Sparse and structured visual attention

Martins, P.H.; Niculae, V.; Marinho, Z.; Martins, A.F.T.

DOI
10.48550/arXiv.2002.05556
10.1109/ICIP42928.2021.9506028

Publication date
2021

Document Version
Author accepted manuscript

Published in
2021 IEEE International Conference on Image Processing

SPARSE AND STRUCTURED VISUAL ATTENTION

Pedro Henrique Martins[Ψ] *Vlad Niculae* [Ϋ] *Zita Marinho*[ϻϠ] *André F. T. Martins*[ΨϞϧ]

[Ψ] Instituto de Telecomunicações [Ϋ] IvI, University of Amsterdam [ϻ] Priberam Labs
[Ϡ]Institute of Systems and Robotics [Ϟ]LUMLIS (Lisbon ELLIS Unit) [ϧ]Unbabel

ABSTRACT

Visual attention mechanisms are widely used in multimodal tasks, as visual question answering (VQA). One drawback of softmax-based attention mechanisms is that they assign some probability mass to all image regions, regardless of their adjacency structure and of their relevance to the text. In this paper, to better link the image structure with the text, we replace the traditional softmax attention mechanism with two alternative sparsity-promoting transformations: *sparsemax*, which is able to select only the relevant regions (assigning zero weight to the rest), and a newly proposed *Total-Variation Sparse Attention* (TVMAX), which further encourages the joint selection of adjacent spatial locations. Experiments in VQA show gains in accuracy as well as higher similarity to human attention, which suggests better interpretability.

Index Terms— Attention, Structured Sparsity, Total Variation

1. INTRODUCTION

Vision-language tasks, as visual question answering (VQA), require combining natural language understanding with object and scene recognition. While general purpose architectures can be powerful [1, 2], the ability to incorporate structural bias is a desirable feature to better link the language and vision components and produce more interpretable decisions. How can we encourage models to look at the relevant objects only, avoiding distractions?

The current state of the art for these tasks is based on deep neural networks with **visual attention** [3, 4, 5, 6, 7]. These models use attention mechanisms to select either grid features generated by convolutional neural networks (CNNs) pretrained on image recognition datasets [8], or CNN features of bounding boxes. While bounding boxes have the advantage that the attention mechanism can attend to full objects, they require an external object segmentation model, which has a computational cost. In this paper, we propose new **selective visual attention mechanisms** over grid features, which owe their ability to select compact objects to the encouragement of joint selection of neighboring regions.

A key component of attention mechanisms is the transformation that maps scores into probability values, with softmax being the standard choice [1]. A downside of softmax is that it is **strictly dense**, *i.e.*, it devotes some attention probability mass to *every* region in the image. This makes the model less interpretable and, for complex images, it may lead to a "lack of focus". This is visible in the example of Fig. 1: the model using softmax attends, always, to the entire image and, consequently, wrongly predicts that no one is crossing the bridge.

In this work, we introduce novel selective visual attention mechanisms by endowing them with a new capability: that of **selecting only the relevant features of the image**. To this end, we first propose replacing softmax with **sparsemax** [9]. With sparsemax, the attention weights obtained are sparse, leading to the selection (non-zero attention) of only a few relevant features. While sparsemax has been applied successfully to NLP to attend over *words* [10, 9, 11], its application to attention over *image regions* is so far unexplored. However, as can be seen in the example of Fig. 1, despite leading to an increased focus on the relevant features, sparsemax selects discontiguous regions of the image which prevents the model from attending to full objects and reduces interpretability.

Thus, to further encourage the weights of related adjacent spatial locations to be the same (*e.g.*, parts of an object), we introduce a new attention mechanism: **Total-Variation Sparse Attention** (which we dub TVMAX), inspired by prior work in **structured sparsity** [12, 13]. Two key results of our paper (§2.3 and Propositions 1–2) show that TVMAX can be evaluated by composing a proximal operator with a sparsemax projection, and that its Jacobian has a closed-form expression. This leads to an efficient implementation of its forward and backward passes.

With TVMAX, sparsity is allied to the ability of selecting *compact* regions, improving interpretability, as shown in Fig. 1. Experiments, in VQA, show that TVMAX leads to improved accuracy while having attention maps more similar to human attention, suggesting higher interpretability.[1]

2. SELECTIVE ATTENTION

Attention mechanisms [1] have the ability to dynamically attend to relevant input features, such as regions of an image. To permit end-to-end training with gradient backpropagation,

