TAKING A HINT: LEVERAGING EXPLANATIONS TO MAKE VISION AND LANGUAGE MODELS MORE GROUNDED

Ramprasaath R. Selvaraju[1] **Stefan Lee**[1] **Yilin Shen**[2] **Hongxia Jin**[2] **Shalini Ghosh**[2]
Dhruv Batra[1,3] **Devi Parikh**[1,3]
[1]Georgia Institute of Technology, [2]Samsung Research America, [3]Facebook AI Research
{ramprs, steflee, dbatra, parikh}@gatech.edu
{yilin.shen, hongxia.jin, shalini.ghosh}@samsung.com

ABSTRACT

Many vision and language models suffer from poor visual grounding – often falling back on easy-to-learn language priors rather than associating language with visual concepts. In this work, we propose a generic framework which we call Human Importance-aware Network Tuning (HINT) that effectively leverages human supervision to improve visual grounding. HINT constrains deep networks to be sensitive to the same input regions as humans. Crucially, our approach optimizes the alignment between human attention maps and gradient-based network importances, ensuring that models learn not just to look at but rather rely on visual concepts that humans found relevant for a task when making predictions. We demonstrate our approach on Visual Question Answering and Image Captioning tasks, achieving state-of-the-art for the VQA-CP dataset which penalizes over-reliance on language priors.

1 INTRODUCTION

Many popular and well-performing models for multi-modal, vision-and-language tasks exhibit poor visual grounding – failing to appropriately associate words or phrases with the image regions they denote and relying instead on superficial linguistic correlations (2; 1; 6; 7). For example, answering the question "What color are the bananas?" with yellow regardless of their ripeness evident in the image. When challenged with datasets that penalize reliance on these sort of biases (2; 6), state-of-the-art models demonstrate significant drops in performance despite there being no change to the set of visual and linguistic concepts about which models must reason.

In addition to these diagnostic datasets, another powerful class of tools for observing this shortcoming has been gradient-based explanation techniques (12; 16; 11) which allow researchers to examine which portions of the input models rely on when making decisions. Applying these techniques has shown that vision-and-language models often focus on seemingly irrelevant or contextual image regions that differ significantly from where human subjects fixate for the same tasks.

While somewhat dissatisfying, these findings are not wholly surprising – after all, standard practices do not provide any guidance for visual grounding. Instead, models are trained on input-output pairs and must resolve grounding from co-occurrences – a challenging task, especially in the presence of more direct and easier to learn correlations in language. To combat this tendency, we explore how to provide grounding supervision directly.

Towards this end, we introduce a generic, second-order approach that updates model parameters to better align gradient-based explanations with human attention maps. Our approach which we call Human Importance-aware Network Tuning (HINT) enforces a ranking loss between human annotations of input importance and gradient-based explanations produced by a deep network – updating model parameters via a gradient-of-gradient step. Importantly, this constrains models to not only look at the correct regions but to also be sensitive to the content present when making predictions. This forces models to base their decisions on the same regions as humans, providing explicit grounding supervision. While we explore applying HINT to vision-and-language problems, this approach is general and can be applied to focus model decisions on specific inputs in any context.

Contributions. To summarize our contributions, we

- introduce Human Importance-aware Network Tuning (HINT), a general approach for constraining the sensitivity of deep networks to specific input regions and demonstrate it results in significantly improved visual grounding for two vision and language tasks, and
- set a new state-of-the-art on the bias-sensitive VQA Under Changing Priors (VQA-CP) dataset.

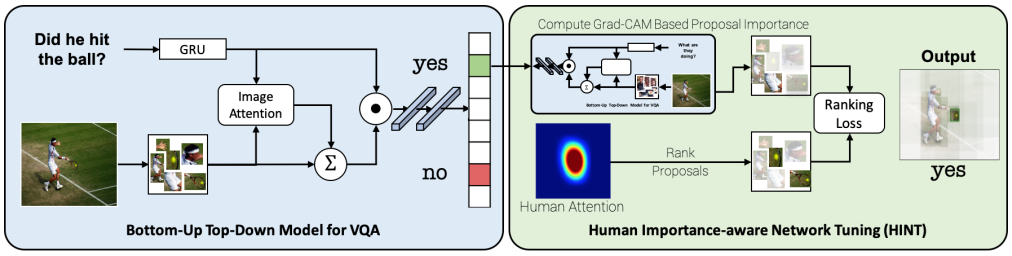

Figure 1: Our Human Importance-aware Network Tuning (HINT) approach: Given an image and a question like "Did he hit the ball?", we pass them through the Bottom-up Top-down architecture shown in the left half. For the example shown, the model incorrectly answers 'no'. For the ground-truth answer 'yes', we determine the proposals important for the decision through Grad-CAM. We rank the proposals through human attention and provide a ranking loss in order to align the network's importance with human importance. Tuning the model through HINT makes the model not only answer correctly, but also look at the right regions.

