SPARSE AND STRUCTURED VISUAL ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual attention mechanisms have been widely used in image captioning models. In this paper, to better link the image structure with the generated text, we replace the traditional softmax attention mechanism by two alternative sparsity-promoting transformations: sparsemax and Total-Variation Sparse Attention (TVMAX). With sparsemax, we obtain sparse attention weights, selecting relevant features. In order to promote sparsity and encourage fusing of the related adjacent spatial locations, we propose TVMAX. By selecting relevant groups of features, the TVMAX transformation improves interpretability. We present results in the Microsoft COCO and Flickr30k datasets, obtaining gains in comparison to softmax. TVMAX outperforms the other compared attention mechanisms in terms of human-rated caption quality and attention relevance.

1 INTRODUCTION

The goal of **image captioning** is to generate a fluent textual caption that describes a given image (Farhadi et al., 2010; Kulkarni et al., 2011; Vinyals et al., 2015; Xu et al., 2015). Image captioning is a multimodal task: it combines text generation with the detection and identification of objects in the image, along with their relations. While neural encoder-decoder models have achieved impressive performance in many text generation tasks (Bahdanau et al., 2015; Vaswani et al., 2017; Chorowski et al., 2015; Chopra et al., 2016), it is appealing to design image captioning models where structural bias can be injected to improve their **adequacy** (preservation of the image's information), therefore strengthening the link between their language and vision components.

State-of-the-art approaches for image captioning (Liu et al., 2018a;b; Anderson et al., 2018; Lu et al., 2018) are based on encoder-decoders with visual attention. These models pay attention either to the features generated by convolutional neural networks (CNNs) pretrained on image recognition datasets, or to detected bounding boxes. In this paper, we focus on the former category: **visual attention** over features generated by a CNN. Without explicit object detection, it is up to the attention mechanism to identify relevant image regions, in an unsupervised manner.

A key component of attention mechanisms is the transformation that maps scores into probabilities, with softmax being the standard choice (Bahdanau et al., 2015). However, softmax is **strictly dense**, i.e., it devotes some attention probability mass to *every* region of the image. Not only is this wasteful, it also leads to "lack of focus": for complex images with many objects, this may lead to vague captions with substantial repetitions. Figure 1 presents an example in which this is visible: in the caption generated using softmax (top), the model attends to the whole image at every time step, leading to a repetition of "bowl of fruit." This undesirable behaviour is eliminated by using our alternative solutions: sparsemax (middle) and the newly proposed TVMAX (bottom).

In this work, we introduce novel 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** (Martins & Astudillo, 2016). While sparsemax has been previously used in NLP for attention mechanisms *over words*, it has never been applied to computer vision to attend over *image regions*. With sparsemax, the attention weights obtained are sparse, leading to the selection (non-zero attention) of only a few relevant features. Second, 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** (Tibshirani et al., 2005; Bach et al., 2012). With TVMAX, sparsity is allied to the ability of selecting *compact* regions. According to our human evaluation experiments,

