Human-like Visual Question Answering with Multimodal Transformers

Erik S. McGuire
DePaul University
Chicago, IL
erik.s.mcguire@gmail.com

Abstract

Recently, research has been focusing on multimodal models which fuse image and language data to ground the learning of representations. One popular multimodal task is Visual Question Answering (VQA), which requires choosing the correct answer given an image and a question. In addition, datasets such as VQA-HAT (Human ATtention) enable researchers to study where human subjects attend to images when completing the VQA task. These data can also be used to supervise attention, inducing human biases in how machines attend to the same image-question pairs for the VQA task. In this work, we investigate the attention supervision of a multimodal transformer model, LXMERT, specifically its cross-modal attentions. We study the performance of the supervised model and compare the human and machine attentions. We find that performance is maintained despite successfully influencing the model to attend in a more human-like manner.

1. Introduction

A current trend in deep learning is the development of multimodal vision-and-language backbone models from the combination of conventional CNN and RNN architectures, and more recently, the development of transformer-based backbones such as ViLBERT (Lu et al., 2019), UNITER (Chen et al., 2020), VisualBERT (Li et al., 2019), VL-BERT (Su et al., 2020), and LXMERT (Tan and Bansal, 2019), where generalizable joint visiolinguistic embeddings are learned from the outset through a fusion of modal streams using various co-attentional strategies, rather than combining separately trained vision and language models (Li et al., 2020). In part, such research is motivated by recognition that a key aspect of intelligent language use is *grounding*, where a context for making meaning can be modeled through multisensory input.

Meanwhile, researchers have explored how to implement human behavior to augment models for vision-and-language tasks such as Visual Question Answering (VQA), where given an image paired with potentially multiple questions, the correct answer is selected. Neural models use information from breaking down images to accomplish this task, and the most performant models implement attention-based strategies for recognizing the most useful regions of images; given attention's importance, some studies have explored the use of human attention to explicitly influence the mechanism, finding that models aren't typically focusing on the same image regions and thus might benefit from explicit human attention supervision (Das et al., 2017; Qiao et al., 2017).

1.1. Contributions

In this work, we seek to induce human-like biases in the multi-head self-attention distributions produced by LXMERT[1]'s cross-modal attentions, by using multi-task fine-tuning of the base uncased model for VQA as the main task, with supervised attention as the auxiliary task. Specifically:

1. We preprocess Human-Like ATtention (VQA-HLAT) maps obtained from Qiao et al. (2017) into comparable attention scores.
2. We inject this data into LXMERT by computing a penalty based on the differences between model and human attention scores for loaded samples.
3. We supervise attention with this penalty during training, adding the attention losses to the main classification loss with a tunable trade-off coefficient such that we create a HAT model to compare with a baseline VQA model.
4. We log and save results for analysis of similarity between human and model attention in relation to task performance on validation data.

Results show that while performance was not improved, models fine-tuned with human features performed similarly to baseline while attention distributions

[1] https://huggingface.co/lxmert-base-uncased

were shifted toward the human ground truth data. We provide[2] our experiments and data in the form of scripts and Jupyter notebooks.

2. Related work

All of the aforementioned BERT-based multimodal models, RNN architectures, and more recently, the development of transformer-based backbones such as ViLBERT (Lu et al., 2019), UNITER (Chen et al., 2020), VisualBERT (Li et al., 2019), VL-BERT (Su et al., 2020), and LXMERT (Tan and Bansal, 2019) use similar methods and pre-training tasks. Each architecture combines the pre-trained BERT language system with image region features extracted from Faster R-CNN (Ren et al., 2015), a convolutional neural network which proposes regions in images where detected objects are bounded. The use of Faster R-CNN in this way stems from Bottom-Up and Top-Down Attention (BUTD; Anderson et al. 2018), where standard top-down attention weighting is combined with bottom-up region proposals.

Language and image embeddings are then joined through self-attention and co-attention or cross-attention blocks using masked modeling tasks where masked language modeling (MLM) is conditioned on visual features, and masked region modeling (MRM) is informed by language features. The models also use an alignment task, where images and questions are matched. Models can then be fine-tuned for multimodal tasks such as Visual Question Answering (VQA). For these tasks, the standard [CLS] token is used as the joint representation for classification. LXMERT, used in this work, consists of three transformer encoders: for language, for detected image objects, and a cross-modal encoder which uses cross-attention described above to combine representations used for cross-modal tasks. LXMERT uses datasets built from MS COCO (Lin et al., 2014) and Visual Genome (Krishna et al., 2017) images, as well as Visual Question Answering (VQA 2.0; Antol et al. 2015) and Graph Question Answering (GQA; Hudson and Manning 2019).