2 Related Work

Model Interpretability. There has been significant recent interest in building machine learning models that are transparent and interpretable in their decision making process. For deep networks, several works propose explanations based on internal states or structures of the network (15; 11). Most related to our work is the approach of Selvaraju (11) which computes neuron importance as part of a visual explanation pipeline. In this work, we enforce that these importance scores match importances provided by domain experts.

Human Attention for VQA. Das (5) collected human attention maps for a subset of the VQA dataset (4). Given a question and a blurry image, humans were asked to interactively deblur regions in the image until they could confidently answer. In this work we utilize these maps, enforcing the gradient-based visual explanations of model decisions to match the human attention closely.

3 Human Importance-aware Network Tuning

The premise of our work is as follows– humans tend to rely on some portion of the input more than others when making decisions. Our approach ensures that those portions of input are relevant for the model as well. HINT computes the important concepts through gradient-based explanations and tunes the network parameters so as to align with the concepts deemed important by humans. We use the generic term 'decision' to refer to both the answer in the case of VQA and the words generated at each time step in the case of image captioning. While our approach is generic and can be applied to any architecture, below we describe HINT in context of the Bottom-up Top-down model for VQA and captioning. The Bottom-up Top-down model architecture can be seen in the left half of 1 is a variant of the traditional attention mechanism, where the attention is at the level of objects and other salient image regions giving significant improvements in VQA and captioning performance.

3.1 Human Importance

In this step we align the expert knowledge obtained from humans into a form corresponding to the network inputs. The Bottom-up Top-down model (3) takes in as input, region proposals. For a given image and question (in case of VQA) we compute an importance score for each of the proposals for the correct decision based on the normalized human attention map energy inside the proposal box relative to the normalized energy outside the box.

More concretely, consider an importance map $A^d \in \mathbb{R}^{h \times w}$ that indicates the spatial regions of support for a decision d with higher values in A^d. Given a proposal region r with area a_r, we can write the normalized importance inside and outside r for decision d as

$$E^d_+(r) = \frac{1}{a_r} \sum_{(i,j) \in r} A^d_{ij} \quad \text{and} \quad E^d_-(r) = \frac{1}{hw - a_r} \sum_{(i,j) \notin r} A^d_{ij}$$

respectively. We compute the overall importance score for k for decision d as: $s^d_k = \frac{E^d_+(k)}{E^d_+(k) + E^d_-(k)}$

Human attention for VQA and captioning. For VQA, we use the human attention maps collected by Das (5) for a subset of the VQA (4) dataset. While human attention maps do not exist for image captioning, COCO dataset (8) has segmentation annotations for 80 everyday occurring categories. We use an object category to word mapping that maps object categories like <person> to a list of potential fine-grained labels like ["child", "man", "woman", ...] similar to (9). We map a total of 830 visual words existing in COCO captions to 80 COCO categories. We then use the segmentation annotations for the 80 categories as human attention for this subset of matching words.

Model	VQA-CP `test`				VQAv2 `val`			
	Overall	Yes/No	Number	Other	Overall	Yes/No	Number	Other
SAN (14)	24.96	38.35	11.14	21.74	52.41	70.06	39.28	47.84
UpDn (3)	39.49	45.21	11.96	42.98	62.85	80.89	42.78	54.44
GVQA (2)†	31.30	57.99	13.68	22.14	48.24	72.03	31.17	34.65
UpDn + Attn. Align	39.37	43.02	11.89	45.00	63.24	80.99	42.55	55.22
UpDn + AdvReg (10)†	41.17	65.49	**15.48**	35.48	62.75	79.84	42.35	55.16
UpDn + HINT (ours)	**47.78**	**70.04**	10.68	**46.31**	**63.38**	**81.18**	**42.99**	**55.56**

Table 1: Results on compositional (VQA-CP) and standard split (VQAv2). We see that our approach (HINT) gets a significant boost of over 8% from the base UpDn model on VQA-CP and minor gains on VQAv2. The Attn. Align baseline sees similar gains on VQAv2, but fails to improve grounding on VQA-CP. † results taken from corresponding papers.