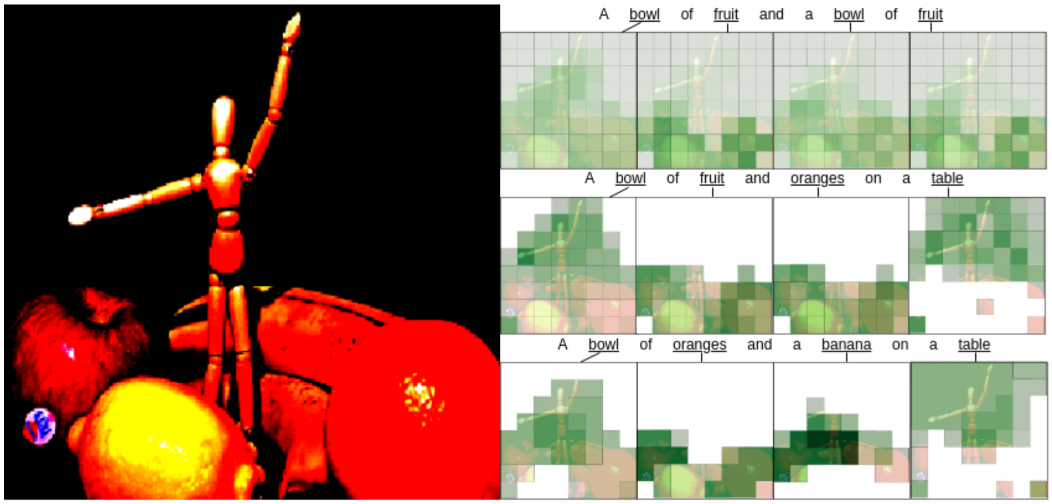

Figure 1: Example of captions generated using softmax (top), sparsemax (middle) and TVMAX attention (bottom). Shading denotes the attention weight, with white for zero attention. The darker the green is, the higher the attention weight is. The full sequences are presented in Appendix C.

this leads to better **interpretability**, since the model's behaviour is better understood by looking at the selected image regions when a particular word is generated. It also leads to a better selection of the relevant features, and consequently to the improvement of the generated captions.

This paper introduces three main contributions:

- We propose a novel visual attention mechanism using sparse attention, based on sparsemax (Martins & Astudillo, 2016), that improves the quality of the generated captions and increases interpretability.
- We introduce a new attention mechanism, TVMAX, that encourages sparse attention over contiguous 2D regions, giving the model the capability of selecting compact objects. We show that TVmax can be evaluated by composing a proximal operator with a sparsemax projection, and we provide a closed-form expression for its Jacobian. This leads to an efficient implementation of its forward and backward pass.
- We perform an empirical and qualitative comparison of the various attention mechanisms considered. We also carry out a human evaluation experiment, taking into account the generated captions as well as the perceived relevance of the selected regions.

2 Selective Visual Attention

Attention mechanisms have the ability to select the relevant features, in this case spatial locations. This requires a mapping from importance scores to a distribution, $z \in \mathbb{R}^k \mapsto p \in \Delta^k$, where $\Delta^k := \left\{ p \in \mathbb{R}^k \mid \sum_{i=1}^{k} p_i = 1, p \geqslant 0 \right\}$ denotes the simplex (the set of all probability distributions over k values). The standard choice for this mapping is softmax, defined as:

$$[\mathrm{softmax}(z)]_i = \frac{\exp(z_i)}{\sum_j \exp(z_j)}. \tag{1}$$

However, as softmax is strictly positive, its output is dense. Thus, the model must pay some attention to the whole image and, consequently, assign lower attention weights to the relevant regions. This motivates our proposed **selective** visual attention mechanisms, which, by being sparse, are able to better isolate the relevant image regions.

2.1 SPARSEMAX

To achieve selective capabilities, we propose the use of sparsemax (Martins & Astudillo, 2016), a sparse mapping consisting in the Euclidean projection of z onto the probability simplex:

$$\text{sparsemax}(z) := \arg\min_{p \in \Delta^k} \frac{1}{2}\|p - z\|_2^2, \tag{2}$$

which allows to obtain sparse outputs with a small increase in complexity. Output sparsity is an attractive property for attention mechanisms, since some features do not provide relevant information for the current prediction. In the image captioning case, using sparsemax allows focusing only on the spatial locations of the image that are relevant to the word being generated, assigning zero attention weight to all other regions.

2.2 SPARSE AND STRUCTURED VISUAL ATTENTION

To generate descriptive captions, the model should identify the objects present in the image. Thus, when generating object-related words, the attention mechanism should assign high weights to the regions of the image containing the object. However, sparsemax is unstructured and index-invariant, leading it to select discontinuous regions. To overcome this, we propose a new visual attention mechanism, TVMAX. TVMAX is a non-trivial generalization of fusedmax (Niculae & Blondel, 2017), a transformation based on fused lasso, to the 2D case. To this end, we first extend fusedmax even more generally, to arbitrary graphs.

2.2.1 GENERALIZED FUSED LASSO

Let $w \in \mathbb{R}^k$, and let $I = \{1, \ldots, k\}$. Consider a graph over I defined by its edges $E \subseteq I \times I$, where an edge between i and j means we want to encourage w_i to be close to w_j. For simplicity we use $i \sim j$ as shorthand for $(i, j) \in E$.