[1]Code available at `https://github.com/deep-spin/TVmax`

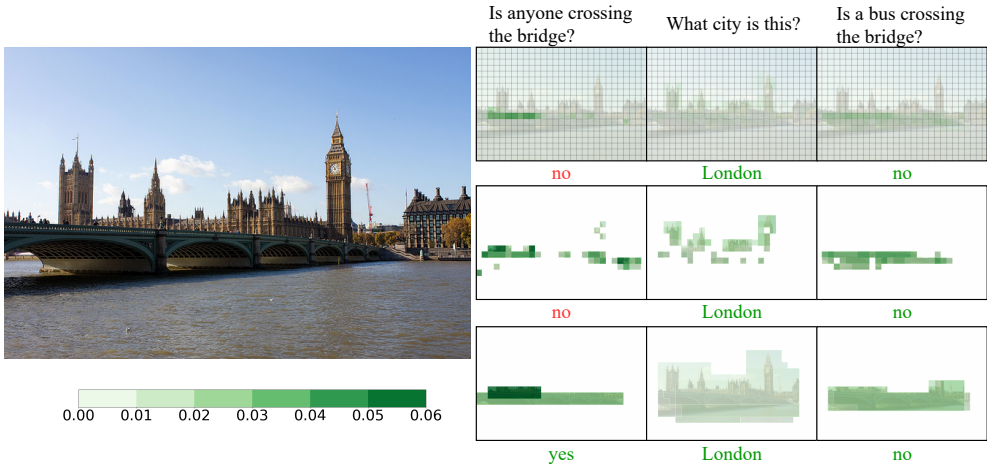

Fig. 1: VQA example using softmax (top), sparsemax (middle), and the proposed attention mechanism: TVMAX (bottom).

they require a differentiable mapping from importance scores $\boldsymbol{z} \in \mathbb{R}^k$ to a distribution $\boldsymbol{p} \in \triangle^k$, where $\triangle^k := \{\boldsymbol{p} \in \mathbb{R}^k \mid \sum_{i=1}^k p_i = 1, \boldsymbol{p} \geqslant \mathbf{0}\}$ denotes the probability simplex. The standard choice is the softmax transformation, defined as $[\mathsf{softmax}(\boldsymbol{z})]_i = \frac{\exp(z_i)}{\sum_j \exp(z_j)}$. Since softmax is strictly positive, its output is **dense**: it always assigns some probability mass to all image regions, even irrelevant ones. This accumulation of low probabilities may "distract" the model, preventing it from fully attending to the most relevant parts. This motivates our proposed **selective** visual attention mechanisms.

2.1. Sparsemax

To achieve selective capabilities, we propose the use of **sparsemax** [9], a sparse mapping consisting in the Euclidean projection of $\boldsymbol{z}$ onto the simplex: $\mathsf{sparsemax}(\boldsymbol{z}) := \arg\min_{\boldsymbol{p} \in \triangle^k} \frac{1}{2}\|\boldsymbol{p} - \boldsymbol{z}\|_2^2$. Sparsemax encourages sparse outputs, corresponding to the boundary of $\triangle^k$. This is an attractive property for visual attention mechanisms, where often only a few features provide relevant information.

2.2. Sparse and Structured Visual Attention

Since the model, often, needs to identify the full objects present in the image, the selected regions should be encouraged to have a compact structure. However, sparsemax is **unstructured** and **index-invariant**, leading it to select discontiguous regions. To overcome this, we propose a new visual attention mechanism, **TVMAX**. TVMAX is a (non-trivial) generalization of fusedmax [14], a 1D transformation based on fused lasso, to the 2D case. For ease of exposition, we first describe how fused lasso is extended to arbitrary graphs, and then we particularize to the 2D case.

Let $\boldsymbol{w} \in \mathbb{R}^k$ be a vector of weights, and (V, E) be an undirected graph, where $V = \{1, \ldots, k\}$ and $E \subseteq \{(i,j) \in V^2 \mid i < j\}$. The generalized fused lasso penalty [13] is defined as $\Omega_E(\boldsymbol{w}) = \sum_{(i,j)\in E} |w_i - w_j|$. Minimizing Ω_E encourages "fused" solutions, *i.e.*, it encourages $w_i = w_j$ for $(i,j) \in E$. In particular, its proximal operator can be seen as a **fused signal approximator**, seeking a vector $\boldsymbol{w}$ that approximates $\mathbf{x}$ well and that is encouraged to be fused:

$$\mathsf{prox}_{\lambda\Omega_E}(\mathbf{x}) = \arg\min_{\boldsymbol{w}\in\mathbb{R}^d} \frac{1}{2}\|\boldsymbol{w} - \mathbf{x}\|_2^2 + \lambda\Omega_E(\boldsymbol{w}). \quad (1)$$

Computing the value of $\mathsf{prox}_{\lambda\Omega_E}$ is non-trivial in general [15], but for certain edge configurations efficient algorithms exist:

- If E forms a chain, the problem is called **1D total variation** and the penalty is defined as $\Omega_{1D}^{TV}(\boldsymbol{w}) := \sum_{i=1}^{k-1} |w_{i+1} - w_i|$. It can be solved in $\mathcal{O}(k)$ time using the *taut string algorithm* [16, 17]. We use the quasilinear algorithm of [18], which is very fast in practice.
- If the indices are aligned on a 2D grid, as in an image, the problem is called **2D total variation** and the penalty is defined as $\Omega_{2D}^{TV}(\boldsymbol{W}) := \sum_i \Omega_{1D}^{TV}(\boldsymbol{w}_{i,:}) + \sum_j \Omega_{1D}^{TV}(\boldsymbol{w}_{:,j})$, where $\boldsymbol{w}_{i,:}$ and $\boldsymbol{w}_{:,j}$ denote the rows and columns of $\boldsymbol{W}$. Unlike the 1D case, exact algorithms are not available. However, for an input of size $a \times b$, it is possible to *split* the penalty into a column-wise and b row-wise 1D problems, and apply iterative methods, as the proximal Dykstra algorithm [17, 19].

