An Evaluation of Image-based Verb Prediction Models against Human Eye-tracking Data

Spandana Gella and **Frank Keller**
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB, UK
spandana.gella@ed.ac.uk, keller@inf.ed.ac.uk

Abstract

Recent research in language and vision has developed models for predicting and disambiguating verbs from images. Here, we ask whether the predictions made by such models correspond to human intuitions about visual verbs. We show that the image regions a verb prediction model identifies as salient for a given verb correlate with the regions fixated by human observers performing a verb classification task.

1 Introduction

Recent research in language and vision has applied fundamental NLP tasks in a multimodal setting. An example is word sense disambiguation (WSD), the task of assigning a word the correct meaning in a given context. WSD traditionally uses textual context, but disambiguation can be performed using an image context instead, relying on the fact that different word senses are often visually distinct. Early work has focused on the disambiguation of nouns (Loeff et al., 2006; Saenko and Darrell, 2008; Chen et al., 2015), but more recent research has proposed visual sense disambiguation models for verbs (Gella et al., 2016). This is a considerably more challenging task, as unlike objects (denoted by nouns), actions (denoted by verbs) are often not clearly localized in an image. Gella et al. (2018) propose a two-stage approach, consisting of a verb prediction model, which labels an image with potential verbs, followed by a visual sense disambiguation model, which uses the image to determine the correct verb senses.

While this approach achieves good verb prediction and sense disambiguation accuracy, it is not clear to what extend the model captures human intuitions about visual verbs. Specifically, it is interesting to ask whether the image regions that the model identifies as salient for a given verb correspond to the regions a human observer relies on when determining which verb is depicted. The output of a verb prediction model can be visualized as a heatmap over the image, where hot colors indicate the most salient areas for a given task (see Figure 2 for examples). In the same way, we can determine which regions a human observes attends to by eye-tracking them while viewing the image. Eye-tracking data consists a stream of gaze coordinates, which can also be turned into a heatmap. Model predictions correspond to human intuitions if the two heatmaps correlate.

In the present paper, we show that the heatmaps generated by the verb prediction model of Gella et al. (2018) correlate well with heatmaps obtained from human observers performing a verb classification task. We achieve a higher correlation than a range of baselines (center bias, visual salience, and model combinations), indicating that the verb prediction model successfully identifies those image regions that are indicative of the verb depicted in the image.

2 Related Work

Most closely related is the work by Das et al. (2016) who tested the hypothesis that the regions attended to by neural visual question answering (VQA) models correlate with the regions attended to by humans performing the same task. Their results were negative: the neural VQA models do not predict human attention better than a baseline visual salience model (see Section 3). It is possible that this result is due to limitations of the study of Das et al. (2016): their evaluation dataset, the VQA-HAT corpus, was collected using mouse-tracking, which is less natural and less sensitive than eye-tracking. Also, their participants did not actually perform question answering, but were given a question and its answer, and then had to mark up the relevant image regions. Das et al. (2016) report a human-

Proceedings of NAACL-HLT 2018, pages 758–763
New Orleans, Louisiana, June 1 - 6, 2018.

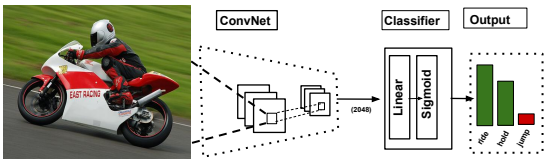

Figure 1: A schematic view of our multilabel verb classification model.

human correlation of 0.623, which suggests low task validity.

Qiao et al. (2017) also use VQA-HAT, but in a supervised fashion: they train the attention component of their VQA model on human attention data. Not surprisingly, this results in a higher correlation with human heatmaps than Das et al.'s (2016) unsupervised approach. However, Qiao et al. (2017) fail to compare to a visual salience model (given their supervised setup, such the salience model would also have to be trained on VQA-HAT for a fair comparison).