The generalized fused lasso penalty (Tibshirani et al., 2005) is defined as:

$$\Omega_E(w) = \sum_{i \sim j} |w_i - w_j|. \tag{3}$$

Minimizing Ω_E encourages "fused" solutions, i.e., it encourages $w_i = w_j$ for $i \sim j$. In particular, its proximal operator[1] can be seen as a fused signal approximator, seeking a vector w that approximates z well (in terms of Euclidean distance) and that is encouraged to be fused:

$$\text{prox}_{\Omega_E}(z) = \arg\min_{w \in \mathbb{R}^d} \frac{1}{2}\|w - z\|^2 + \Omega_E(w). \tag{4}$$

Computing the value of prox_{Ω_E} is non-trivial in general (Xin et al., 2016), but for certain edge configurations, described below, efficient algorithms exist.

- If E forms a chain, i.e., $i \sim j \iff i = j - 1$, the problem is called *1D total variation* and can be solved in $O(k)$ time using the *taut string algorithm* (Davies & Kovac, 2001; Barbero & Sra, 2014). We use the quasilinear algorithm of Condat (2013), which is very fast in practice.
- If the indices are aligned on a 2D grid, as in an image, $i \sim j$ holds iff. j is to the right or immediately below i, the problem is called *2D total variation*. 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. We may then apply a number of iterative methods, for instance *proximal Dykstra* (Barbero & Sra, 2014)[2].

[1] The proximal operator is defined in Eq. 11 of Appendix A.1.

[2] We use the implementation readily available in the copt library, available at http://openopt.github.io/copt/.

2.2.2 TVMAX

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. TVMAX is defined as follows:

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

$$\mathrm{TVMAX}(z) := \operatorname*{arg\,min}_{p \in \Delta^k} \frac{1}{2}\|p - z\|_2^2 + \lambda \Omega^{\mathrm{TV}}_{2\mathrm{D}}(p); \tag{5}$$

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

Note that Eq. 5 differs from Eq. 4 in which the variable p is further constrained to lie in the probability simplex. We show next how the forward and backward passes can be efficiently computed.

2.2.3 GENERALIZED FUSED SPARSE ATTENTION

To construct *generalized fused sparse attention*, we follow Niculae & Blondel (2017) and define

$$\mathrm{gfusedmax}_E(z) := \operatorname*{arg\,min}_{p \in \Delta} \|p - z\|_2^2 + \Omega_E(p). \tag{6}$$

This can be seen as a *constrained* fused lasso approximator, because the solution p must be a probability distribution vector. While the optimization function is very similar to Eq. 4, the additional constraint that $p \in \Delta$ increases complexity. Fortunately, the following result holds:

Proposition 1 (Computing generalized fusedmax)

$$\mathrm{gfusedmax}_E(z) = \mathrm{proj}_\Delta\left(\mathrm{prox}_{\Omega_E}(z)\right). \tag{7}$$

The proof is given in Appendix A.2.

Proposition 1 also provides a shortcut for deriving the Jacobian of generalized fusedmax via the chain rule: denoting by J_F the Jacobian of prox_{Ω_E}, we have

$$\frac{\partial\,\mathrm{gfusedmax}}{\partial z} = J_{\mathrm{gfusedmax}} = J_{\mathrm{sparsemax}}(\mathrm{prox}_{\Omega_E}(z))\, J_F(z).$$

As we already know how to compute $J_{\mathrm{sparsemax}}$ (Appendix A.1), we may concentrate our effort on deriving the simpler J_F (Eq. 9).

Proposition 2 (Group-wise characterization of prox_{Ω_E}). Let $w^\star := \mathrm{prox}_{\Omega_E}$, and denote by G_i the set of indices fused to w_i in the solution. G_i may be defined recursively:

1. $i \in G_i$ for all i, and
2. $j \in G_i$ if there exists $m \in G_i$ such that $m \sim j$ and $w^\star_m = w^\star_j$.