Split	Questions	Images
train	443,757	82,784
val	214,354	40,504
test	447,793	81,434

Table 1. Statistics for the VQA dataset.

VQA Visual Question Answering was designed to be a more "AI-complete" task–a task whose difficulty putatively requires a more general, human-like level of intelligence–which requires more fine-grained multimodal understanding than image captioning. The system is meant to take images and open-ended, naturalistic questions as input, and produce natural language answers as output.

Question	Answer
What do the white letters on the car say?	bus
Are there any signs that prevent left turns?	yes
What color is the umbrella?	yellow
Which direction is the arrow pointing?	left

Table 2. Questions and Answers for the input image given by inference on LXMERT pre-trained on VQA[4].

However, evaluations based on the original VQA dataset found that language provided strong priors which biased models to use artifacts in the text signal over rich visual understanding, e.g. simply answering "yes" in response to any question which began with certain phrases, such as "Do you see... ?" In response, VQA 2.0 was released, which contained complementary images for questions which entailed different answers, such as pairing "Who is wearing glasses?" with two images of a male/female couple, in which the man is wearing glasses in one, the woman in the other, ideally forcing the model to process more of the visual information (Goyal et al., 2017).

In VQA 2.0, there are approximately 5 answers per question on average, for over 1 million questions, and most answers consist of a single word. Typically for fine-tuning on VQA, a vocabulary of 3,129 answers is created from the most frequent correct candidates, with multi-class cross-entropy loss after selecting the most probable answer.

2.1. Human attention

VQA-HAT Inspired by the idea of how image areas are used by models for the VQA task and the role of attention mechanisms, Das et al. (2017) investigate which regions human annotators look at to accomplish the task, and whether deep learning models attend to the same regions. To do so, the authors curate a set of VQA Human ATtention maps (VQA-HAT) by having humans (800 workers on Mechanical Turk) sharpen relevant areas of blurred images, with the intent that these maps be used for evaluation and for explicit training of models.

The authors note a residual bias in the HATs stemming from a well-known center bias in saliency research, where observers tend to look toward the center of images or salient objects. They mitigate this by removing maps which correlate highly with an exemplar dataset

[2] http://github.com/erikmcguire

[4] https://huggingface.co/unc-nlp/lxmert-vqa-uncased

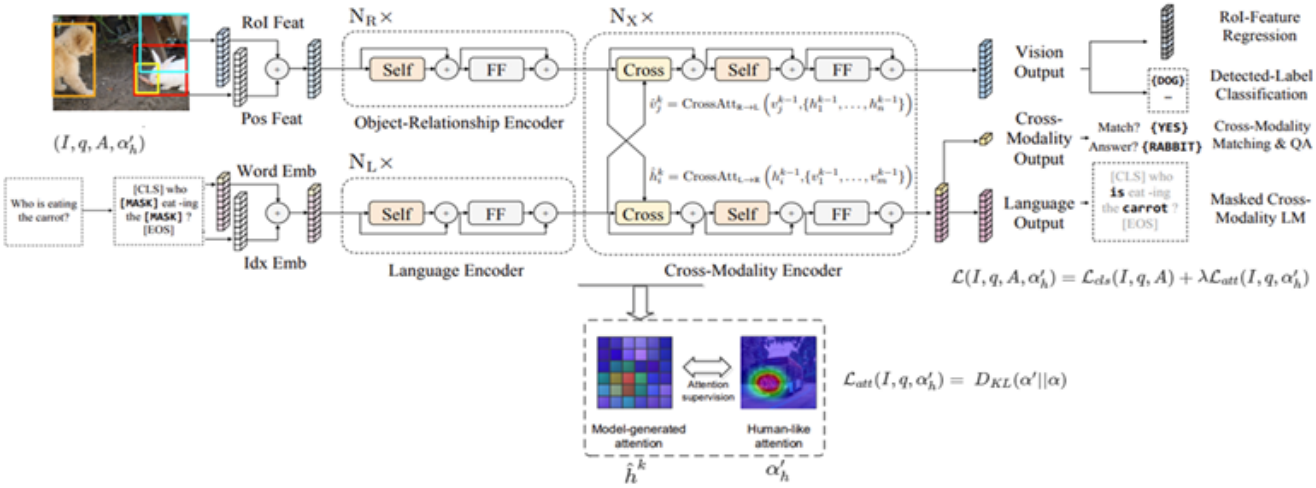

Figure 1. A collage created from diagrams in Tan and Bansal (2019) and Qiao et al. (2017), with LaTeX annotations.

that strongly contains this bias (Judd et al., 2009). Other studies (Hayes and Henderson, 2019; Peacock et al., 2020) suggest that scene meaning is more strongly predictive of human observations than saliency[5], including when center bias is controlled, so we might expect that these filtered human attention maps may be more indicative of semantic processing in observers.