TVMAX combines 2D total variation (TV2D) regularization with sparsemax. This way, it promotes sparsity and encourages the attention weights of adjacent spatial locations to be the same, selecting contiguous regions of the image.

Definition 1 (TVMAX). *Let $\boldsymbol{z} \in \mathbb{R}^k$, such that the indices of $\boldsymbol{z}$ can be decomposed into rows and columns. The* TVMAX *transformation is defined as*

$$\text{TVMAX}(\boldsymbol{z}) := \arg\min_{\boldsymbol{p}\in\triangle^k} \frac{1}{2}\|\boldsymbol{p} - \boldsymbol{z}\|_2^2 + \lambda\Omega_{2D}^{TV}(\boldsymbol{p}), \quad (2)$$

where λ is a hyper-parameter controlling the amount of fusion ($\lambda = 0$ recovers sparsemax) and Ω_{2D}^{TV} is the 2D TV penalty.

2.3. TVMAX's Forward and Backward Passes

In order to use the TVMAX transformation as a component in a neural network, we need efficient forward and backpropagation algorithms. We will now derive these algorithms for a more general case, the **generalized fused sparse attention**. We follow [14] and define

$$\mathsf{gfusedmax}_E(\mathbf{x}) \coloneqq \operatorname*{arg\,min}_{\boldsymbol{p}\in\triangle^k} \|\boldsymbol{p}-\mathbf{x}\|_2^2 + \lambda\Omega_E(\boldsymbol{p}). \quad (3)$$

This can be seen as a *constrained* fused lasso approximator, because the solution $\boldsymbol{p}$ must be a probability distribution vector. While the optimization function is very similar to Eq. 1, note the additional constraint $\boldsymbol{p} \in \triangle^k$. Fortunately, the following result holds.

Proposition 1. *The generalized fusedmax can be expressed as* $\mathsf{gfusedmax}_E(\mathbf{x}) = \mathsf{proj}_{\triangle^k}\left(\mathsf{prox}_{\lambda\Omega_E}(\mathbf{x})\right)$.

Proof. This result is an extension of Proposition 2 in [14], and also follows from Corolary 4 of [20]. By taking $f = \iota_{\triangle}$,[2] and noting that $\iota_{\triangle}$ is symmetric: if $\boldsymbol{p} \in \triangle$, then any vector $\boldsymbol{p}'$ obtained by permuting $\boldsymbol{p}$ is also in $\triangle$, because its values remain non-negative and sum to 1. □

Proposition 1 shows that gfusedmax's forward pass can be computed simply by composing the proximal step of fused lasso with the forward pass of sparsemax. It also provides a shortcut for deriving the Jacobian of gfusedmax via the *chain rule*. Denoting by $\boldsymbol{J}_F$ the Jacobian of $\mathsf{prox}_{\lambda\Omega_E}$, we have:

$$\frac{\partial\,\mathsf{gfusedmax}}{\partial \boldsymbol{z}} = \boldsymbol{J}_{\mathsf{sparsemax}}(\mathsf{prox}_{\lambda\Omega_E}(\boldsymbol{z}))\boldsymbol{J}_F(\boldsymbol{z}). \quad (4)$$

$\boldsymbol{J}_{\mathsf{sparsemax}}$ has been derived by [9]: $\boldsymbol{J}_{\mathsf{sparsemax}}(\boldsymbol{z}) = \mathsf{diag}\,\boldsymbol{s} - \frac{1}{\|\boldsymbol{s}\|_1}\boldsymbol{s}\boldsymbol{s}^\top$, where $s_j = 1$ if $\mathsf{sparsemax}(z)_j > 0$ and $s_j = 0$ otherwise. The next proposition completes the puzzle, giving a full characterization of $\boldsymbol{J}_F$.

Proposition 2 (Group-wise characterization of $\mathsf{prox}_{\lambda\Omega_E}$)**.** *Let $\boldsymbol{w}^\star \coloneqq \mathsf{prox}_{\lambda\Omega_E}(\boldsymbol{z})$, and denote by G_i the set of indices fused to w_i in the solution, defined recursively:*

1. *$i \in G_i$ for all i, and*
2. *$j \in G_i$ $\exists m \in G_i$ such that edge $(m, j) \in E$ and $w_m^\star = w_j^\star$.*

Define $s_{ij} = \mathsf{sign}(w_i^\star - w_j^\star)$. Then, we have

$$w_i^\star = \frac{1}{|G_i|}\sum_{j\in G_i}\left(z_j + \sum_{\substack{(m,j)\in E,\\ m\notin G_i}}\lambda s_{mj} - \sum_{\substack{(j,m)\in E,\\ m\notin G_i}}\lambda s_{jm}\right). \quad (5)$$

[2] $\iota_{\triangle}$ is the indicator function of set $\triangle$.