The work that is perhaps closest to our own work is Hahn and Keller (2016), who use a reinforcement learning model to predict eye-tracking data for text reading (rather than visual processing). Their model is unsupervised (there is no use of eye-tracking data at training time), but achieves a good correlation with eye-tracking data at test time.

Furthermore, a number of authors have used eye-tracking data for training computer vision models, including zero shot image classification (Karessli et al., 2017), object detection (Papadopoulos et al., 2014), and action classification in still images (Ge et al., 2015; Yun et al., 2015) and videos (Dorr and Vig, 2017). In NLP, some authors have used eye-tracking data collected for text reading to train models that perform part-of-speech tagging (Barrett et al., 2016a,b), grammatical function classification (Barrett and Søgaard, 2015), and sentence compression (Klerke et al., 2016).

3 Fixation Prediction Models

Verb Prediction Model (M) In our study, we used the verb prediction model proposed by Gella et al. (2018), which employs a multilabel CNN-based classification approach and is designed to simultaneously predict all verbs associated with an image. This model is trained over a vocabulary that consists of the 250 most common verbs in the TUHOI, Flickr30k, and COCO image description datasets. For each image in these datasets, we obtained a set of verb labels by extracting all the verbs from the ground truth descriptions of the image (each image comes with multiple descriptions, each of which can contribute one or more verbs).

Our model uses a sigmoid cross-entropy loss and the ResNet 152-layer CNN architecture. The network weights were initialized with the publicly available CNN pretrained on ImageNet[1] and fine-tuned on the verb labels. We used stochastic gradient descent and trained the network with a batch size of one for three epochs. The model architecture is shown schematically in Figure 1.

To derive fixation predictions, we turned the output of the verb prediction model into heatmaps using the class activation mapping (CAM) technique proposed by Zhou et al. (2016). CAM uses global average pooling of convolution feature maps to identify the important image regions by projecting back the weights of the output layer onto the convolutional feature maps. This technique has been shown to achieve competitive results on both object localization and localizing the discriminative regions for action classification.

Center Bias (CB) We compare against a center bias baseline, which simulates the task-independent tendency of observers to make fixations towards the center of an image. This is a strong baseline for most eye-tracking datasets (Tatler, 2007). We follow Clarke and Tatler (2014) and compute a heatmap based on a zero mean Gaussian with a co-variance matrix of $\begin{pmatrix} \sigma^2 & 0 \\ 0 & \nu\sigma^2 \end{pmatrix}$, where $\sigma^2 = 0.22$ and $\nu = 0.45$ (the values suggested by Clarke and Tatler 2014).

Visual Salience (SM) Models of visual salience are meant to capture the tendency of the human visual system to fixate the most prominent parts of a scene, often within a few hundred milliseconds of exposure. A large number of salience models have been proposed in the cognitive literature, and we choose the model of Liu and Han (2016), as it currently achieves the highest correlation with human fixations on the MIT300 benchmark out of 77 models (Bylinskii et al., 2016).

The deep spatial contextual long-term recurrent convolutional network (DSCLRCN) of Liu and Han (2016) is trained on SALICON (Jiang et al., 2015), a large human attention dataset, to infer salience for arbitrary images. DSCLRCN learns powerful local feature representations while simul-