VQA-HLAT Qiao et al. (2017) noted the potential training benefits of VQA-HAT for attention-based models, as well as its limitations: containing maps for only 10% of VQA 1.0. To resolve this, the authors chose to generate Human-Like ATtention maps (VQA-HLAT) using a Human Attention Network (HAN) trained to predict maps from the original VQA-HAT training dataset of 58,475 samples, using Mean Squared Error (MSE) loss between model attention weights and human maps after softmax.

The model was validated on the VQA-HAT dev set of 1,374 samples, where the authors found it surpassed human prediction performance and correlated best with human attentions. The trained HAN was then used to generate the HLAT dataset, consisting of human-like maps for all image-QA pairs in the VQA 2.0 dataset which can be used for attention supervision in VQA, which in their work produced more accurate results, with a more relevant focus on objects in images.

3. Method

For convenience due to its implementation in the HuggingFace library and provision of accessible pretrained models, we use LXMERT in our experiments.

[5]Salient image features attract attention in a bottom-up fashion, sometimes conflicting with top-down attention such as from task instructions (Lindsay, 2020).

COCO Split	Question-Image Pairs
trainval	658,111
testdev	107,394
testing	447,793

Table 3. Statistics for the VQA 2.0 HLAT dataset.

LXMERT With the processes noted in §2, our basic LXMERT VQA framework, modified from the paper's original code[6], is as follows:

1. Load train, dev, and test splits' image IDs and question-answer data from JSON files.
2. Extract 36 image region-of-interest (RoI) features with a set of 4 detected object bounding box coordinates using F-RCNN. We load these image feature vectors ($\in \mathbb{R}^{2048}$) obtained for VQA images, matching images to the IDs found in the split QA data.
3. Load VQA-HLAT maps and match to the image-QA data. These 196 vectors are reshaped to 14x14 maps and then resized to image dimensions. The bounding box coordinates $[x1, y1, x2, y2]$ are used to extract the softmax human attentions for the 36 detected objects. The human attentions are very small, as they were normalized over the entire image. After they were interpolated when resized, we found it useful when obtaining a scalar for each box to sum the respective box weights and then divide this by the sum of the entire image weights, so that each human score for a box reflected each box's proportionate mass in the image. The results no longer summed