Define $s_{ij} = \mathrm{sign}(w^\star_i - w^\star_j)$. Then, the solution has the expression

$$w^\star_i = \frac{1}{|G_i|}\left(\sum_{j \in G_i} z_j + \sum_{\substack{m \sim j \\ m \notin G_i}} s_{mj} - \sum_{\substack{j \sim m \\ m \notin G_i}} s_{jm}\right). \tag{8}$$

Proposition 2 shows how to easily compute a generalized Jacobian of gfusedmax since small perturbations in z never change the groups G_i nor the signs of across-group differences s_{ij}, differentiating Eq. 8 yields

$$J_{F\,i,j} = \frac{\partial w^\star_i}{\partial z_j} = \begin{cases} \frac{1}{|G_i|}; & j \in G_i; \\ 0; & j \notin G_i. \end{cases} \tag{9}$$

This generalizes Lemma 1 of Niculae & Blondel (2017) to generalized fused lasso, with a simpler proof, given in Appendix A.3.

2.2.4 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. 7 we have that TVMAX's Jacobian can be computed as $J_{\text{TVMAX}} = J_{sp}(\text{prox}_{\text{TV}_{2D}}(z))J_{tv}(z)$, where J_{sp} is the sparsemax's Jacobian and J_{tv} is the Jacobian of the Total Variation proximal operator.[3] As derived in Proposition 2, $(J_{tv})_{i,j} = \frac{1}{n_{ij}}$ if i and j are fused in a group with n_{ij} elements, and 0 otherwise.

The backward pass intuitively involves "spreading" the credit assigned to one image location evenly across all locations fused with it. This can be implemented by Algorithm 1 in $O(k^2 \log k)$ for the worst case. When positions are fused it is much faster. The algorithm is inspired by flood filling algorithms (Burtsev & Kuzmin, 1993).

Algorithm 1 TVMAX backward pass (Jacobian-vector products)

Input: $p = \text{TVMAX}(z)$, $dp \in \mathbb{R}^k$.
Output: $dz = J^{\top}_{\text{TVMAX}}(dp) \in \mathbb{R}^k$ — # chain rule
Initialize: $N \leftarrow \emptyset$ — # neighbours stack
    $V \leftarrow \emptyset$ — # visited positions
    $G \leftarrow \emptyset$ — # current group
    $s = 0$ — # intermediate value used for J_{TVMAX}'s computation
$dw \leftarrow (J^{sp})^{\top} dp$ — # Eqs. 14 and 15 of A.1
while $|V| < k$ **do** — # check if all positions have been visited
    pick $(i_0, j_0) \notin V$, push (i_0, j_0) to N — # get not visited position and add it to neighbours stack
    while N not empty **do**
        pop (i, j) from N — # get element from neighbours stack
        if $p_{i,j} = p_{i_0,j_0}$ **then** — # check if element is fused
            $G \leftarrow G \cup \{(i, j)\}$; $V \leftarrow V \cup \{(i, j)\}$ — # add neighbour to group and to visited positions
            $s \leftarrow s + (dw)_{i,j}$ — # sum of the dw of each element of the group
            for all neighbours (i', j') of (i, j) **do**
                if $(i', j') \notin V$ **then** push (i', j') to N — # add not visited neighbours of (i, j) to the stack
    if G not empty **then**:
        $(dz)_{i,j} \leftarrow \frac{s}{|G|}$ for all $(i, j) \in G$ — # compute J_{TVMAX} for elements in group G
        $G \leftarrow \emptyset$
        $s = 0$

3 IMAGE CAPTIONING MODEL

To compare the proposed attention mechanisms, we use a straight-forward simple encoder-decoder model with visual attention, inspired by Liu et al. (2018a). The model is sketched in Figure 2.

Given an image, we use a residual CNN pretrained on ImageNet (He et al., 2016; Russakovsky et al., 2014) to get a feature map with spatial dimension of size 8×8 and channel dimension of size 2048, that go through a fine-tuned feedforward layer yielding $g = 512$ feature maps. The visual feature matrix $V = [v_1, v_2, \ldots, v_k]$, with $v_i \in \mathbb{R}^g$ and $k = 64 = 8 \times 8$, contains the image information used to generate the corresponding caption. Following Liu et al. (2018a), we use input and output attention to select the relevant features for the current generation. To generate the word at position t, the input attention, α_t, is computed using the LSTM's previous hidden state $h_{t-1} \in \mathbb{R}^d$. First, a similarity score $z_{t,i}$, $i \in \{1, \ldots, k\}$, is computed between h_{t-1} and the i^{th} image cell via a feedforward transformation (Bahdanau et al., 2015), as $z_{t,i} = w^{\top} \tanh(\text{affine}([v_i; h_{t-1}]))$, for all k image cells. Then, α_t is obtained by normalizing the k-dimensional vector of scores z_t with softmax, $\alpha_t = \text{softmax}(z_t)$. Using these attention weights, a vector representation of the image to be used as input of the LSTM, is obtained, $s_t = V\alpha_t$. The output attention e_t, is computed in the same way as above, but applied to the current LSTM hidden state h_t, instead of