[1] https://github.com/KaimingHe/deep-residual-networks

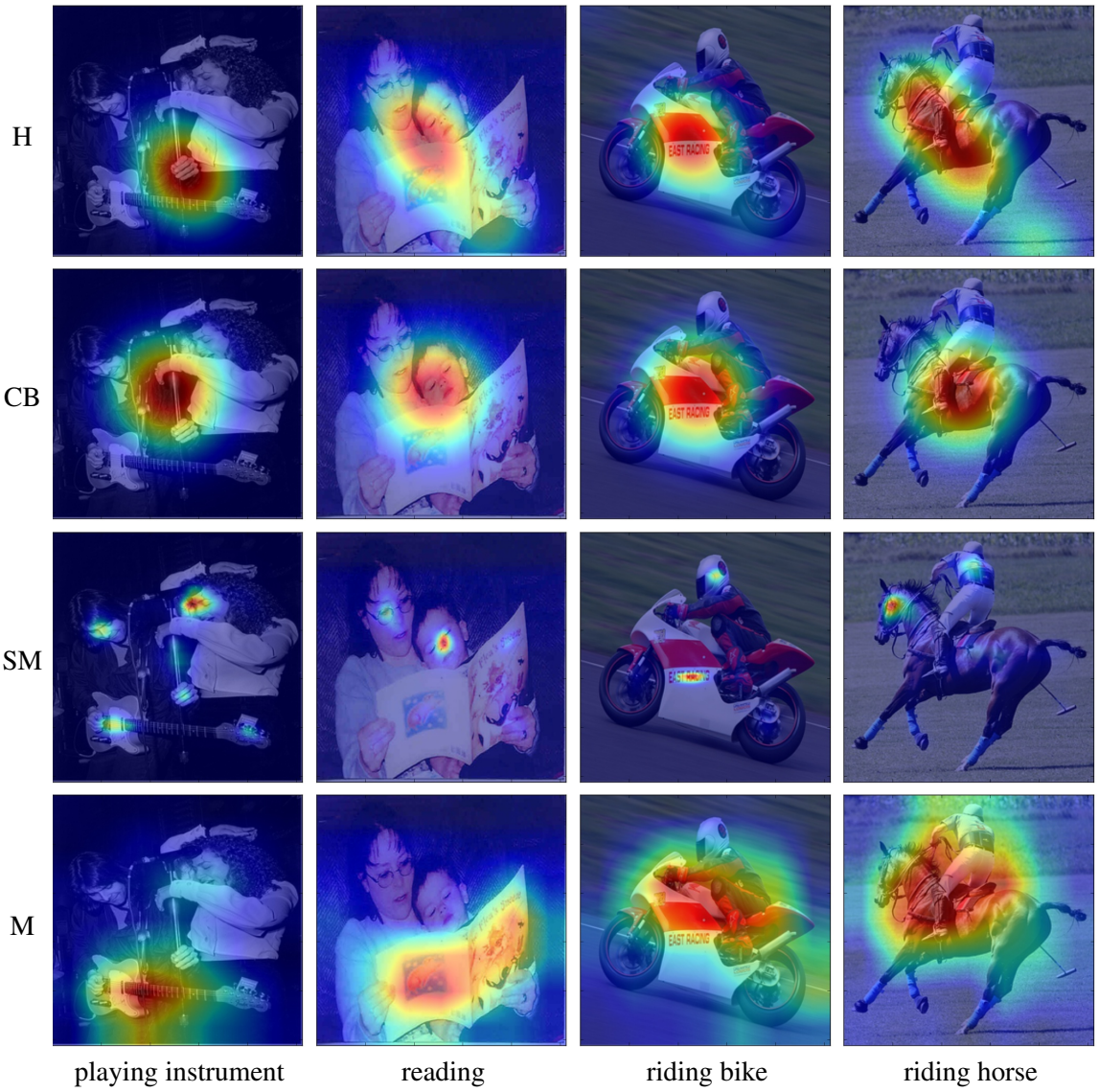

Figure 2: Heatmaps visualizing human fixations (H), Center Bias (CB), salience model (SM) predictions, and verb model (M) prediction for randomly picked example images. The SM heatmaps are very focused, which is a consequence of that model being trained on SALICON, which contains focused human attention maps. However, our evaluation uses rank correlation, rather than correlation on absolute attention scores, and is therefore unaffected by this issue.

taneously incorporating global context and scene context to compute a heatmap representing visual salience. Note that salience models are normally tested using free viewing tasks or visual search tasks, not verb prediction. However, salience can be expected to play a large role in determining fixation locations independent of task, so DSCLRCN is a good baseline to compare to.