Proof. The subgradient optimality conditions of Eq. 1 are [21]:

$$w_i^\star - z_i + \sum_{(i,k)\in E}\lambda t_{ik} - \sum_{(k,i)\in E}\lambda t_{ki} = 0 \qquad 1 \le i \le d. \quad (6)$$

where $t_{ij} = s_{ij}$ if $w_i^\star \neq w_j^\star$, otherwise t_{ij} is a free variable in $[-1, 1]$. We focus on a single group $G = G_i$. Within a fused group, the solution is constant, *i.e.*, $w_j^\star = w$ for $j \in G$. We separate the sums in Eq. 6 according to whether $k \in G$ or not, and move the "constant" terms to the right hand side, yielding

$$w + \sum_{\substack{(j,k)\in E\\ k\in G}}\lambda t_{jk} - \sum_{\substack{(k,j)\in E\\ k\in G}}\lambda t_{kj} = z_j - \sum_{\substack{(j,k)\in E\\ k\notin G}}\lambda s_{jk} + \sum_{\substack{(k,j)\in E\\ k\notin G}}\lambda s_{kj}, \quad (7)$$

for $j \in G$. Summing up the Eq. 7 over all $j \in G$, we observe that for any edge $(i, j) \in E$ with $i, j \in G$, the term λt_{jk} appears twice with opposite signs (as in Eq. 9 in [22]). Thus,

$$\sum_{j\in G} w = \sum_{j\in G}\left(z_j + \sum_{\substack{(k,j)\in E\\ k\notin G}}\lambda s_{kj} - \sum_{\substack{(j,k)\in E\\ k\notin G}}\lambda s_{jk}\right). \quad (8)$$

Dividing by $|G|$ gives exactly Eq. 5. This reasoning applies to any group G_i. □

Proposition 2 enables easy computation of a generalized Jacobian of gfusedmax: since small perturbations in $\boldsymbol{z}$ never change the groups G_i nor the signs of across-group differences s_{ij}, differentiating Eq. 5 yields

$$(\boldsymbol{J}_F)_{i,j} = \frac{\partial w_i^\star}{\partial z_j} = \begin{cases}\frac{1}{|G_i|}, & j \in G_i,\\ 0, & j \notin G_i.\end{cases} \quad (9)$$

This generalizes Lemma 1 of [14] to arbitrary graphs.

Computation. As we show in Proposition 1, computing TVMAX's forward pass can be done by chaining efficient algorithms for TV2D and sparsemax. From Eq. 4 we have that TVMAX's Jacobian can be computed by composing $\boldsymbol{J}_F$ and $\boldsymbol{J}_{\mathsf{sparsemax}}$. As derived in Proposition 2, $(\boldsymbol{J}_F)_{i,j} = 1/n_{ij}$ if i and j are fused in a group with n_{ij} elements, and 0 otherwise. Thus, the backward pass intuitively involves "spreading" the credit assigned to one region across all regions fused with it. This can be implemented by Alg. 1 in $\mathcal{O}(N_g \log k)$ where N_g is the number of groups of fused regions. In the worst case, when there are no fused regions, the complexity is $\mathcal{O}(k \log k)$. This algorithm is inspired by flood filling algorithms [23].

3. EXPERIMENTS

To compare the attention mechanisms in VQA, we use the encoder-decoder version of modular co-attention networks

		Test-Dev				Test-Standard			
		Y/N	Numb.	Other	Overall	Y/N	Numb.	Other	Overall
bounding boxes	softmax	85.14	49.59	58.72	68.57	85.56	49.54	59.11	69.04
	sparsemax	85.41	50.29	58.62	68.71	85.80	50.18	59.08	69.19
grid	softmax	86.88	52.61	60.15	70.31	86.94	52.88	60.36	70.56
	sparsemax	86.61	52.28	60.04	70.11	86.77	52.66	60.14	70.40
	TVMAX	**86.92**	**53.19**	**60.22**	**70.42**	**86.98**	**53.08**	**60.56**	**70.70**

Table 1: VQA accuracy (per-type and overall) on VQA-2.0 dataset using bounding box features or grid features as input.

Algorithm 1 TVMAX backward pass

Input: $\boldsymbol{p} = \text{TVMAX}(\mathbf{x})$, $\mathrm{d}\boldsymbol{p} \in \mathbb{R}^k$.
Output: $\mathrm{d}\mathbf{x} = \boldsymbol{J}_{\text{TVMAX}}^\top(\mathrm{d}\boldsymbol{p}) \ \in \mathbb{R}^k$
Initialize: $N \leftarrow \varnothing, V \leftarrow \varnothing, G \leftarrow \varnothing, s = 0$
$\mathrm{d}\boldsymbol{w} \leftarrow (\boldsymbol{J}_{\text{sparsemax}})^\top \mathrm{d}\boldsymbol{p}$
while $|\boldsymbol{V}| < k$ **do**
  pick $(i_0, j_0) \notin V$, **push** (i_0, j_0) **to** N
  while N not empty **do**
    pop (i, j) **from** N
    if $p_{i,j} = p_{i_0,j_0}$ **then**
      $G \leftarrow G \cup \{(i,j)\}, V \leftarrow V \cup \{(i,j)\}, s \leftarrow s + (\mathrm{d}\boldsymbol{w})_{i,j}$
      for all neighbours $(i', j') \sim (i, j)$ **do**
        if $(i', j') \notin V$ **then**
          push (i', j') **to** N
  if $\boldsymbol{G}$ not empty **then**
    $(\mathrm{d}\mathbf{x})_{i,j} \leftarrow {}^{s}/_{|G|}$ for all $(i,j) \in G, G \leftarrow \varnothing, s = 0$

[5]. To represent the image we use grid features pre-trained by [7] on Visual Genome [8] with a ResNet-152 as backbone ("grid" in Table 1) or bounding box features extracted with Faster R-CNN [24] pretrained on Visual Genome ("bounding boxes" in Table 1). The different attention mechanisms, softmax, sparsemax, and TVMAX, are used in the output attention layer. All models were trained on VQA-v2 dataset [25] for 15 epochs using Adam [26] with a learning rate of $\min(2.5t \cdot 10^{-5}, 1 \cdot 10^{-4})$ when using bounding boxes and $\min(2.5t \cdot 10^{-5}, 5 \cdot 10^{-5})$ for grid features, where t is the epoch number. After 10 epochs, the learning rate is multiplied by $1/5$ every 2 epochs. We set $\lambda = 0.01$ for TVMAX.