[6]https://github.com/airsplay/lxmert/tree/master/src

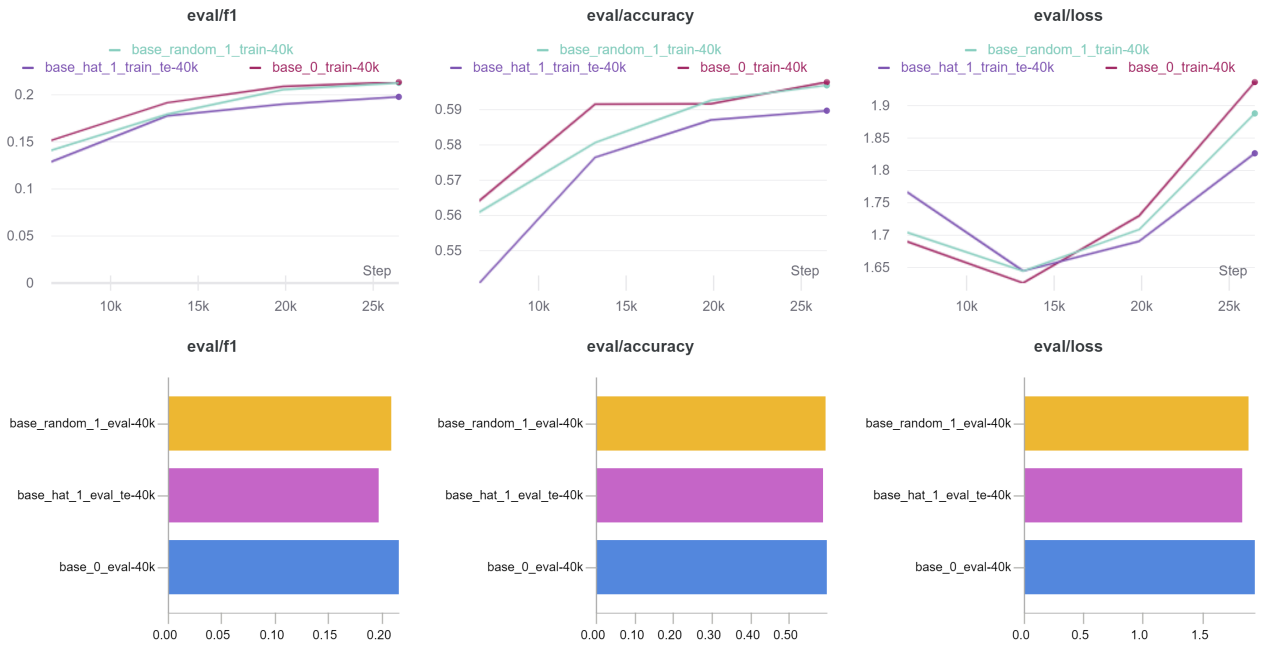

Figure 2. Top: Evaluation loss, accuracy, and F1 for the two models: human-like attention (HLAT) supervised, random supervised, and the baseline model with no attention supervision during training. Bottom: Bar charts of the same metrics for the final checkpoint model (40k steps) evaluated on the test set.

to 1, so renormalization with softmax was used; as the 36 box scores were rather diffuse and uniform, we used softmax with a low temperature to obtain a hardened, peakier distribution.

4. LXMERT outputs attention weights from language, vision, and cross- modalities, given the image features and tokenized input questions.

5. For auxiliary loss, the cross-attention weights are used for attention supervision against VQA-HLAT maps. Cross-attention consists of an attention distribution over the 36 image regions for each token. The [CLS] token's attention scores have been taken in works as sentence-level attention (Clark et al., 2019). We also use the [CLS] token attentions as the as the competitive question-level attention over the paired images from a computational reader of sorts, for comparison with the human attention over the image elicited by the same question.

6. Add the auxiliary attention loss to the main VQA task loss to fine-tune LXMERT on VQA with VQA-HLAT attention supervision.

Supervised attention As seen in Figure 1, the model performs cross-attention between language features h and vision features v in a cross-modal sublayer, which produces three outputs: cross-attended language $\hat{h}$ and vision $\hat{v}$, as well as a cross-modal output, by taking the [CLS] token vector of the crossed language features $\hat{h}$. For QA, this token's features are squashed with $\tanh$ activation to give us pooled cross-modal outputs.

The model's cross-attentions, therefore, come from the language modality after processing in the cross-modal sublayer where it is composited with visual information, with the [CLS] token being the pertinent multimodal representative of the sequence. Accordingly, we use the cross-attentions[7] for this token after softmax as the model attention to be supervised with the human attentions, which were also normalized with softmax to form a distribution over each image and then renormalized over the regions of interest.

We do this by fitting these model attentions α to the target human-like attention maps α' by computing the Kullback-Leibler divergence from α to α', $D_{KL}(\alpha'||\alpha)$, an intuitive objective[8]. since we are comparing distributions, and which has been used in other studies for explicit attention supervision (Qiuxia et al., 2020; Sharan et al., 2019; Sood et al., 2020; Zhang et al., 2019). Thus, given image I, question q, and loaded human attentions

[7]We use the maximum value per attention head, renormalizing by dividing by the sum. Subsequent related experiments found averaging over heads engenders a more robust effect.

[8]Also known as relative entropy, and equivalent to cross-entropy in the case of one-hot encoded labels.

α', we calculate the attention loss L_{att}:

$$\mathcal{L}_{att}(I, q, \alpha'_h) = D_{KL}(\alpha' || \alpha) \quad (1)$$

We compute this for each of LXMERT's 5 layers, averaged. For the main task cross-entropy loss L_{CE} we also use answer A. Thus for the joint loss with auxiliary loss coefficient λ, we sum:

$$\mathcal{L}(I, q, A, \alpha'_h) = \mathcal{L}_{CE}(I, q, A) + \lambda \mathcal{L}_{att}(I, q, \alpha'_h) \quad (2)$$

4. Experiments

Due to computational limitations, we did not perform a hyperparameter search for the attention loss coefficient, and restricted the number of images for training to 40,000, nearly half the total images (see Table 1), giving us 211,712 QA pairs rather than 443,757. We used the default LXMERT validation data of 5,000 images, resulting in 25,475 QA pairs. Test set annotations were not immediately available for offline evaluation in this project, reserved for the online VQA challenge server[9].