[3] J_{tv} is a special case of J_F, when using the graph $\mathcal{G}$ that consists in a 2D grid.

h_{t-1}, and normalized with the different proposed transformations. This produces output visual features $s_t = V e_t$, which are passed through a feedforward layer to yield the image representation $r_t = \tanh(\mathrm{affine}(s_t))$. Finally, the predictive probability of the next word is:

$$P(y_t \mid y_{1:(t-1)}; \mathrm{Image}) \propto \mathrm{softmax}(\mathrm{affine}([r_t; h_t])). \tag{10}$$

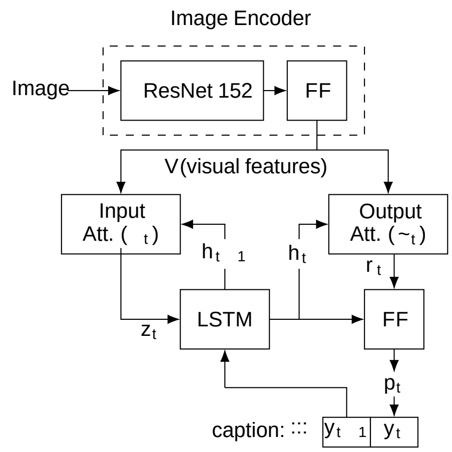

Figure 2: Diagram of the caption generation network.

4 Experiments

Settings. The input images are resized to 256×256 before going through the residual CNN and the feature maps obtained have a size of 8×8. We use an LSTM hidden size of $d = 512$ and a word embedding size of 256, for all models. The models were trained for 50 epochs using the Adam optimizer (Kingma & Ba, 2014) with a learning rate of 0.0001 and a decay of 0.8 and 0.999 for the first and second momentum, respectively. After the 10^{th} epoch, the learning rate starts decaying with a decay factor of 0.99. For TVmax, we set $\lambda = 0.01$.

Datasets and Metrics. We report our results on the Microsoft COCO (MSCOCO) and Flickr30k datasets. MSCOCO is composed of 113,287 images of common objects in context while Flickr30k consists in 31,000 pictures of people involved in everyday activities and events. Each image is annotated with 5 captions. We use the split proposed by Karpathy & Fei-Fei (2015), which stipulates equal validation and test sizes of 5,000 images (MSCOCO) and 1,000 (Flickr30k). The metrics we report are SPICE (Anderson et al., 2016), CIDEr (Vedantam et al., 2015), longest common subsequence ROUGE, (denoted ROUGE$_L$, Lin, 2004), 1– to 4–gram BLEU (denoted BLEU$_4$, Papineni et al., 2002), and METEOR (Banerjee & Lavie, 2005). To investigate whether selective attention alleviates repetition, we also measure the n-gram repetition metric REP (Malaviya et al., 2018).

Table 1: Automatic evaluation of caption generation on MSCOCO and Flickr30k.

	MSCOCO						Flickr30k					
	SPICE	CIDER	ROUGE$_L$	BLEU$_4$	METEOR	REP$\downarrow$	SPICE	CIDER	ROUGE$_L$	BLEU$_4$	METEOR	REP$\downarrow$
softmax	18.4	0.967	52.9	29.9	24.9	3.76	13.5	0.443	44.2	19.9	19.1	6.09
sparsemax	18.9	0.990	53.5	31.5	25.3	3.69	13.7	0.444	44.3	20.7	19.3	5.84
TVmax	18.5	0.974	53.1	29.9	25.1	3.17	13.3	0.438	44.2	20.5	19.0	3.97

Automated metrics. As can be seen in table 1, overall sparsemax and TVmax attention mechanisms achieve better results when compared with softmax, indicating that the use of selective attention leads to better captions. This improvement does not come at a high computational cost: at

Figure 3: Example of captions generated using softmax (top), sparsemax (middle) and TVMAX attention (bottom). Shading denotes the attention weight, with white for zero attention. The darker the green is, the higher the attention weight is. The full sequences are presented in Appendix C.

inference time, models using TVMAX and sparsemax are only 1.3x and 1.1x slower than softmax. Moreover, for TVMAX, automatic metrics results are slightly worse than sparsemax but still superior to softmax on MSCOCO and similar on Flickr30k. We show next that this is compensated with fewer repetitions and higher scores in the human evaluation of the captions and attention relevance.