Results. When using bounding box features, sparsemax outperforms softmax, suggesting that a sparse selection of relevant bounding boxes leads to more accurate answers. When using grid features as input, the model using TVMAX attention outperforms all other models. This shows that having sparse attention in conjunction to encouraging the selection of contiguous regions, not only improves interpretability but also leads to an accuracy improvement in VQA. Moreover, the superior result of TVMAX when compared to sparsemax corroborates our premise that selecting contiguous regions of the image is beneficial. We can also see that, as stated in [7], grid features outperform bounding box features.

Human attention. Finally, to understand if TVMAX leads to higher interpretability, we compared the attention distributions obtained using the different transformations with human attention. To do so, we used the VQA-HAT dataset [27], where human attention is obtained by having annotators unblurring the relevant regions of the images. To compare the attention distributions with the human attention we used the Spearman's rank correlation and the Jensen-Shannon divergence (JS). As shown in Table 2, the attention distributions obtained with TVMAX are more similar to human attention than with softmax and sparsemax. This indicates that TVMAX leads to more interpretable attention distributions.

	Spearman	JS divergence
softmax	0.33	0.64
sparsemax	0.32	0.66
TVMAX	**0.37**	**0.62**

Table 2: Spearman correlation and JS divergence between attention distributions obtained with the different models and human attention.

4. CONCLUSIONS

We propose using sparse and structured visual attention to improve the process of selecting the relevant features. For that, we used sparsemax and introduced TVMAX. By selecting only relevant compact groups of features, TVMAX leads to more interpretable attention distributions, as shown by the higher similarity to human attention. Our experiments in VQA show improvements in accuracy when replacing softmax by sparsemax to attend over bounding boxes and when using TVMAX to attend over grid features.

5. ACKNOWLEDGMENTS

This work was supported by the ERC StG DeepSPIN 758969, by the FCT through contract UIDB/50008/2020 and contract PD/BD/150633/2020 in the scope of the Doctoral Program FCT - PD/00140/2013 NETSyS, and Lisboa 2020 through ERDF, within project TRAINER (Nº 045347). We thank Pedro M. Q. Aguiar for helpful discussion and feedback.

6. REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. ICLR*, 2015.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Proc. NeurIPS*, 2017.

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang, "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering," in *Proc. CVPR*, 2018.

[4] Hao Tan and Mohit Bansal, "LXMERT: Learning Cross-Modality Encoder Representations from Transformers," in *Proc. EMNLP*, 2019.

[5] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian, "Deep Modular Co-Attention Networks for Visual Question Answering," in *Proc. CVPR*, 2019.

[6] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu, "UNITER: UNiversal Image-TExt Representation Learning," in *Proc. ECCV*, 2020.

[7] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen, "In defense of grid features for visual question answering," in *Proc. CVPR*, 2020.

[8] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *IJCV*, 2017.

[9] Andre Martins and Ramon Astudillo, "From Softmax to Sparsemax: A Sparse Model of Attention and Multi-Label Classification," in *Proc. ICML*, 2016.

[10] Chaitanya Malaviya, Pedro Ferreira, and André FT Martins, "Sparse and Constrained Attention for Neural Machine Translation," in *Proc. ACL*, 2018.

[11] Ben Peters, Vlad Niculae, and André FT Martins, "Sparse Sequence-to-Sequence Models," *In Proc. ACL*, 2019.

[12] Francis Bach, Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, et al., "Optimization with sparsity-inducing penalties," *Foundations and Trends in Machine Learning*, 2012.

[13] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight, "Sparsity and smoothness via the fused lasso," *Journal of the Royal Statistical Society*, 2005.

[14] Vlad Niculae and Mathieu Blondel, "A Regularized Framework for Sparse and Structured Neural Attention," in *Proc. NeurIPS*, 2017.

[15] Bo Xin, Yoshinobu Kawahara, Yizhou Wang, Lingjing Hu, and Wen Gao, "Efficient generalized fused lasso and its applications," *ACM TIST*, 2016.

[16] P. Laurie Davies and Arne Kovac, "Local extremes, runs, strings and multiresolution," *The Annals of Statistics*, 2001.

[17] Álvaro Barbero and Suvrit Sra, "Modular proximal optimization for multidimensional total-variation regularization," 2014.

[18] Laurent Condat, "A direct algorithm for 1-d total variation denoising," *IEEE Signal Processing Letters*, 2013.

[19] Gideon Dresdner Fabian Pedregosa, Geoffrey Negiar, "http://openopt.github.io/copt/," 2020.

