Multimodal Continuous Visual Attention Mechanisms

António Farinhas[1] André F. T. Martins[1,3,4] Pedro M. Q. Aguiar[2,3]

[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]Unbabel, Lisbon, Portugal

Outline

Visual attention mechanisms are an important component of deep learning models.

- Most models for visual attention operate over discrete domains (Bahdanau et al., 2015).
- Recently, **continuous attention mechanisms** have been proposed, limiting the attention to a simple unimodal density (Martins et al., 2020).

This paper: we introduce **multimodal** continuous attention mechanisms.

From Discrete to Continuous Attention

Discrete Attention

Images are represented using L feature vectors in $\mathbb{R}^D$ (*e.g.*, grid-level or object-level representations).

- Feature matrix $V \in \mathbb{R}^{D \times L}$
- Score vector $f = [f_1, \dots, f_L]^\top \in \mathbb{R}^L$
- Probability vector via $p = \mathrm{softmax}(f)$

Output:

- Weighted average $c = Vp \in \mathbb{R}^D$

How many planes are in this photograph?
2

$\Longrightarrow$

Continuous Attention

Images are represented as smooth functions in 2D.

- Feature function $V_B(x) = B\psi(x)$
- Score function $f(x) = -\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)$
- Probability density $p(x) = \mathcal{N}(x; \mu, \Sigma)$

Output:

- $c = \mathbb{E}_p[V_B(x)] = B\int_{\mathbb{R}^2} p(x)\psi(x) \in \mathbb{R}^D$

How many planes are in this photograph?
3

This paper: multimodal continuous attention

We let the attention density be a mixture of unimodal distributions, specifically Gaussians

$$p(x) = \sum_{k=1}^{K} \pi_k p_k(x). \quad (1)$$

Forward step. The context vector is a mixture of the context representations for each component,

$$c = \mathbb{E}_p[B\psi(x)] = \sum_{k=1}^{K} \pi_k \underbrace{\mathbb{E}_{p_k}[B\psi(x)]}_{c_k} = \sum_{k=1}^{K} \pi_k c_k. \quad (2)$$

Backward step. Linear combination of unimodal attention mechanisms.

How many planes are in this photograph?
2

The EM algorithm with weighted data

Parameters: Centers of grid regions and weights $\mathcal{X} = \{(x_\ell, w_\ell)\}_{\ell=1}^{L}$, initialization $\Theta(K) = \{(\pi_k, \mu_k, \Sigma_k)\}_{k=1}^{K}$, iterations I.