Table 2: Human evaluation results with different attention mechanisms on MSCOCO.

	CAPTION (1-5)	ATTENTION RELEVANCE (1-5)
softmax	3.50	3.38
sparsemax	3.71	3.89
TVMAX	3.87	4.10

Human rating. The caption evaluation consisted in attributing a score from 1 to 5 to the caption of each model while the attention evaluation consisted in scoring the relevancy of the attended areas, from 1 to 5, when generating the non stop words of the captions. A full description of the human assessment can be found in Appendix B.

Despite performing slightly worse than sparsemax under automated metrics, TVMAX outperforms sparsemax and softmax in the caption human evaluation and the attention relevance human evaluation, reported in Table 2. The superior score on attention relevance shows that TVMAX is better at selecting the relevant features and its output is more interpretable. Additionally, the better caption evaluation results demonstrate that the ability to select compact regions induces the generation of better captions. We next explore possible explanations for the TVMAX superior results.

Repetition. Figure 1 illustrates that softmax attention is prone to spuriously repeating references to the same object. Selective attention mechanisms like sparsemax and especially TVMAX reduce repetition, as measured by the REP metric reported in Table 1. This expected success can be attributed to their ability to select compact regions exclusively and can be one of the causes of the human evaluation results. This happens even though TVMAX generates longer sentences than sparsemax and softmax (9.5 against 9.0 words on average) and shows the bene t of promoting structured and sparse attention simultaneously.

Object detection. Using the MSCOCO object detection ground truth, we compared the percentage of objects present in the image that are referred to in the captions, using each attention mechanism. With TVMAX 28.2% of the reference objects are referred, against 27.5% and 27.4% for sparsemax and softmax, repectively. This shows that promoting high attention to groups of spatial locations of the image leads to a more precise identi cation of the objects.

Sparsity. The average image area that receives zero attention is 34% for sparsemax and 25% for TVMAX. To illustrate where the models attend to, we display the output attention in Figures 1 and 3. As expected, softmax weights are spread widely across the image, ending up missing the relevant regions. In contrast, sparsemax and TVMAX weights are zero for the non-relevant spatial locations.

Qualitative comparison. As the image of Figure 1 contains various similar objects, the softmax model (top) generates a incoherent, repetition-laden caption. In contrast, the sparsemax (middle) and TVMAX (bottom) models better identify the relevant parts of the image, generating coherent and descriptive captions. Moreover, the groups obtained with TVMAX are clearly visible and more aligned to object boundaries, offering better interpretability, as revealed by human attention assessment. In Figure 3 it can also be noticed that with TVMAX (bottom) the model correctly identi ed "a group of people" instead of "a soccer player" as with sparsemax (middle) and softmax (top). This indicates its superior ability to correctly de ne the relevant groups of features and that this ability leads to improved captions.

5 RELATED WORK

Image captioning. In the last years, neural models with visual attention mechanisms have been receiving increased interest. Several researchers have been studying diverse attention mechanisms in order to re ne visual information for image captioning. Xu et al. (2015) proposed the use of hard attention, which only attends to one region at each step. However, to generate descriptive captions the model should, often, focus on more than one region. In addition, hard attention is non-differentiable, requiring imitation learning or Monte Carlo policy gradient approximations.Anderson et al. (2018) proposed bottom-up attention, using an object detection model designed to identify bounding boxes of objects, and top-down attention, selecting the relevant bounding-boxes. Wang et al. (2019) proposed an hierarchical attention network composed by a patch detector, object detector, and concept detector. Using object detection models is less demanding on the attention mechanism, since it only has to select the boxes the model should attend to. However, such models are limited by the bounding boxes position's accuracy. Gao et al. (2019) introduced a deliberate attention network to re ne the attended visual features. Yet, the attention distribution remained dense.