[20] Yaoliang Yu, "On decomposing the proximal map," in *Proc. NeurIPS*, 2013.

[21] Jerome Friedman, Trevor Hastie, Holger Höfling, and Robert Tibshirani, "Pathwise coordinate optimization," *The Annals of Applied Statistics*, 2007.

[22] Vivian Viallon, Sophie Lambert-Lacroix, Holger Höfling, and Franck Picard, "Adaptive Generalized Fused-Lasso: Asymptotic Properties and Applications," 2013.

[23] Sergei Burtsev and Ye.P. Kuzmin, "An efficient flood-filling algorithm," *Computers & Graphics*, 1993.

[24] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Proc. NIPS*, 2015.

[25] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh, "Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering," in *Proc. CVPR*, 2017.

[26] Diederik P. Kingma and Jimmy Ba, "Adam: A Method for Stochastic Optimization.," 2014.

[27] Abhishek Das, Harsh Agrawal, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra, "Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions?," in *Proc. EMNLP*, 2016.

A. EXAMPLES

Additional VQA examples, using the softmax, sparsemax, and TVMAX attention, are presented in Figures 2, 3, 4, and 5.

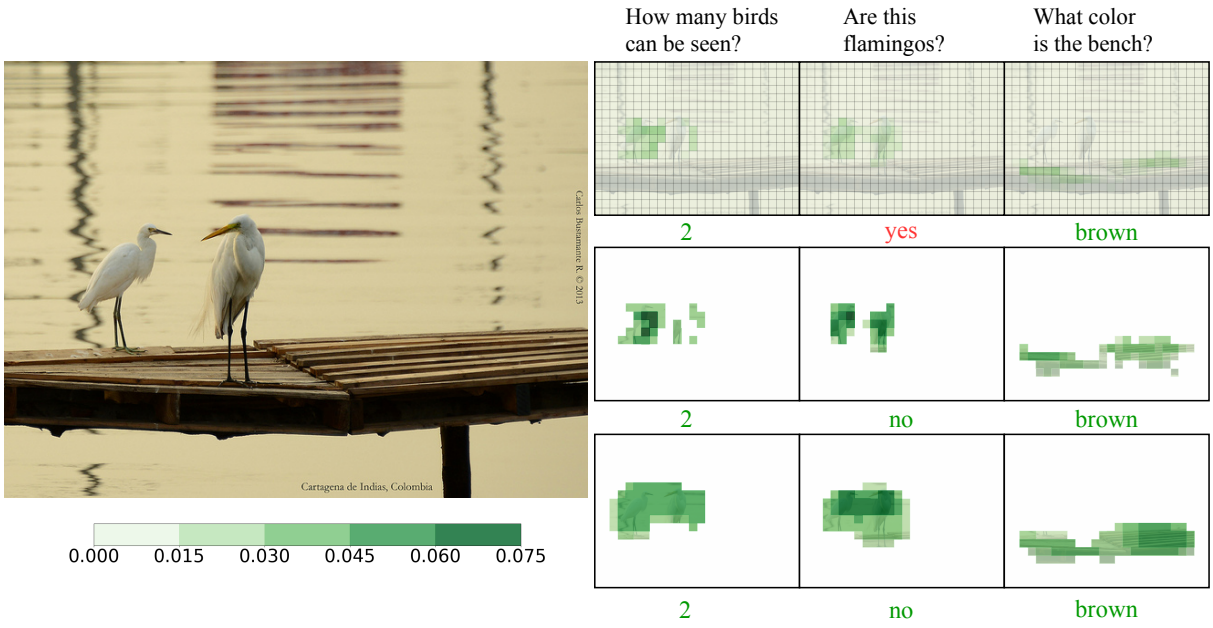

Fig. 2: VQA using softmax (top), sparsemax (middle) and TVMAX attention (bottom). Shading denotes the attention weight, with white for zero attention.

