Multimodal Continuous Visual Attention Mechanisms

Anonymous ICCV submission

Paper ID ****

Abstract

*Visual attention mechanisms are a key component of neural network models for computer vision. By focusing on a discrete set of objects or image regions, these mechanisms identify the most relevant features and use them to build more powerful representations. Recently, continuous-domain alternatives to discrete attention models have been proposed, which exploit the continuity of images. These approaches model attention as simple unimodal densities (*e.g. *a Gaussian), making them less suitable to deal with images whose region of interest has a complex shape or is composed of multiple non-contiguous patches. In this paper, we introduce a new continuous attention mechanism that produces multimodal densities, in the form of mixtures of Gaussians. We use the EM algorithm to obtain a clustering of relevant regions in the image, and a description length penalty to select the number of components in the mixture. Our densities decompose as a linear combination of unimodal attention mechanisms, enabling closed-form Jacobians for the backpropagation step. Experiments on visual question answering in the VQA-v2 dataset show competitive accuracies and a selection of regions that mimics human attention more closely in VQA-HAT, substantiating the impact of the structure induced by our visual prior. We present several examples that suggest how multimodal attention maps are naturally more interpretable than their unimodal counterparts, showing the ability of our model to automatically segregate objects from ground in complex scenes.*

1. Introduction

Visual attention mechanisms are an important component of deep learning models [26, 1, 22, 28]. They appear as a way to mimic the human visual system, which selectively attends to the most relevant parts of visual stimuli, enabling processing large amounts of information in parallel [19].

A neural network with an attention mechanism automatically learns the relevance of any element of the input by generating a set of weights and taking them into account while performing the proposed task. In addition to boosting the performance of a model, attention mechanisms can provide insights into the model's decision process, being suitable for interpretability purposes [25, 5]. In particular, the visualization of attention weights can help us analyze the outputs of a neural network and possibly understand some unpredictable outcomes [8].

Most models for visual attention operate over discrete domains, where images are split into a finite set of regions or pixels [26, 1, 22, 28]. However, this sometimes leads to lack of focus, where the attention distribution over the image becomes too scattered. Discrete attention mechanisms disregard the fact that images are inherently "continuous" objects. Recently, **continuous attention mechanisms** have been proposed [14], which are able to attend over continuous domains and to select compact regions of interest in the image, such as ellipses. Nevertheless, this approach (which we review in §2) is limited in which it models attention with a simple unimodal density, making it less suitable to deal with images whose region of interest has a complex shape or is composed of multiple non-contiguous patches.

In this paper, we address the limitation above by introducing **multimodal** continuous attention mechanisms, in the form of mixtures of unimodal distributions (§3). These mechanisms are able to generate more flexible attention maps while enjoying the best properties of their unimodal counterparts. In particular, we inject *prior knowledge as structural bias*, restricting the attention density to a mixture of Gaussians. We use the Expectation-Maximization (EM) algorithm to obtain a clustering of relevant regions in the image (§4), and we apply a description length penalty to select the number of components in the mixture (§5). Crucially, our densities decompose as a linear combination of unimodal attention mechanisms, enabling tractable and efficient forward and gradient backpropagation steps.

Our experiments in visual question answering show competitive accuracy results in the VQA-v2 dataset [10] (§6). More compelling is the fact that the proposed models lead to more interpretable decisions, being able, for example, to attend to multiple objects without becoming overly unfocused, as illustrated by the example in Figure 1. To obtain a quantitative measure of how well artificial

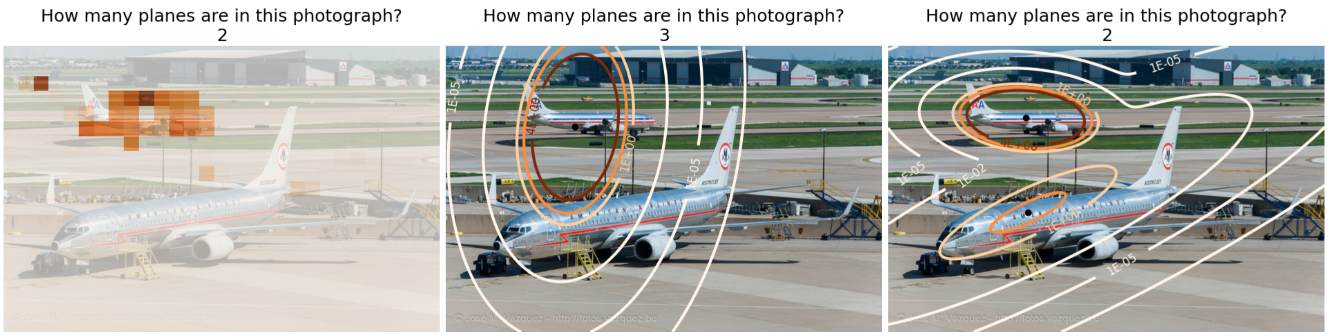
How many planes are in this photograph?
2
How many planes are in this photograph?
3
1E-05
4E+00
1E+00
1E-05
© José M. Vázquez - http://fotos.vazquez.be
How many planes are in this photograph?
2
1E-05
1E+00
4E+00
1E-05
1E-02
1E+00
1E-05
© José M. Vázquez - http://fotos.vazquez.be