Sparse attention. In several tasks only a few features are relevant for the current prediction. This can be attained when using sparse attention. Various prior works have proposed sparse attention mechanisms with promising results, (Xu et al., 2015; Martins & Astudillo, 2016; Malaviya et al., 2018; Peters et al., 2019). Niculae & Blondel (2017) proposed 1D fusedmax, which incorporates the fused lasso, so that adjacent words are encouraged to have the same attention weight. In this work, the authors were able to improve interpretability without sacri cing performance, obtaining superior results on textual entailment and summarization. We derive a generalized fused attention mechanism, extending 1D fusedmax.

6 CONCLUSIONS AND FUTURE WORK

We propose using sparse and structured visual attention, in order to improve the process of selecting the features relevant to the caption generation. For that, we used sparsemax and introduced TVMAX. Results on the image captioning task, show that the attention mechanism is able to select better features when using sparsemax or TVMAX. Furthermore, in the human assessment and attention analysis we see that the improved selection of the relevant features as well as the ability to group spatial features lead to the generation of better captions, while improving the model's interpretability.

In future work, TVMAX attention can be applied to other multimodal problems such as visual question answering. It can also be applied in other tasks for which we have prior knowledge of the data's stucture, for instance graphs or trees.

REFERENCES

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic propositional image caption evaluation. In *Proc. ECCV*, 2016.

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.

Francis Bach, Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, et al. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4(1):1–106, 2012.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proc. ICLR*, 2015.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proc. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005.

Álvaro Barbero and Suvrit Sra. Modular proximal optimization for multidimensional total-variation regularization. *preprint arXiv:1411.0589*, 2014.

Sergei Burtsev and Ye.P. Kuzmin. An efficient flood-filling algorithm. *Computers & Graphics*, 17: 549–561, 09 1993.

Sumit Chopra, Michael Auli, and Alexander M Rush. Abstractive sentence summarization with attentive recurrent neural networks. In *Proc. NAACL-HLT*, 2016.

Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. In *Proc. NeurIPS*, 2015.

Laurent Condat. A direct algorithm for 1-d total variation denoising. *IEEE Signal Processing Letters*, 20(11):1054–1057, 2013.

P. Laurie Davies and Arne Kovac. Local extremes, runs, strings and multiresolution. *The Annals of Statistics*, 29(1):1–48, 2001. ISSN 00905364.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *Proc. ECCV*, 2010.

Jerome Friedman, Trevor Hastie, Holger Höfling, and Robert Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007.

Lianli Gao, Kaixuan Fan, Jingkuan Song, Xianglong Liu, Xing Xu, and Heng Tao Shen. Deliberate attention networks for image captioning. 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, 2016.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *preprint arXiv:1412.6980*, 2014.

Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. Baby talk: Understanding and generating image descriptions. In *Proc. CVPR*, 2011.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004.

Fenglin Liu, Xuancheng Ren, Yuanxin Liu, Houfeng Wang, and Xu Sun. simNet: Stepwise image-topic merging network for generating detailed and comprehensive image captions. *Preprint arXiv:1808.08732*, 2018a.

Xihui Liu, Hongsheng Li, Jing Shao, Dapeng Chen, and Xiaogang Wang. Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. *preprint arXiv:1803.08314*, 2018b.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *Proc. CVPR*, 2018.

Chaitanya Malaviya, Pedro Ferreira, and André FT Martins. Sparse and constrained attention for neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 370–376, 2018.

Andre Martins and Ramon Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proc. ICML*, 20–22 Jun 2016.

Vlad Niculae and Mathieu Blondel. A regularized framework for sparse and structured neural attention. In *Proc. NeurIPS*, 2017.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proc. ACL*, 2002.

Ben Peters, Vlad Niculae, and André FT Martins. Sparse sequence-to-sequence models. *In Proc. ACL*, 2019.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2014.

Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. NeurIPS*, 2017.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *Proc. CVPR*, 2015.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proc. CVPR*, 2015.

Weixuan Wang, Zhihong Chen, and Haifeng Hu. Hierarchical attention network for image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 8957–8964, 2019.