3.2 Network Importance

We define Network Importance as the importance (weight) that the given trained network places on spatial regions of the input when forced to make a decision. Selvaraju (11) proposed an approach to compute the importance of last convolutional layer's neurons. In their work, they compute the importance of last convolutional layer neurons as they serve as the best compromise between high level semantics and detailed spatial information. Since proposals usually look at objects and salient/semantic regions of interest while providing a good spatial resolution, we naturally extend (11) to compute importance over proposals. Given a proposal r, its embedding P^r, its importance for predicting the ground-truth decision d_{gt}, can be computed as,

$$\alpha^r_{d_{gt}} = \overbrace{\sum_{i=1}^{|P|}}^{\text{global pooling}} \underbrace{\frac{\partial o_{d_{gt}}}{\partial P^r_i}}_{\text{gradients via backprop}} \tag{1}$$

where $o_{d_{gt}}$ is a one-hot encoding containing the score for the ground-truth decision (answer in VQA and the visual word in case of captioning). Note that we compute the importance for the ground-truth decision, and not the predicted. Human attention for incorrect decisions are not available and are intuitively non-existent, as there exists no evidence for incorrect predictions in the image.

3.3 Human-Network Importance Alignment

At this stage, we now have two sets of importance scores – one from the human attention and another from network importance – that we would like to align. We focus on the relative rankings of the proposals, applying a ranking loss – specifically, a variant of Weighted Approximate Rank Pairwise (WARP) loss. At a high level, our ranking loss searches all possible pairs of proposals and finds those pairs where the pair-wise ranking based on network importance disagrees with the ranking from human importance. Let $\mathcal{S}$ denote the set of all such misranked pairs. For each pair in $\mathcal{S}$, the loss is updated with the absolute difference between the network importance score for the proposals pair. In order to stabilize training we observe that it is necessary to have the task loss – cross entropy loss in case of VQA and negative log-likelihood for image captioning. So the HINT loss becomes,

$$\mathcal{L}_{\mathcal{HINT}} = \sum_{(r',r)\in\mathcal{S}} \left|\alpha^{r'}_{-} - \alpha^{r}_{+}\right| + L_{Task} \tag{2}$$

The first term encourages the network to base decisions on the correct regions and the second term encourages it to actually make the right decision.

4 Experiments and Analysis

4.1 HINT for VQA

Table 1 shows results on VQA-CP and VQAv2 for HINT and prior work. We summarize key results:

HINT reduces language-bias. For VQA-CP, our HINTed UpDown model significantly improves over its base architecture alone by 8 percentage point gain in overall accuracy. Further, it outperforms existing approaches based on the same UpDn architecture (41.17 for AdvReg vs 47.78 for HINT), setting a new state-of-the-art for this problem. We do note that our approach uses additional supervision in the form of human attention maps for 6% of training images.

HINT improves grounding without reducing standard VQA performance. Unlike previous approaches for language-bias reduction which cite trade-offs in performance between the VQA and VQA-CP splits (10; 2), we find our HINTed UpDn model improves on standard VQA – making HINT the first ever approach to show simultaneous improvement on both splits.

Attn. Align is ineffective compared to HINT. A surprising (to us at least) finding and motivating observation of this work is that directly supervising model attention (as in Attn. Align) is ineffective at reducing language-bias and improving visual grounding as measured by VQA-CP.

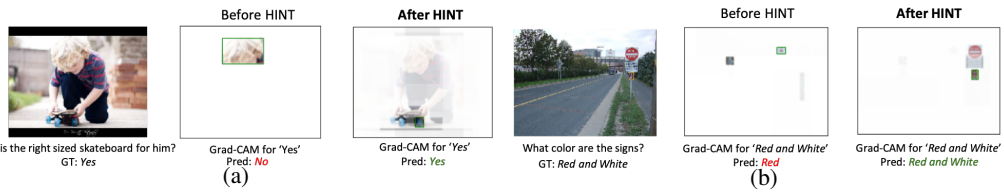

Figure 2: Qualitative comparison of models before and after applying HINT. The left column shows the input image along with the question and the ground-truth (GT) answer from the VQA-CP val split. In the middle column, for the base model we show the explanation visualization for the GT answer along with the model's answer. Similarly we show the explanations and predicted answer for the HINTed models in the third column. We see that the HINTed model looks at more appropriate regions and answers better.

