Sparse and Continuous Attention Mechanisms

André Martins[1,3,5] António Farinhas[1] Marcos Treviso[1] Vlad Niculae[4,1] Pedro Aguiar[2,3] Mário Figueiredo[1,3]

[1]Instituto de Telecomunicações, Instituto Superior Técnico, Lisbon, Portugal [2]Instituto de Sistemas e Robótica, Instituto Superior Técnico, Lisbon, Portugal
[3]LUMLIS (Lisbon ELLIS Unit), Lisbon, Portugal [4]Informatics Institute, University of Amsterdam, The Netherlands [5]Unbabel, Lisbon, Portugal

Outline

Attention mechanisms are a powerful component in neural networks.
- Key to recent successes in MT, NLP, and vision tasks.
- So far: attention over a **finite set** (words, pixel regions, etc.)

This work: We generalize attention to *arbitrary sets*, possibly continuous.

This Paper: From Discrete to Continuous Attention

(Bahdanau et al., 2015, ICLR)

Finite set $S = \{1, \dots, L\}$

Three ingredients:
- Score vector $f \in \mathbb{R}^L$
- Transformation from f to probability vector $p \in \triangle^L$
- Value matrix $V \in \mathbb{R}^{D \times L}$

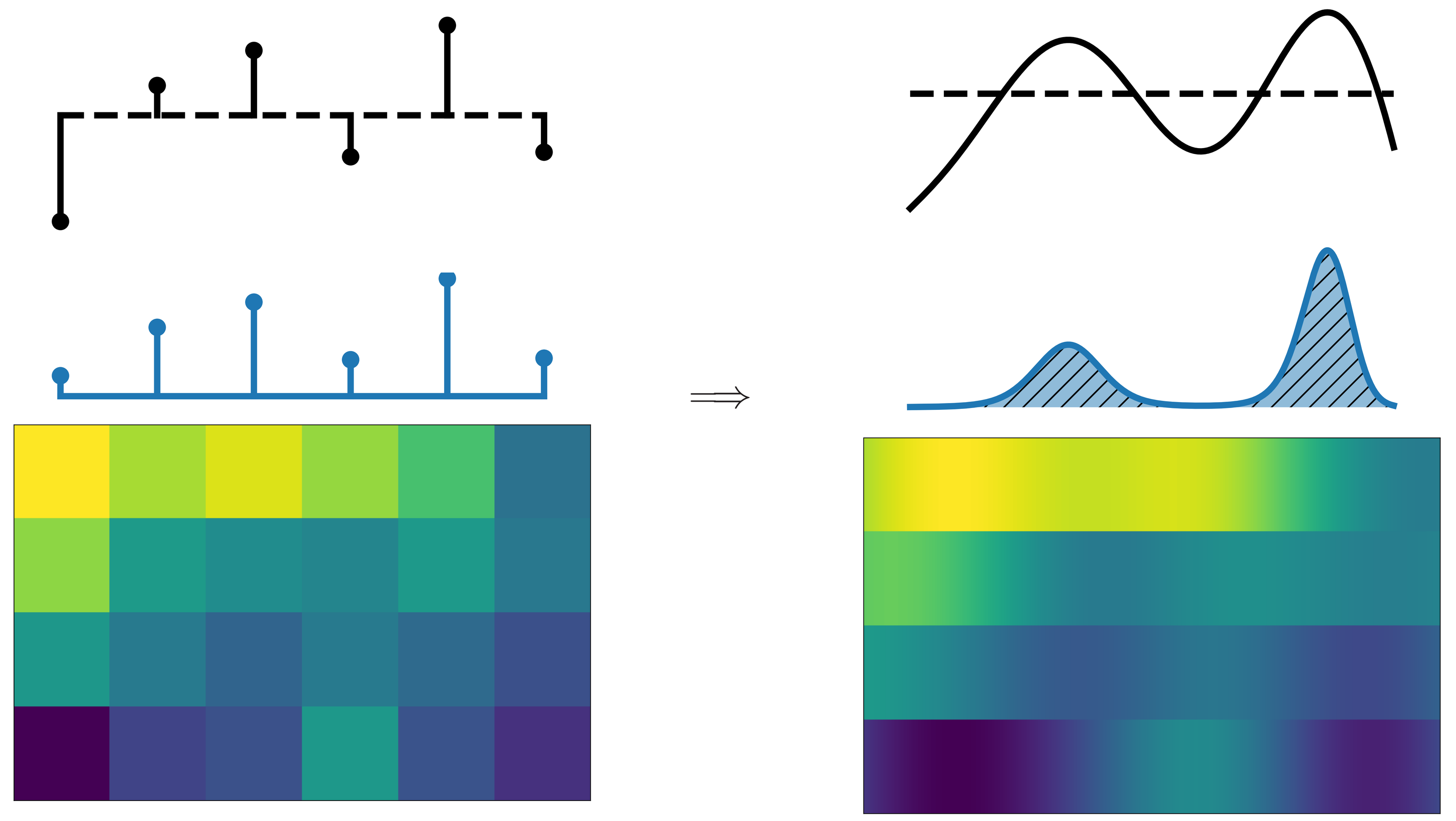
Output:
- Weighted average $Vp \in \mathbb{R}^D$

