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization, CoRR (2016).
- [16] M. I. Posner, S. E. Petersen, The attention system of the human brain, Annual review of neuroscience (1990).
- [17] A. Das, H. Agrawal, L. Zitnick, D. Parikh, D. Batra, Human attention in visual question answering: Do humans and deep networks look at the same regions?, Computer Vision and Image Understanding (2017).
- [18] Y. Rong, W. Xu, Z. Akata, E. Kasneci, Human attention in fine-grained classification, arXiv preprint arXiv:2111.01628 (2021).
- [19] F. Liu, X. Ren, X. Wu, S. Ge, W. Fan, Y. Zou, X. Sun, Prophet attention: Predicting attention with future attention, in: NeurIPS, 2020. URL: <https://proceedings.neurips.cc/paper/2020/file/13fe9d84310e77f13a6d184dbf1232f3-Paper.pdf>.
- [20] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, CoRR (2015). URL: <http://arxiv.org/abs/1502.03044>. arXiv:1502.03044.
- [21] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, T.-S. Chua, Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning, in: CVPR, 2017.
- [22] J. Lu, C. Xiong, D. Parikh, R. Socher, Knowing when to look: Adaptive attention via a visual sentinel for image captioning, in: CVPR, 2017.
- [23] Q. You, H. Jin, Z. Wang, C. Fang, J. Luo, Image captioning with semantic attention, in: CVPR, 2016.
- [24] C. Liu, J. Mao, F. Sha, A. Yuille, Attention correctness in neural image captioning, in: AAAI, 2017.
- [25] L. Zhou, Y. Kalantidis, X. Chen, J. J. Corso, M. Rohrbach, Grounded video description, in: CVPR, 2019.
- [26] S. Barratt, Interpnet: Neural introspection for interpretable deep learning, ArXiv (2017).
- [27] M. Schuster, K. K. Paliwal, Bidirectional recurrent neural networks, IEEE transactions on Signal Processing (1997).
- [28] C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie, The Caltech-UCSD Birds-200-2011 Dataset, Technical Report, California Institute of Technology, 2011.
- [29] S. E. Reed, Z. Akata, B. Schiele, H. Lee, Learning deep representations of fine-grained visual descriptions, CoRR (2016). URL: <http://arxiv.org/abs/1605.05395>. arXiv:1605.05395.
- [30] M. Andriluka, L. Pishchulin, P. Gehler, B. Schiele, 2d human pose estimation: New benchmark and state of the art analysis, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.