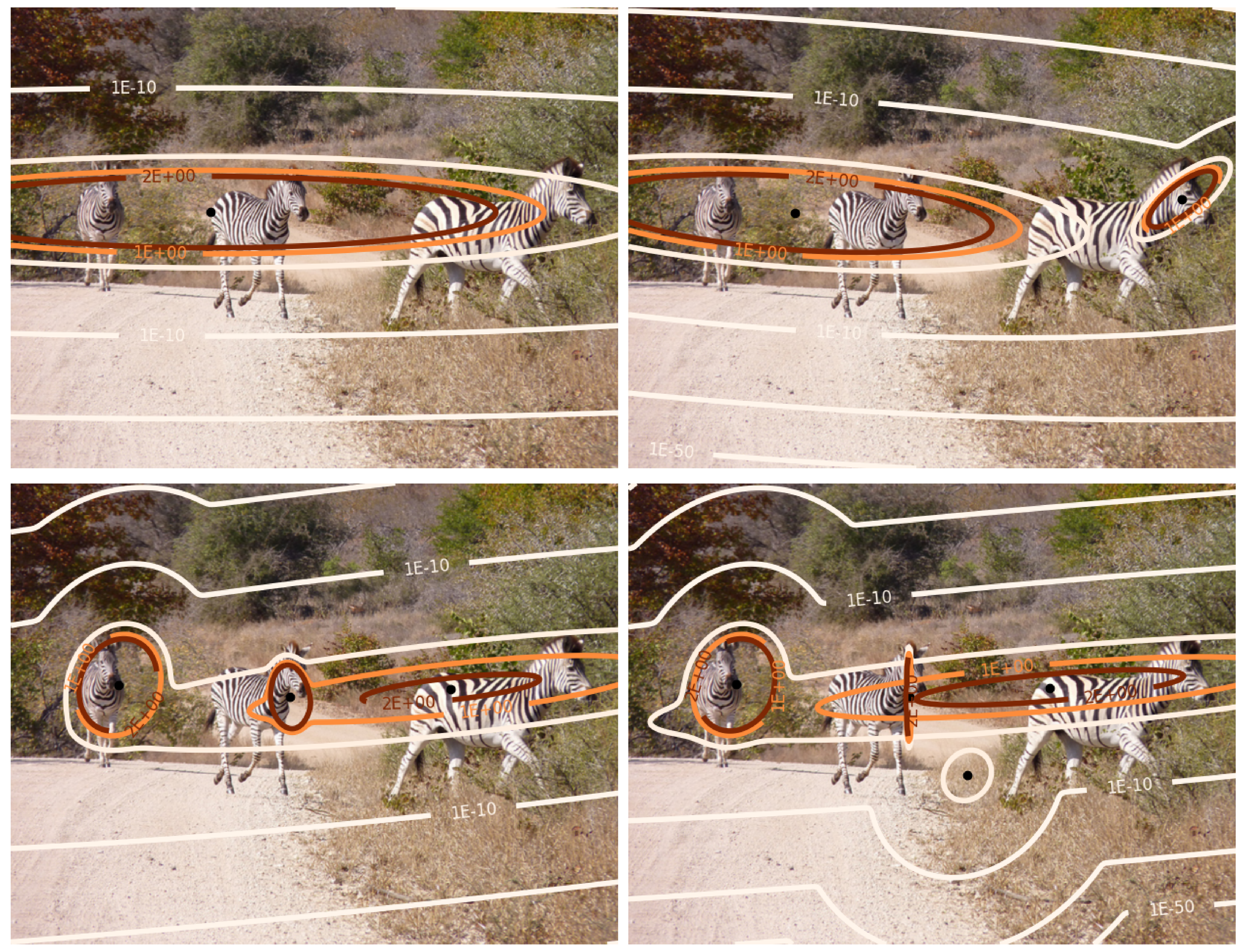
Function $\mathtt{WeightedEM}(\mathcal{X}, \Theta(K), I)$:
- **for** $i \leftarrow 1$ **to** I **do**
  - **for** $\ell \leftarrow 1$ **to** L **do**
    - **for** $k \leftarrow 1$ **to** K **do**
      - $\gamma_{\ell k} \leftarrow \frac{\pi_k \mathcal{N}(x_\ell|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_\ell|\mu_j, \Sigma_j)}$ `// Evaluate the responsibilities`
    - **end**
  - **end**
  - **for** $k \leftarrow 1$ **to** K **do**
    - $\pi_k \leftarrow \sum_{\ell=1}^{L} w_l \gamma_{\ell k}$ `// Re-estimate the mixing coefficients`
    - $\mu_k \leftarrow \frac{1}{\pi_k}\sum_{\ell=1}^{L} w_\ell \gamma_{\ell k} x_\ell$, $\Sigma_k \leftarrow \frac{1}{\pi_k}\sum_{\ell=1}^{L} w_\ell \gamma_{\ell k}(x_\ell - \mu_k)(x_\ell - \mu_k)^\top$ `// Re-estimate the means and covariances`
  - **end**
- **end**
- **return** $\Theta = \{(\pi_k, \mu_k, \Sigma_k)\}_{k=1}^{K}$

Estimating the number of components

Parameters: Centers of grid regions and weights $\mathcal{X} = \{(x_\ell, w_\ell)\}_{\ell=1}^{L}$, initialization $\Theta(K) = \{(\pi_k, \mu_k, \Sigma_k)\}_{k=1}^{K}$, iterations I.