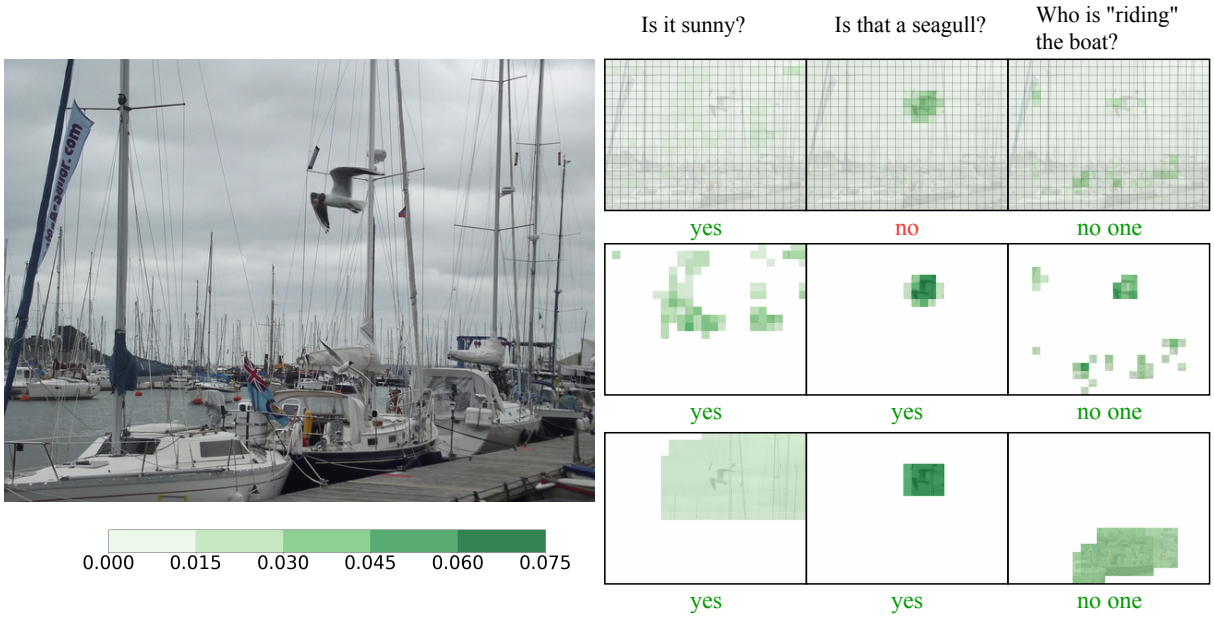

Fig. 3: VQA using softmax (top), sparsemax (middle) and TVMAX attention (bottom). Shading denotes the attention weight, with white for zero attention.

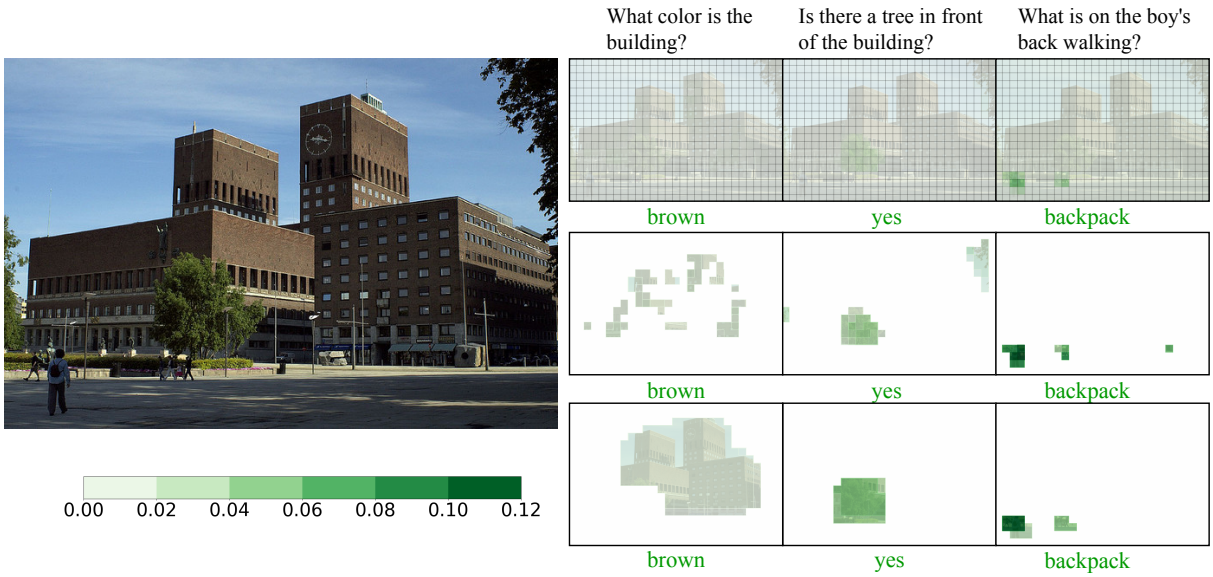

Fig. 4: VQA using softmax (top), sparsemax (middle) and TVMAX attention (bottom). Shading denotes the attention weight, with white for zero attention.