Our work:

Measure space S (*e.g.* continuous)

Three ingredients:
- Score *function* $f : S \to \mathbb{R}$
- Transformation from f to *density* $p : S \to \mathbb{R}_+$, $\int_S p = 1$
- Value *function* $V : S \to \mathbb{R}^D$

Output:
- $\mathbb{E}_p[V(t)] = \int_S p(t)V(t) \in \mathbb{R}^D$



Score Function $f(t)$ and Value Function $V(t)$

Score function: Parametrized as $f_\theta(t) = \theta^\top \phi(t)$, where
- $\phi(t) \in \mathbb{R}^M$ are *basis functions*.
- Parameters $\theta \in \mathbb{R}^M$ are output by a neural network.

Example: $\phi(t) = [t, \mathrm{vec}(tt^\top)]$ and $\theta = [\Sigma^{-1}\mu, \mathrm{vec}(-\frac{1}{2}\Sigma^{-1})]$ lead to a quadratic form

$$f_\theta(t) = -\frac{1}{2}(t-\mu)^\top \Sigma^{-1}(t-\mu).$$

Value function: Parametrized as $V_B(t) = B\psi(t)$, where
- $\psi(t) \in \mathbb{R}^N$ are *basis functions* (e.g., Gaussian RBFs)
- $B \in \mathbb{R}^{D \times N}$ fit to measurements by ridge regression (see paper).

Ω-Regularized Prediction Map (Ω-RPM)

Transforms **score function** f into **probability density** $p \equiv \hat{p}_\Omega[f]$:

$$\hat{p}_\Omega[f] = \operatorname*{argmax}_p \mathbb{E}_p[f(t)] - \Omega(p), \qquad \Omega \text{ convex regularizer.}$$

$-\Omega$ **Shannon/differential entropy** $\Longrightarrow$ **softmax/Gibbs distributions** (exponential families):

$$\hat{p}_\Omega[f] = \mathrm{softmax}(f), \qquad \hat{p}_\Omega[f](t) = \exp(f(t) - \tau)$$

$-\Omega_\alpha$ **Tsallis α-entropy** $\Longrightarrow$ **α-entmax** (deformed exponential family):

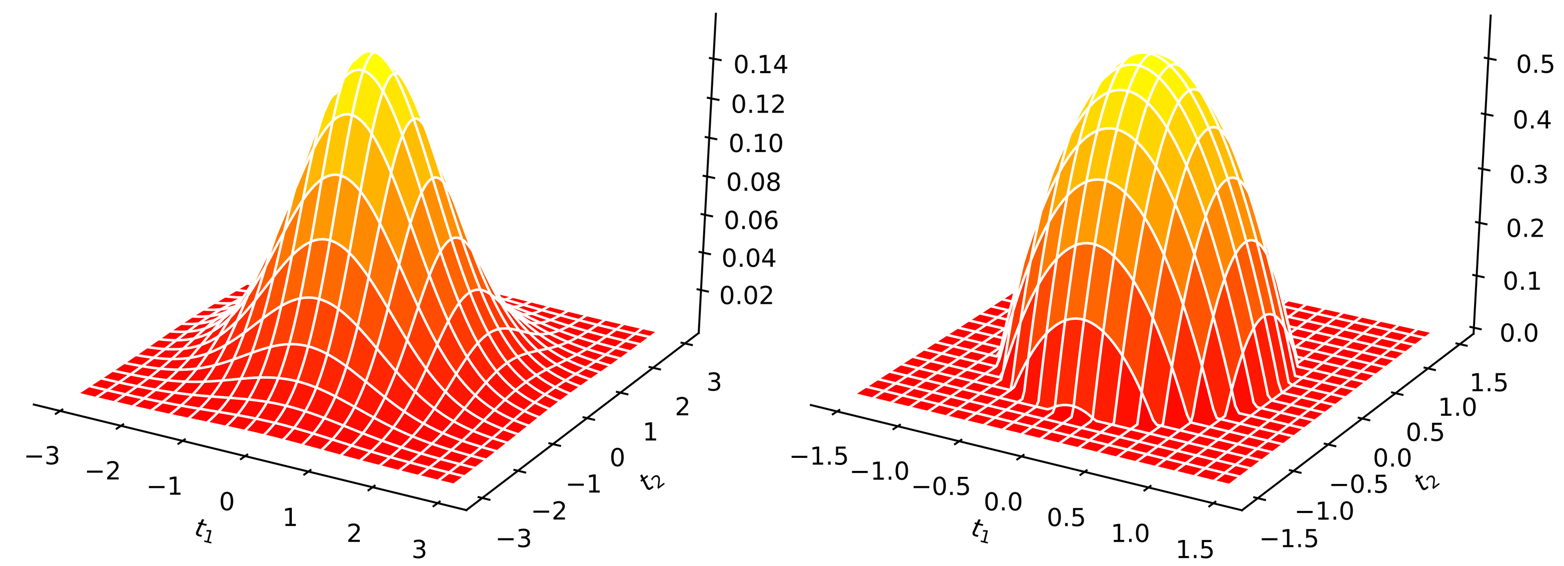
$$\hat{p}_{\Omega_\alpha}[f](t) = \left[1 + (\alpha - 1)(f(t) - \tau)\right]_+^{\frac{1}{\alpha-1}}$$