Bo Xin, Yoshinobu Kawahara, Yizhou Wang, Lingjing Hu, and Wen Gao. Efficient generalized fused lasso and its applications. *ACM Trans. Intell. Syst. Technol.*, 7(4):60:1–60:22, May 2016. ISSN 2157-6904.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proc. ICML*, 2015.

Yaoliang Yu. On decomposing the proximal map. In *Proc. NeurIPS*, 2013.

A FORWARD AND BACKWARD PASS OF 2D FUSEDMAX ATTENTION.

A.1 PRELIMINARIES

The **proximal operator** of a function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is defined as

$$\mathrm{prox}_f(z) = \operatorname*{arg\,min}_{w \in \mathbb{R}^d} f(z) + \frac{1}{2}\|z - w\|_2^2, \tag{11}$$

and it is guaranteed to have a unique solution, thanks to the strong convexity of the Euclidean distance.

The **indicator function** of a set $C \subseteq \mathbb{R}^d$ is the function

$$\iota_C : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}, \qquad \iota_C(w) := \begin{cases} 0, & w \in C, \\ \infty, & w \notin C. \end{cases} \tag{12}$$

The **projection onto a convex set** $C \subseteq \mathbb{R}^d$ is defined as

$$\mathrm{proj}_C(z) := \operatorname*{arg\,min}_{w \in C} \frac{1}{2}\|z - w\|_2^2 = \mathrm{prox}_{\iota_C}(z), \tag{13}$$

showing that the proximal operator can be seen as a generalization of projection.

The **sparsemax** attention mapping (Martins & Astudillo, 2016) is the projection onto the simplex,

$$\mathrm{sparsemax}(z) := \mathrm{proj}_{\triangle}(z) = \operatorname*{arg\,min}_{p \in \triangle} \frac{1}{2}\|p - z\|^2. \tag{14}$$

A necessary component for using sparsemax for attention is its Jacobian, the matrix of its partial derivatives $(J_{\mathrm{sparsemax}})_{i,j} = \frac{\partial\, \mathrm{sparsemax}(z)_i}{\partial z_j}$. Martins & Astudillo (2016) derive its expression

$$J_{\mathrm{sparsemax}}(z) = \mathrm{diag}\, s - \frac{1}{\|s\|_1} s s^\top, \tag{15}$$

where $s_j = 1$ if $\mathrm{sparsemax}(z)_j > 0$ and $s_j = 0$ otherwise.

A.2 PROOF OF PROPOSITION 1

Proof. This result is a slight extension of Proposition 2 in Niculae & Blondel (2017), and also follows from Corrolary 4 of Yu (2013), by taking $f = \iota_{\triangle}$, and noting that $\iota_{\triangle}$ is symmetric: if $p \in \triangle$, then any vector p' obtained by permuting p is also in $\triangle$, because its values remain non-negative and sum to 1. ☐

A.3 PROOF OF PROPOSITION 2

Let $w^\star := \mathrm{prox}_{\Omega_E}$, and denote by G_i the set of indices fused to w_i in the solution. Define $s_{ij} = \mathrm{sign}(w_i^\star - w_j^\star)$.

Proof. The subgradient optimality conditions of Eq. 4 are: (Friedman et al., 2007)

$$w_i^\star - z_i + \sum_{k : i \sim k} t_{ik} - \sum_{k : k \sim i} t_{ki} = 0, \qquad 1 \leq i \leq d. \tag{16}$$

where $t_{ij} = \mathrm{sign}(w_i^\star - w_j^\star)$ if $w_i^\star \neq w_j^\star$, otherwise t_{ij} is a free variable in $[-1, 1]$.

We focus on a single group $G = G_i$, dropping the index i for brevity. Within a fused group, the solution is constant, i.e., $w_j^\star = w$ for $j \in G$. We separate the sums in Eq. 16 according to whether $k \in G$ or not, and move the "constant" terms to the right hand side, yielding the system

$$w + \sum_{\substack{j \sim k \\ k \in G}} t_{jk} - \sum_{\substack{k \sim j \\ k \in G}} t_{kj} = z_j + \sum_{\substack{k \sim j \\ k \notin G}} s_{kj} - \sum_{\substack{j \sim k \\ k \notin G}} s_{jk}, \qquad j \in G. \tag{17}$$