4 Eye-tracking Dataset

The PASCAL VOC 2012 Actions Fixation dataset (Mathe and Sminchisescu, 2013) contains 9,157 images covering 10 action classes (phoning, reading, jumping, running, walking, riding bike, riding horse, playing instrument, taking photo, using computer). Each image is annotated with the eye-fixations of eight human observers who, for each image, were asked to recognize the action depicted and respond with one of the class labels. Participants were given three seconds to freely view an image while the x- and y-coordinates of their gaze positions were recorded. (Note that the original dataset also contained a control condition in which four participants performed visual search; we do not use the data from this control condition.) In Figure 2 (row H) we show examples of heatmaps generated from the human fixations in the Mathe and Sminchisescu (2013) dataset. For details on

Verb	Images	Rank correlations							
		H	CB	SM	M	CB+SM	CB+M	M+SM	M+CB+SM
phoning	221	0.911	0.599	0.361	0.562	0.598	**0.654**	0.569	0.652
reading	231	0.923	0.589	0.404	0.544	0.598	**0.655**	0.558	0.655
jumping	201	0.930	0.612	0.300	0.560	0.609	**0.650**	0.561	0.647
running	154	0.934	0.548	0.264	0.536	0.545	**0.604**	0.536	0.602
walking	195	0.938	0.553	0.311	0.535	0.552	**0.611**	0.537	0.609
riding bike	199	0.925	0.580	0.329	0.518	0.578	**0.622**	0.527	0.621
riding horse	206	0.910	0.593	0.351	0.532	0.588	**0.604**	0.532	0.601
playing instrument	229	0.925	0.571	0.350	0.478	0.568	**0.596**	0.484	0.593
taking photo	205	0.925	**0.656**	0.354	0.508	0.647	0.630	0.514	0.628
using computer	196	0.916	0.633	0.389	0.525	0.626	**0.655**	0.533	0.652
overall	2037	0.923	0.592	0.344	0.529	0.591	**0.628**	0.535	0.626

Table 1: Table of average rank correlation scores for the verb prediction model (M), compared with the upper bound of average human-human agreement (H), center bias (CB) baseline (Clarke and Tatler, 2014), and salience map (SM) baseline (Liu and Han, 2016). Results are reported on the validation set of the PASCAL VOC 2012 Actions Fixation data (Mathe and Sminchisescu, 2013). The best score for each class is shown in **bold** (except upper bound). Model combination are by mean of heatmaps.

the eye-tracking setup used, including information on measurement error, please refer to Mathe and Sminchisescu (2015), who used the same setup as Mathe and Sminchisescu (2013).

While actions and verbs are distinct concepts (Ronchi and Perona, 2015; Pustejovsky et al., 2016; Gella and Keller, 2017), we can still use the PASCAL Actions Fixation data to evaluate our model. When predicting a verb, the model presumably has to attend to the same regions that humans fixate on when working out which action is depicted – all the actions in the dataset are verb-based, hence recognizing the verb is part of recognizing the action.

5 Results

To evaluate the similarity between human fixations and model predictions, we first computed a heatmap based on the human fixations for each image. We used the PyGaze toolkit (Dalmaijer et al., 2014) to generate Gaussian heatmaps weighted by fixation durations. We then computed the heatmap predicted by our model for the top-ranked verb the model assigns to the image (out of its vocabulary of 250 verbs). We used the rank correlation between these two heatmaps as our evaluation measure. For this, both maps are converted into a 14×14 grid, and each grid square is ranked according to its average attention score. Spearman's ρ is then computed between these two sets of ranks. This is the same evaluation protocol that Das et al. (2016) used to evaluate the heatmaps generated by two question answering models with unsupervised attention, viz., the Stacked Attention Network (Yang et al., 2016) and the Hierarchical Co-Attention Network (Lu et al., 2016). This makes their rank correlations and ours directly comparable.