Particular cases: (continuous) softmax ($\alpha = 1$) and sparsemax ($\alpha = 2$).

Blondel et al. (2020, JMLR),Martins and Astudillo (2016, ICML), Peters et al. (2019, ACL), Tsallis (1988)

Example: Gaussian and Truncated Paraboloid (2D)

With $t \in \mathbb{R}^2$ and $f(t) = -\frac{1}{2}(t-\mu)^\top\Sigma^{-1}(t-\mu)$:



(continuous softmax, $\alpha = 1$) (continuous sparsemax, $\alpha = 2$)

Truncated paraboloid has sparse, varying support!

Key Result I: Forward Pass

Assuming
- Quadratic score $f_\theta(t) = \theta^\top\phi(t) = -\frac{1}{2}(t-\mu)^\top\Sigma^{-1}(t-\mu)$
- Value function $V_B(t) = B\psi(t)$ where $\psi(t)$ are Gaussian RBFs

Then:

Continuous softmax ($\alpha = 1$):
- $\mathbb{E}_{\hat{p}_\Omega[f_\theta]}[V_B(t)]$ becomes product of Gaussians $\Longrightarrow$ closed form.

Continuous sparsemax ($\alpha = 2$):
- $\mathbb{E}_{\hat{p}_\Omega[f_\theta]}[V_B(t)]$ closed form in 1D, easy to compute numerically in 2D.

Key Result II: Backprop

How to backpropagate?
For any α, Jacobian is a "generalized covariance" (see paper):

$$\frac{\partial \mathbb{E}_{\hat{p}_\Omega[f_\theta]}[V_B(t)]}{\partial \theta} = B^\top \mathrm{cov}_{\hat{p}_\Omega[f_\theta], 2-\alpha}(\phi(t), \psi(t)).$$