Ablations Because of the aforementioned limitations, we did not use the authors' pre-fine-tuned VQA LXMERT model for comparison, as they were able to train on far more data. Instead, we fine-tuned the base LXMERT model with human and random attention supervision (HAT, Random), and without attention supervision (Baseline), creating three models. The current setup does not use a convex combination of losses, so we added the full attention loss by setting the λ coefficient to 1.0, training with default LXMERT hyperparameters as described in §A.1. After training, we used the final checkpoints for evaluation.

5. Results

Due to limited data relative to the published LXMERT work, accuracies for baseline and supervised attention VQA models after 26k steps hit a ceiling at ∼59%, well below the LXMERT paper's accuracy of 72.5%. Models had very similar losses. Figure 2 charts show evaluation results during training, provided by Weights & Biases[10].

Permutation tests (Dror et al., 2018) were performed on results to assess statistical significance, along with analysis of effect size (Sullivan and Feinn, 2012) and statistical power (Card et al., 2020), as seen in Figure 3 and more comprehensively in our notebooks, as *p*-value alone gives an impartial understanding, as it depends in part on sample size, unlike effect size, with power giving us better understanding of the role of the sample size. This testing in conjunction with evaluation metrics allows us to see that the significance for the baseline model versus the HAT model on the validation set is likely a product of the large sample size, as there is a very small effect size and low power. This also shows with the random versus HAT model (p-value 0.001, effect size -0.12).

6. Analysis

Attention similarity As seen in Table 5, we measure behavioral similarity or object overlap (Sen et al., 2020). We compare the top-k regions for questions using a variety of k values, for items scored by LXMERT cross-attentions after fine-tuning and the scores given by human-like data. To do this, we run the models on validation and test sets, saving the [CLS] cross-attentions we obtained in §3 and computing the percentage of matches for machine vs. respective HLAT scores in the top k values for each image-question pairing over all k values per batch (e.g., batch size 32×3 when $k = 3$).

We can see that the attention-supervised LXMERT overlaps far more with the VQA-HLAT data. Importantly, there is a similar increase in overlap on the test set[11], on which the model's attentions were not supervised, apparently demonstrating generalization.

7. Discussion and future work

In this work, we were able to fine-tune a multimodal transformer model on the VQA task with cross-modal attention supervised by human-like attention maps without reduction in task performance, while increasing similarity in attention distributions between models and humans. Oddly, when supervising with random attention distributions, the model performed the same. This suggests that performance was preserved not because the supervisory signal was as useful as what the model learns on its own, but instead that cross-modal attentions do not affect performance as immediately as we might expect, although after using only half the data and without deeper analysis, it's premature to make assumptions.

Recall from §2.1 that VQA-HAT and consequently VQA-HLAT were designed to mitigate the center bias, and that when the center bias is controlled, it has been found that semantics play a stronger role. Potentially, by inducing models to attend to images based on data with an attenuated center bias, we might be forcing models to rely more on semantics in the visual modality, a purported counterbalance to the language bias that VQA 2.0 attempted to correct.

[9] https://visualqa.org/challenge.html
[10] https://wandb.ai/

[11] As with training, the test set run was limited to 40k images and 219,872 their paired QA samples, due to memory limitations