4.2 HINT FOR IMAGE CAPTIONING

Our implementation of the Bottom-up Top-down captioning model achieves a CIDEr (13) score of 1.06 on the standard split and 0.90 on the robust split. Upon applying HINT to the base model trained on the robust split, we obtain a CIDEr score of 0.92, an improvement of 0.02 over the base model. For the model trained on the standard split, performance drops by 0.02 in CIDEr score (1.04 compared to 1.06). As we show in the following sections, the lack of improvement in score does not imply a lack of change – we find the model shows significant improvements at grounding (Fig. 3).

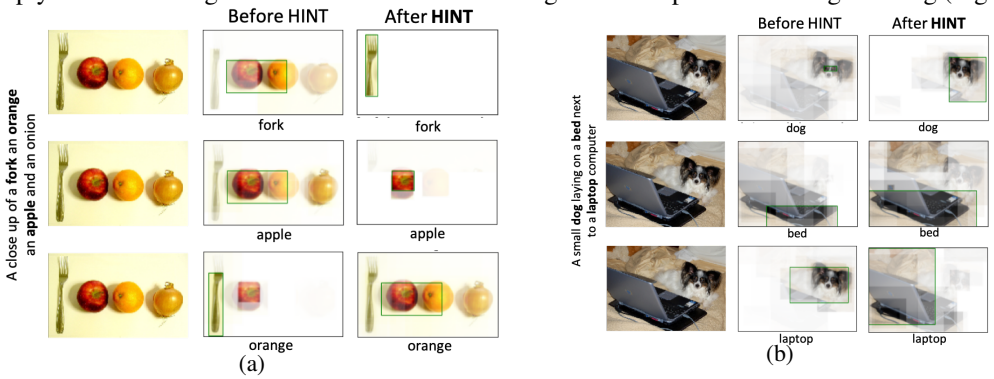

Figure 3: Qualitative comparison of captioning models before and after applying HINT. The left column shows the input image along with the ground-truth caption from the COCO robust split. In the middle column, for the base model we show the explanation visualization for the visual word mentioned below. Explanations for the HINTed models are in the third column. We see that the HINTed model looks at more appropriate regions. For example in (a) note how the HINTed model correctly localizes the fork, apple and the orange accurately when generating the corresponding visual words, but the base model fails to do so.

5 CONCLUSION

We presented Human Importance-aware Network Tuning (HINT), a general framework for aligning network sensitivity to spatial input regions that humans deemed as being relevant to a task. We demonstrated this method's effectiveness at improving visual grounding in vision and language tasks such as Visual Question Answering and Image Captioning. We also show that better grounding not only improves the generalization capability of models to arbitrary test distributions, but also improves the trust-worthiness of model.

REFERENCES

[1] A. Agrawal, D. Batra, and D. Parikh. Analyzing the behavior of visual question answering models. In *EMNLP*, 2016.
[2] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi. Dont just assume; look and answer: Overcoming priors for visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
[3] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.
[4] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. VQA: Visual Question Answering. 2015.
[5] A. Das, H. Agrawal, C. L. Zitnick, D. Parikh, and D. Batra. Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions? 2016.
[6] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017.
[7] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. 2017.
[8] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. 2014.
[9] J. Lu, J. Yang, D. Batra, and D. Parikh. Neural baby talk. In *CVPR*, 2018.
[10] S. Ramakrishnan, A. Agrawal, and S. Lee. Overcoming language priors in visual question answering with adversarial regularization. In *Neural Information Processing Systems (NIPS)*, 2018.
[11] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017.
[12] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328, 2017.
[13] R. Vedantam, C. L. Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. *CoRR*, abs/1411.5726, 2014.
[14] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 21–29, 2016.
[15] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. 2014.
[16] J. Zhang, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff. Top-down Neural Attention by Excitation Backprop. 2016.