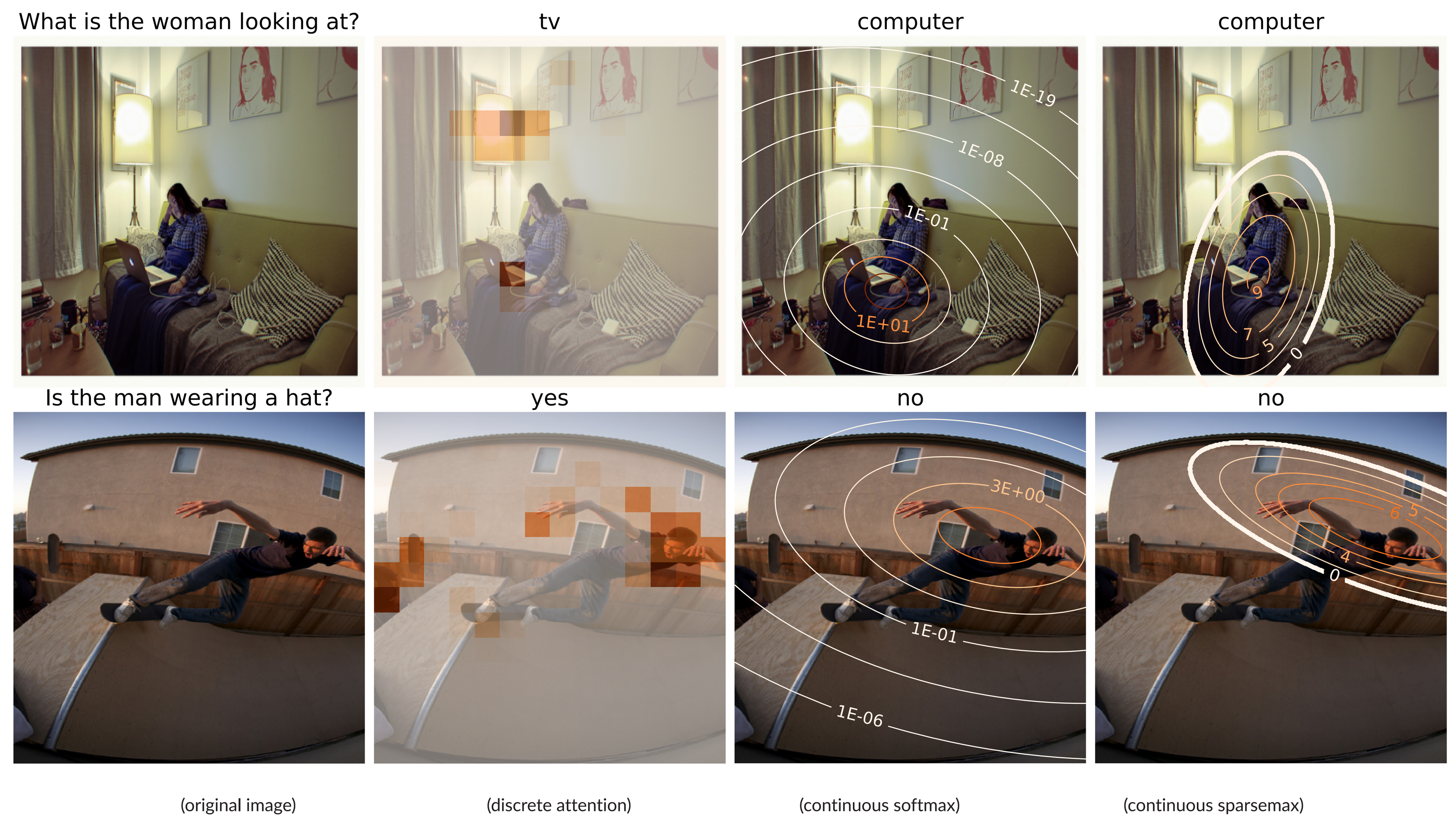
Also tractable for the two cases above.

Experiments

- 1D continuous attention for NLP (document classification and NMT)
- 2D continuous attention for vision (VQA-v2)

	Doc. Class. IMDB (%)	NMT De-En IWSLT (BLEU)	VQA-v2 Test-Dev (%)	VQA-v2 Test-Std (%)
Discrete softmax	90.78	23.92	65.83	66.13
Continuous softmax	90.98	24.00	**65.96**	**66.27**
Continuous sparsemax	**91.10**	**24.25**	65.79	66.10

VQA: Attention Maps



Conclusions

- We generalized attention and Ω-RPMs to continuous domains
- When Ω is a α-Tsallis regularizer: continuous and *sparse* densities
- Forward and backprop efficient for $\alpha \in \{1, 2\}$ with quadratic scores and Gaussian RBFs
- Proof of concept (1D/2D): document classification, NMT, and VQA.
- **Future work:** Multimodal attention (*mixtures* of Gaussians or TPs)

Open-source code:

https://github.com/deep-spin/mcan-vqa-continuous-attention

References

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proc. of ICLR*.

Blondel, M., Martins, A. F., and Niculae, V. (2020). Learning with fenchel-young losses. *Journal of Machine Learning Research*, 21(35):1–69.

Martins, A. F. and Astudillo, R. F. (2016). From softmax to sparsemax: A sparse model of attention and multi-label classification. In *Proc. of ICML*.

Peters, B., Niculae, V., and Martins, A. F. (2019). Sparse sequence-to-sequence models. In *Proc. of ACL*.

Tsallis, C. (1988). Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, 52:479–487.

Acknowledgments: ERC StG DeepSPIN 758969 "Deep Structured Prediction in NLP", P2020 program MAIA (contract 045909), FCT (contract UIDB/50008/2020).