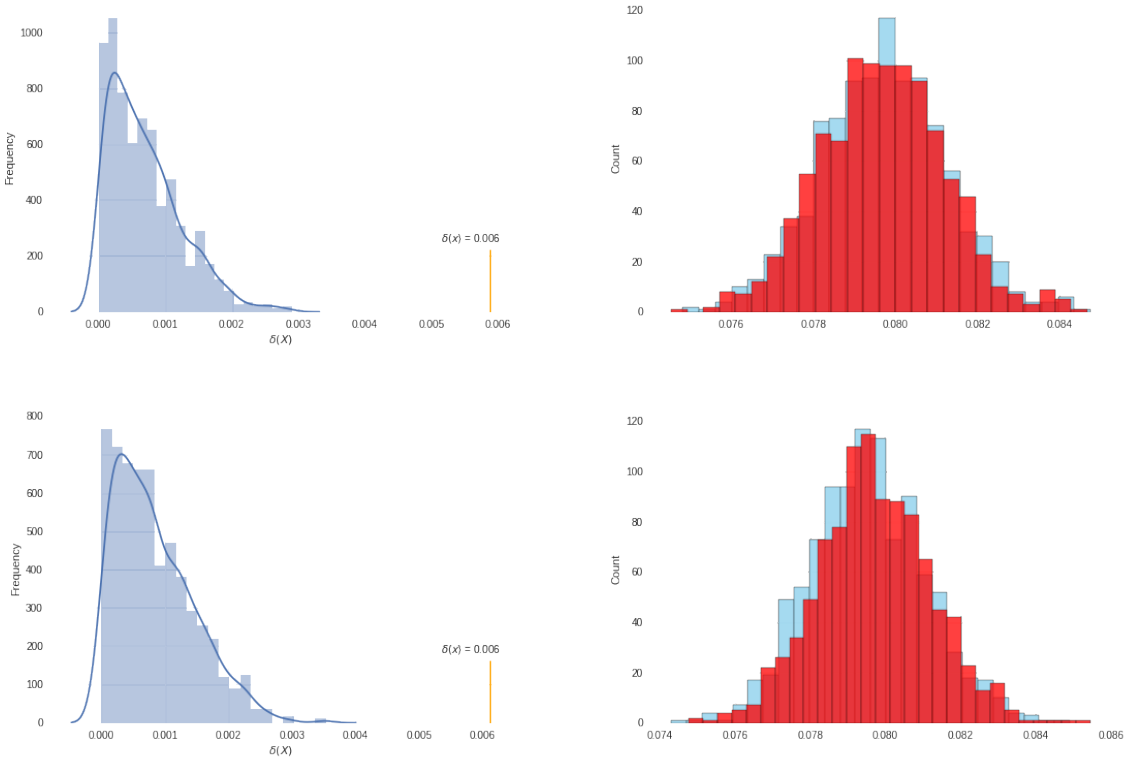

Figure 3. Top: Permutation test results comparing baseline model with model using VQA-HLAT supervision. Left: Sample and original delta (gold) for the dev set, showing the random accuracy differences are most often lower than the original, and thus the results are significant (p-value 0.001), which we might expect with a large sample size of 25,475. Right: Histogram showing the distributions of permutation scores for baseline (blue) and VQA-HAT (red) after 1,000 rounds on the dev set. Complementing p-value with effect size reveals little difference in performance: there is a very small positive effect size (Cohen's $d = 0.02$ ($\leq 85\%$ overlap; confidence interval [-0.071, 0.105] and statistical power of 48.33%. Bottom: Comparing random with HLAT.

model	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$
Base	3.43	6.17	8.93	11.59	14.33	17.10	19.76	22.46
Random	3.63	6.82	10.22	13.46	16.59	19.68	22.73	25.55
HLAT	73.02	76.17	78.39	79.54	80.24	80.68	81.07	81.74

Table 4. Overlapping top-k for model vs. human attentions on dev set. We can see that the model with HLAT supervision has the most overlap in every case.

model	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$
Base	8.94	13.12	16.50	19.51	22.27	24.86	27.39	29.84
Random	4.24	8.03	11.64	15.15	18.55	21.78	24.83	27.74
HLAT	73.26	75.88	78.11	79.51	80.27	80.82	81.37	82.00

Table 5. Overlapping top-k for model vs. human attentions on test set. Note that the attention was not supervised for this set.

For some cases when the link from the question to the image is ambiguous, human subjects may have been able to focus on the appropriate areas despite what for the machine is a lack of text clues. In these cases, we can imagine the human attention data provides the necessary signal.

7.1. Future work

Future work should experiment with other multimodal visual grounding tasks and attempt to correlate results with human judgments. Fully training on all of the VQA data and experimenting with different trade-off λ parameters would also be useful. We have also set up the system to save test predictions in a format suitable for submission to the VQA benchmark servers, in which case we can evaluate how performance fared on the test set, as the labels for the test set are unavailable to the public.

One ethical consideration would be neurodivergent versus neurotypical attention. Plausibly there are many ways of attending effectively, with a spectrum of strengths and weakness to be explored. For example, autistic subjects have been found to attend less to faces than neurotypical subjects, with autistic females attending more to faces than males with autism (Harrop et al., 2019).

Gaze data is also considered a reasonable proxy for human attention (Zhang et al., 2020), and it may be worthwhile to incorporate eye-tracking data into the model. Zhang et al. (2019) use a cosine-decay function during VQA training with attention supervision, to allow models to develop their own attentional preferences as they learn, which is an operation that offers intriguing possibilities for authors who may want to customize the ways their models attend.