Function $\mathtt{ModelSelection}(\mathcal{X}, \{\Theta(k)\}_{k=1}^{k_{\max}}, I, \lambda)$:
- **for** $k \leftarrow 1$ **to** $k_{\max}$ **do**
  - $\hat{\Theta}_k \leftarrow \mathtt{WeightedEM}(\mathcal{X}, \Theta(k), I)$ `// Obtain parameters using WeightedEM`
  - $\log p(\mathcal{X}|\hat{\Theta}_k) \leftarrow \sum_{\ell=1}^{L} w_\ell \log\left\{\sum_{k=1}^{K} \hat{\pi}_k \mathcal{N}(x_\ell|\hat{\mu}_k, \hat{\Sigma}_k)\right\}$, $\mathcal{C}(\hat{\Theta}_k, k) \leftarrow -2\log p(\mathcal{X}|\hat{\Theta}_k) + \lambda k$ `// Evaluate criterion`
- **end**
- $k^\star = \mathrm{argmin}_k\{\mathcal{C}(\hat{\Theta}_k, k)\}$ `// Choose the optimum number of components`
- **return** $k^\star, \hat{\Theta}_{k^\star}$

How many zebras facing in the left direction?



Attention model

Each attention density is a **K-component mixture of Gaussians**.

- At training time, we pick the number of components *randomly* from a uniform distribution, up to a predefined maximum.
- At test time, we select the *optimum* $K^\star$ from a set of possible choices, using a model selection criterion.

Experiments: Visual Question Answering (VQA)

- Unimodal continuous attention faces difficulties in complex scenes with **multiple regions of interest** far from each other. Multimodal attention densities tend to perform better.
- For a **single complex-shaped interest region**, discrete attention may be too scattered and unimodal attention too focused. Multimodal continuous attention is a good compromise.

What does she have her cups on? bowl — What does she have her cups on? sink — What does she have her cups on? toilet

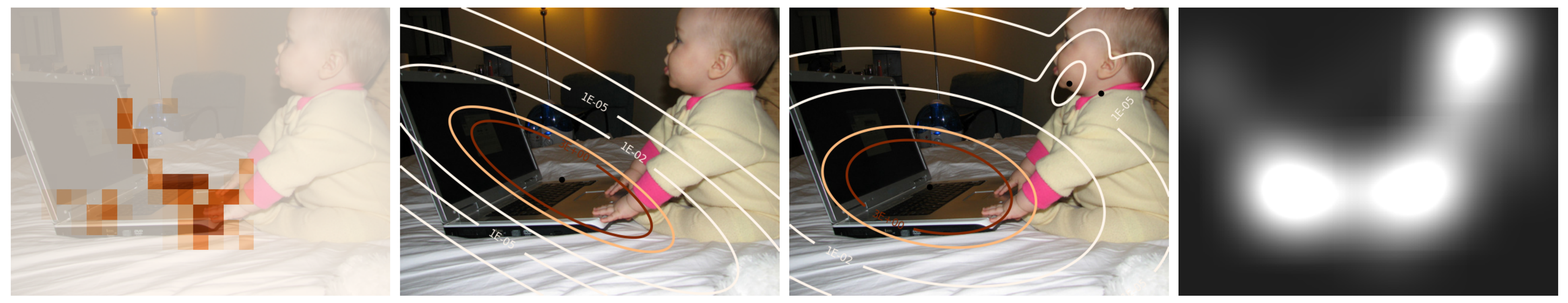
Human attention

- The attention distributions obtained with multimodal continuous attention are **more similar to human attention** than the ones obtained with discrete or unimodal attention.

Attention	JS divergence ↓
Discrete softmax	0.64
Unimodal continuous	0.59
Multimodal continuous	**0.54**

- Humans **sequentially look for regions** in the image, until they found the information they need. Our model replicates this process by identifying multiple regions of interest.

Is the baby using the computer?



Conclusions

- New continuous attention mechanisms that produce multimodal densities with tractable and efficient forward and gradient backpropagation steps.
- Weighted version of the EM algorithm to obtain a selection of relevant regions. Penalized likelihood method to select the number of components in the mixture.
- Experiments on VQA mimic human attention and present increased interpretability.
- **Future work:** Mixtures of sparse family distributions and other vision tasks.

Open-source code:
`https://github.com/deep-spin/vqa-multimodal-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*.

Martins, A., Farinhas, A., Treviso, M., Niculae, V., Aguiar, P., and Figueiredo, M. (2020). Sparse and Continuous Attention Mechanisms. In *Advances in Neural Information Processing Systems*, volume 33, pages 20989–21001.

Acknowledgments: ERC StG DeepSPIN 758969 "Deep Structured Prediction in NLP", P2020 program MAIA (contract 045909), LARSyS-FCT Pluri-anual funding 2020-2023, FCT (contract UIDB/50008/2020).