In Table 1 we present the correlations between human fixation heatmaps and model-predicted heatmaps. All results were computed on the validation portion of the PASCAL Actions Fixation dataset. We average the correlations for each action class (though the class labels were not used in our evaluation), and also present overall averages. In addition to our model results, we also give the correlations of human fixations with (a) the center bias baseline, and (b) the salience model. We also report the correlations obtained by all combinations of our model and these baselines. Finally, we report the human-human agreement averaged over the eight observes. This serves as an upper bound to model performance.

The results show a high human-human agreement for all verbs, with an average of 0.923. This is considerably higher than the human-human agreement of 0.623 that Das et al. (2016) report for their question answering ask, indicating that verb classification is a task that can be performed more reliably than Das et al.'s (2016) VQA region markup task (they also used mouse-tracking rather than eye-tracking, a less sensitive experimental method).

We also notice that the center baseline (CB) generally performs well, achieving an average correlation of 0.592. The salience model (SM) is less convincing, averaging a correlation of 0.344. This

is likely due to the fact that SM was trained on the SALICON dataset; a higher correlation can probably be achieved by fine-tuning the salience model on the PASCAL Actions Fixation data. However, this would no longer be fair comparison with our verb prediction model, which was not trained on fixation data (it only uses image description datasets at training time, see Section 3). Adding SM to CB does not lead to an improvement over CB alone, with an average correlation of 0.591.

Our model (M) on its own achieves an average correlation of 0.529, rising to 0.628 when combined with center bias, clearly outperforming center bias alone. Adding SM does not lead to a further improvement (0.626). The combination of our model with SM performs only slightly better than the model on its own.

In Figure 2, we visualize samples of heatmaps generated from the human fixations, the center-bias, the salience model, and the predictions of our model. We observe that human fixations and center bias exhibit high overlap. The salience model attends to regions that attract human attention independent of task (e.g., faces), while our model mimics human observers in attending to regions that are associated with the verbs depicted in the image. In Figure 2 we can observe that our model predicts fixations that vary with the different uses of a given verb (riding bike vs. riding horse).

6 Conclusions

We showed that a model that labels images with verbs is able to predict which image regions humans attend when performing the same task. The model therefore captures aspects of human intuitions about how verbs are depicted. This is an encouraging result given that our verb prediction model was not designed to model human behavior, and was trained on an unrelated image description dataset, without any access to eye-tracking data. Our result contradicts the existing literature (Das et al., 2016), which found no above-baseline correlation between human attention and model attention in a VQA task.

References

Maria Barrett, Joachim Bingel, Frank Keller, and Anders Søgaard. 2016a. Weakly supervised part-of-speech tagging using eye-tracking data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Volume 2: Short Papers*. Berlin, pages 579–584.

Maria Barrett, Frank Keller, and Anders Søgaard. 2016b. Cross-lingual transfer of correlations between parts of speech and gaze features. In *Proceedings of the 26th International Conference on Computational Linguistics*. Osaka, pages 1330–1339.

Maria Barrett and Anders Søgaard. 2015. Using reading behavior to predict grammatical functions. In *EMNLP Workshop on Cognitive Aspects of Computational Language Learning*. Lisbon, Portugal.

Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. 2016. Mit saliency benchmark. http://saliency.mit.edu/.

Xinlei Chen, Alan Ritter, Abhinav Gupta, and Tom M. Mitchell. 2015. Sense discovery via co-clustering on images and text. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. pages 5298–5306.

Alasdair DF Clarke and Benjamin W Tatler. 2014. Deriving an appropriate baseline for describing fixation behaviour. *Vision research* 102:41–51.