It may be reasonable to evaluate the relative distances of overlapping attentions' magnitudes to other elements within their respective distributions, to assess how similar the scores are for overlaps between models and human attention and the role of peaky versus diffuse, uniform distributions.

The human attention is coarse, so there may be value in creating a sort of hierarchy of coarse to fine attention or vice versa among the heads, decaying the loss in one direction or the other so that the model is allowed to use more precise box attentions at certain layers and/or heads.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086. [§2.]

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433. [§2.]

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. [§5.]

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: UNiversal Image-TExt Representation Learning. [§§1 and 2.]

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does BERT look at? an analysis of BERT's attention. *arXiv preprint arXiv:1906.04341*. [§5.]

Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. 2017. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100. [§§1 and 2.1.]

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392. [§5.]

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. [§2.]

Clare Harrop, Desiree Jones, Shuting Zheng, Sallie Nowell, Robert Schultz, and Julia Parish-Morris. 2019. Visual attention to faces in children with autism spectrum disorder: are there sex differences? *Molecular autism*, 10(1):28. [§7.1.]

Taylor R Hayes and John M Henderson. 2019. Center bias outperforms image salience but not semantics in accounting for attention during scene viewing. *Attention, Perception, & Psychophysics*, pages 1–10. [§2.1.]

Drew A Hudson and Christopher D Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6700–6709. [§2.]

Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. 2009. Learning to predict where humans look. In *2009 IEEE 12th international conference on computer vision*, pages 2106–2113. IEEE. [§2.1.]

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis

Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73. [§2.]

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A simple and performant baseline for vision and language. [§§1 and 2.]

Yikuan Li, Hanyin Wang, and Yuan Luo. 2020. A comparison of pre-trained vision-and-language models for multimodal representation learning across medical images and reports. [§1.]

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer. [§2.]

Grace W Lindsay. 2020. Attention in psychology, neuroscience, and machine learning. *Frontiers in Computational Neuroscience*, 14:29. [§5.]

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. [§§1 and 2.]

Candace E Peacock, Taylor R Hayes, and John M Henderson. 2020. Center bias does not account for the advantage of meaning over salience in attentional guidance during scene viewing. *Frontiers in Psychology*, 11. [§2.1.]

Tingting Qiao, Jianfeng Dong, and Duanqing Xu. 2017. Exploring human-like attention supervision in visual question answering. *arXiv preprint arXiv:1709.06308*. [§§1, 1, 1, and 2.1.]

LAI Qiuxia, Salman Khan, Yongwei Nie, Sun Hanqiu, Jianbing Shen, and Ling Shao. 2020. Understanding more about human and machine attention in deep neural networks. *IEEE Transactions on Multimedia*. [§3.]

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99. [§2.]

Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. 2016. Grad-CAM: Why did you say that? *arXiv preprint arXiv:1611.07450*. [§4.]

Cansu Sen, Thomas Hartvigsen, Biao Yin, Xiangnan Kong, and Elke Rundensteiner. 2020. Human attention maps for text classification: Do humans and neural networks focus on the same words? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4596–4608. [§6.]

Komal Sharan, Ashwinkumar Ganesan, and Tim Oates. 2019. Improving visual reasoning with attention alignment. In *International Symposium on Visual Computing*, pages 219–230. Springer. [§3.]

Ekta Sood, Simon Tannert, Diego Frassinelli, Andreas Bulling, and Ngoc Thang Vu. 2020. Interpreting attention models with human visual attention in machine reading comprehension. *arXiv preprint arXiv:2010.06396*. [§3.]

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: Pre-training of generic visual-linguistic representations. [§§1 and 2.]

Gail M Sullivan and Richard Feinn. 2012. Using effect size—or why the P value is not enough. *Journal of graduate medical education*, 4(3):279–282. [§5.]

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*. [§§1, 2, and 1.]

Ruohan Zhang, Akanksha Saran, Bo Liu, Yifeng Zhu, Sihang Guo, Scott Niekum, Dana Ballard, and Mary Hayhoe. 2020. Human gaze assisted artificial intelligence: A review. In *IJCAI: proceedings of the conference*, volume 2020, page 4951. NIH Public Access. [§7.1.]

Yundong Zhang, Juan Carlos Niebles, and Alvaro Soto. 2019. Interpretable visual question answering by visual grounding from attention supervision mining. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 349–357. IEEE. [§§3 and 7.1.]

A. Supplemental Material

A.1. Model parameters

LXMERT We use LXMERT uncased[12] with the paper's settings: 4 epochs, learning rate 5e-5, and batch size 32.