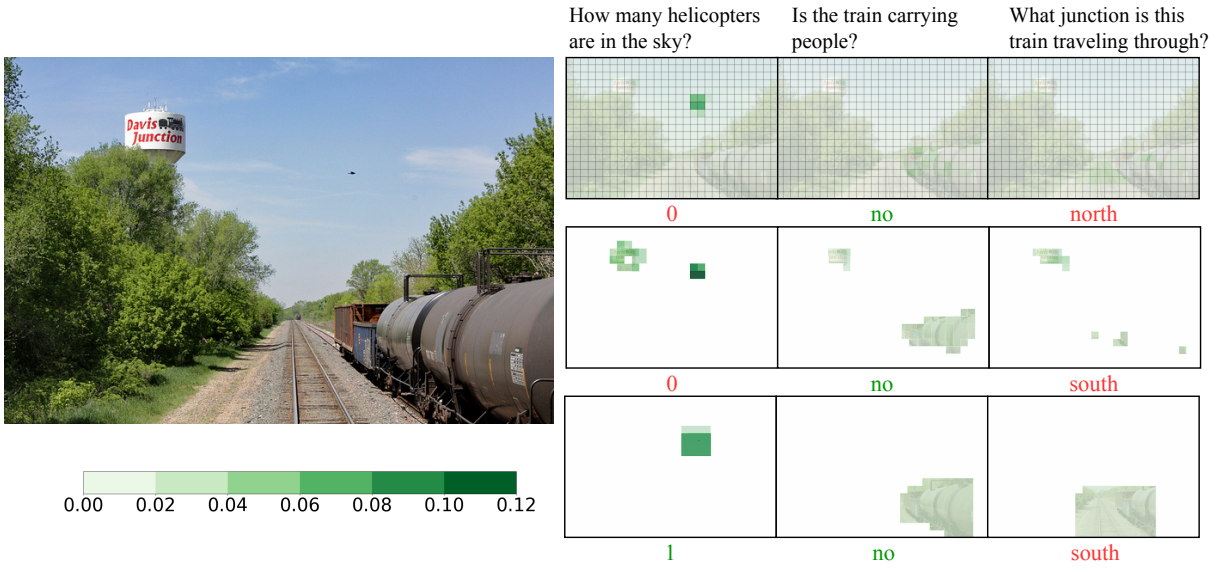

Fig. 5: VQA using softmax (top), sparsemax (middle) and TVMAX attention (bottom). Shading denotes the attention weight, with white for zero attention.

B. HUMAN ATTENTION EXAMPLES

We present in Figure 6 some images of the VQA-v2 validation set with the corresponding human attention from the VQA-HAT dataset and the attention distributions obtained with the different attention mechanisms: softmax, sparsemax, and TVMAX.

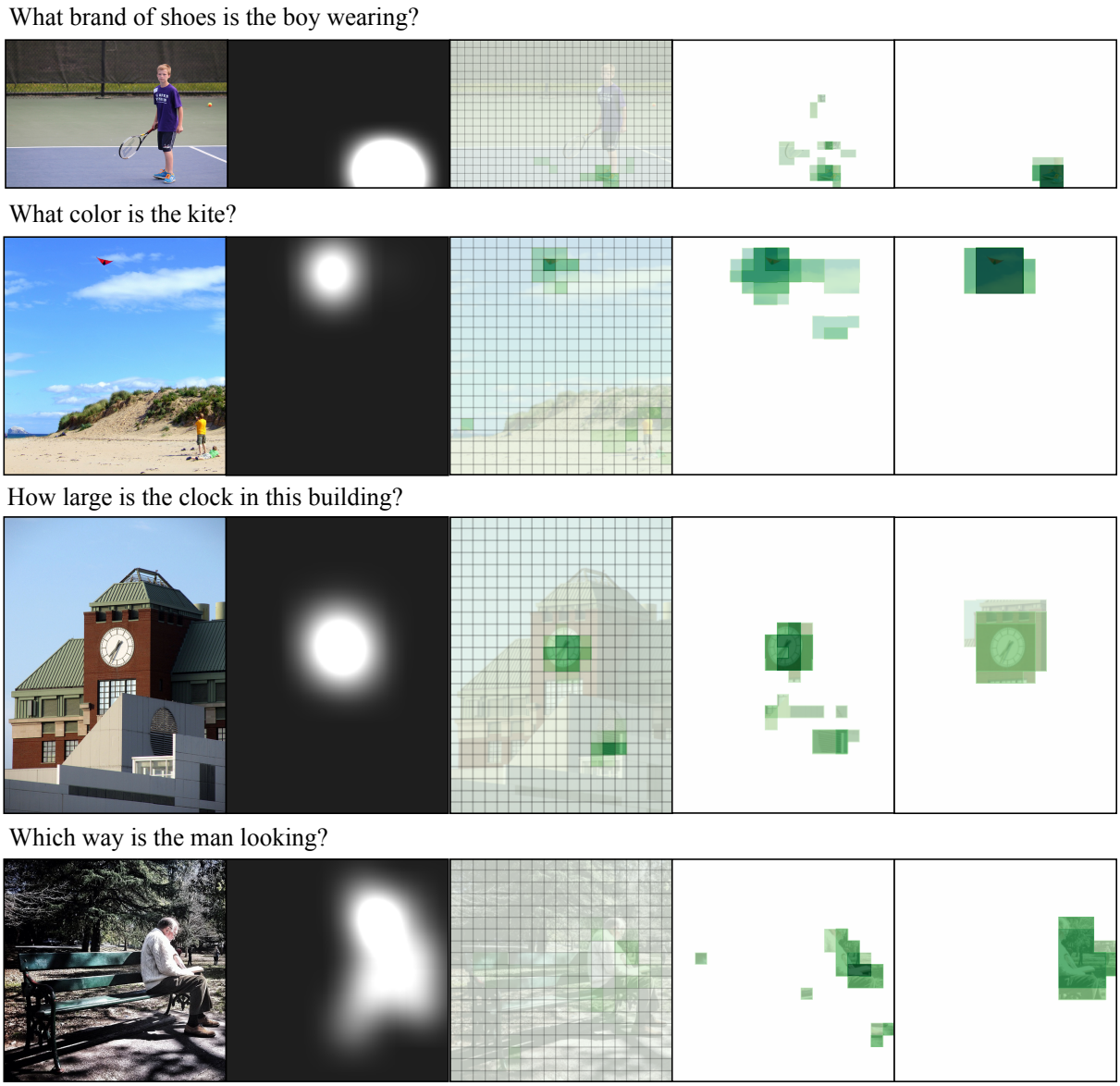

Fig. 6: Examples of human attention and the attention distributions obtained with the different attention mechanisms. The original image in the left, followed by human attention, softmax attention, sparsemax attention, and TVMAX attention.