Edwin S Dalmaijer, Sebastiaan Mathôt, and Stefan Van der Stigchel. 2014. Pygaze: An open-source, cross-platform toolbox for minimal-effort programming of eyetracking experiments. *Behavior research methods* 46(4):913–921.

Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. 2016. Human attention in visual question answering: Do humans and deep networks look at the same regions? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*. pages 932–937.

Michael Dorr and Eleonora Vig. 2017. Saliency prediction for action recognition. In *Visual Content Indexing and Retrieval with Psycho-Visual Models*, Springer, pages 103–124.

Gary Ge, Kiwon Yun, Dimitris Samaras, and Gregory J Zelinsky. 2015. Action classification in still images using human eye movements. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. pages 16–23.

Spandana Gella and Frank Keller. 2017. An analysis of action recognition datasets for language and vision tasks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Volume 2: Short Papers*. Vancouver, pages 64–71.

Spandana Gella, Frank Keller, and Mirella Lapata. 2018. Disambiguating visual verbs. *IEEE Trans. Pattern Anal. Mach. Intell.* .

Spandana Gella, Mirella Lapata, and Frank Keller. 2016. Unsupervised visual sense disambiguation for verbs using multimodal embeddings. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*. pages 182–192.

Michael Hahn and Frank Keller. 2016. Modeling human reading with neural attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*. pages 85–95.

Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. 2015. Salicon: Saliency in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. pages 1072–1080.

Nour Karessli, Zeynep Akata, Bernt Schiele, and Andreas Bulling. 2017. Gaze embeddings for zero-shot image classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. pages 6412–6421.

Sigrid Klerke, Yoav Goldberg, and Anders Søgaard. 2016. Improving sentence compression by learning to predict gaze. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*. San Diego, CA.

Nian Liu and Junwei Han. 2016. A deep spatial contextual long-term recurrent convolutional network for saliency detection. *CoRR* abs/1610.01708.

Nicolas Loeff, Cecilia Ovesdotter Alm, and David A. Forsyth. 2006. Discriminating image senses by clustering with multimodal features. In *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006*. Association for Computational Linguistics, pages 547–554.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnet t, editors, *Advances in Neural Information Processing Systems 29*, Curran Associates, Inc., pages 289–297.

Stefan Mathe and Cristian Sminchisescu. 2013. Action from still image dataset and inverse optimal control to learn task specific visual scanpaths. In *Advances in neural information processing systems*. pages 1923–1931.

Stefan Mathe and Cristian Sminchisescu. 2015. Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.(7):1408–1424.

Dim P Papadopoulos, Alasdair DF Clarke, Frank Keller, and Vittorio Ferrari. 2014. Training object class detectors from eye tracking data. In *European Conference on Computer Vision*. Springer, pages 361–376.

James Pustejovsky, Tuan Do, Gitit Kehat, and Nikhil Krishnaswamy. 2016. The development of multimodal lexical resources. In *Proceedings of the Workshop on Grammar and Lexicon: interactions and interfaces (GramLex)*. pages 41–47.

Tingting Qiao, Jianfeng Dong, and Duanqing Xu. 2017. Exploring human-like attention supervision in visual question answering. ArXiv:1709.06308.

Matteo Ruggero Ronchi and Pietro Perona. 2015. Describing common human visual actions in images. In *Proceedings of the British Machine Vision Conference (BMVC 2015)*. BMVA Press, pages 52.1–52.12.

Kate Saenko and Trevor Darrell. 2008. Unsupervised learning of visual sense models for polysemous words. In *Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8-11, 2008*. pages 1393–1400.

Benjamin W. Tatler. 2007. The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision* 7(14):1–17.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola. 2016. Stacked attention networks for image question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV.

Kiwon Yun, Gary Ge, Dimitris Samaras, and Gregory Zelinsky. 2015. How we look tells us what we do: Action recognition using human gaze. *Journal of vision* 15(12):121–121.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Computer Vision and Pattern Recognition*.