[12] https://huggingface.co/unc-nlp/lxmert-base-uncased

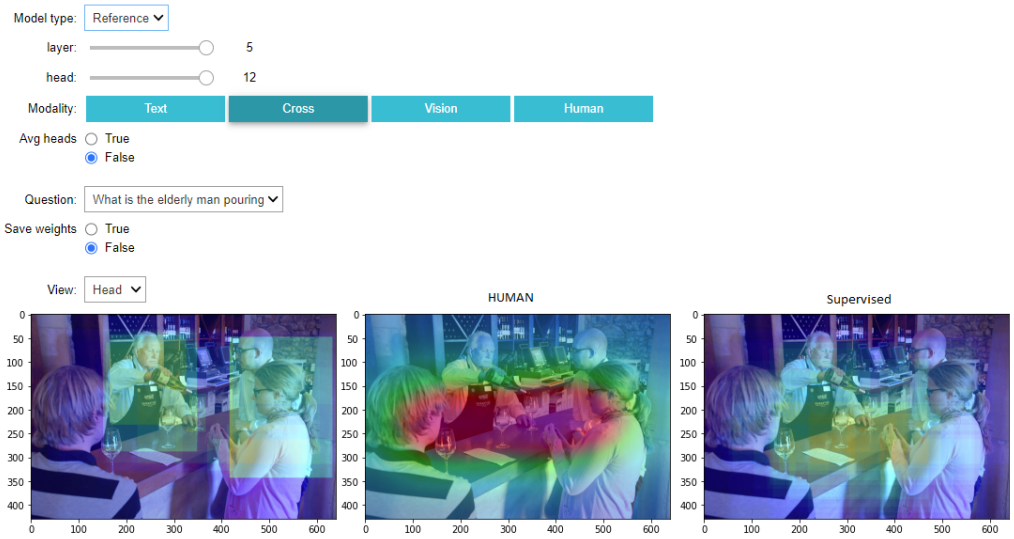

Figure 4. An instance where the baseline model (left) differs substantially from the human attentions, as compared with the attention supervised model. This style of visualization is somewhat inspired by Grad-CAM (Selvaraju et al., 2016).

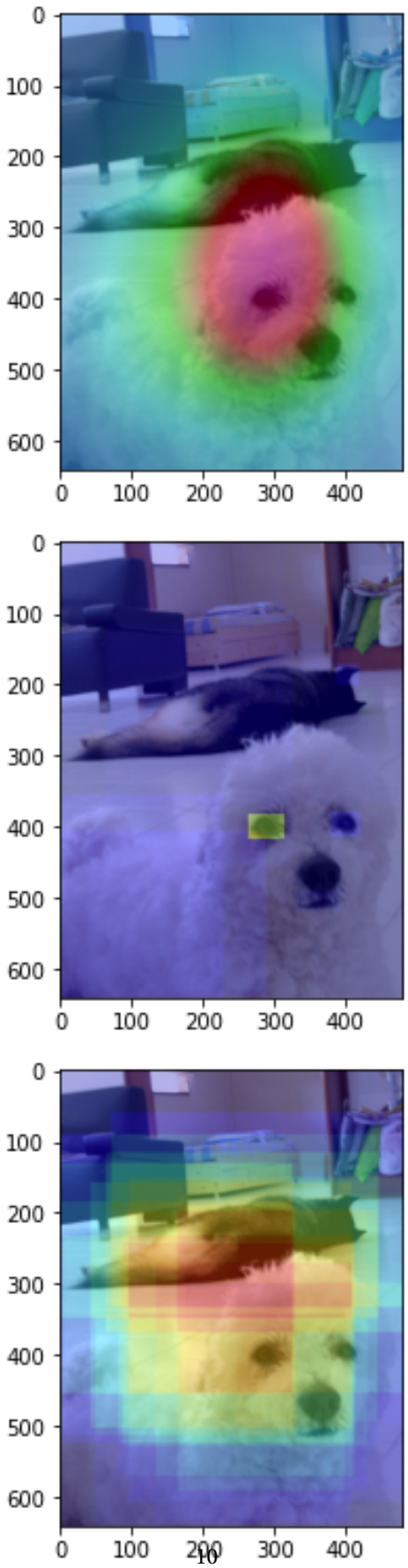

Figure 5. Here we see how much more precise the model attentions are, whereas the human attentions encouraged a coarse attention in response to a question about the dog's eyes. From top to bottom: human, baseline, HLAT-supervised attentions.