Figure 1. Examples of attention maps for VQA. Left: discrete attention. Middle: Unimodal continuous attention. Right: Multimodal continuous attention (ours). For continuous attention models, we identify the means of the Gaussians with black dots.

models represent human attention, we use the VQA-HAT dataset [5]. We conclude that the attention maps provided by the proposed models lead to an overall higher similarity than the ones obtained with discrete or unimodal continuous attention, confirming that the structure induced by our visual prior is appropriate.[1]

2. Continuous attention

2.1. Discrete attention

Attention mechanisms are typically discrete [2, 26]. In vision applications, the starting point is an input image from which L feature vectors in $\mathbb{R}^D$ are extracted (e.g., grid-level or object-level representations), leading to a feature matrix $\boldsymbol{V} \in \mathbb{R}^{D \times L}$. Given some conditioning context (for example a question in natural language), a *score vector* $\boldsymbol{f} = [f_1, \ldots, f_L]^\top \in \mathbb{R}^L$ is computed, where high scores correspond to more relevant parts of the input. These scores are converted into a probability vector $\boldsymbol{p} \in \triangle^L$ (the *attention weights*), where $\triangle^L := \{\boldsymbol{p} \in \mathbb{R}^L \mid \boldsymbol{1}^\top \boldsymbol{p} = 1, \boldsymbol{p} \geq \boldsymbol{0}\}$ is the L-dimensional probability simplex, typically via a softmax transformation, $\boldsymbol{p} = \mathrm{softmax}(\boldsymbol{f})$. Finally, the probability vector is used to compute a weighted average of the input (known as the *context vector*), $\boldsymbol{c} = \boldsymbol{V}\boldsymbol{p} \in \mathbb{R}^D$, that is used to produce the network's decision (for example, an answer to the question).

While discrete attention mechanisms are very flexible, since they allow arbitrary probability mass functions over the input features, this flexibility can be harmful, resulting sometimes in attention maps that are too scattered and lack focus – this may affect prediction accuracy and result in poor interpretability.

2.2. Continuous attention

To avoid the shortcoming above, Martins *et al.* [14] introduced *continuous* attention mechanisms, where images are represented as smooth functions in 2D, instead of being split into regions in a grid.

Feature function. In this framework, instead of the feature matrix $\boldsymbol{V} \in \mathbb{R}^{D \times L}$ above, the image is represented as a continuous *feature function* $\boldsymbol{V} : \mathbb{R}^2 \to \mathbb{R}^D$, where each point in the $\mathbb{R}^2$ plane is assigned a vector representation. This function is linearly parametrized as

$$\boldsymbol{V}_{\boldsymbol{B}}(\boldsymbol{x}) = \boldsymbol{B}\boldsymbol{\psi}(\boldsymbol{x}), \tag{1}$$

where $\boldsymbol{x} = [u, v]^\top$ are coordinates in the image, $\boldsymbol{\psi} : \mathbb{R}^2 \to \mathbb{R}^N$ are N bivariate Gaussian radial basis functions (RBFs) with different means and covariance parameters, and $\boldsymbol{B} \in \mathbb{R}^{D \times N}$ are parameters fit with ridge regression (see [14, §3.1] for details). If $N \ll L$ (fewer basis functions than regions), the continuous representation of the image is more compact than the discrete feature matrix.

Score function and attention density. Likewise, the score vector $\boldsymbol{f} = [f_1, \ldots, f_L]^\top$ above is replaced by a quadratic *score function* $f : \mathbb{R}^2 \to \mathbb{R}$, defined as

$$f(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}), \tag{2}$$

where $\boldsymbol{\mu} \in \mathbb{R}^2$ is a location parameter and $\boldsymbol{\Sigma} \succ \boldsymbol{0}$ is a positive definite matrix in $\mathbb{R}^{2 \times 2}$. This way, relevance is directed to a single location in the image (specified by $\boldsymbol{\mu}$) and it has an elliptical shape, determined by $\boldsymbol{\Sigma}$. The score function is mapped to a probability density $p : \mathbb{R}^2 \to \mathbb{R}_+$ via a regularized prediction mapping [4]. With an entropy regularizer, this results in a Gibbs distribution $p(\boldsymbol{x}) \propto \exp(f(\boldsymbol{x}))$, which for quadratic scores leads to a Gaussian density, $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Evaluation and gradient backpropagation. In continuous attention mechanisms, the output weighted average (context vector) is written as the expectation of the feature function with respect to the probability density,

$$\boldsymbol{c} = \mathbb{E}_p[\boldsymbol{V}_{\boldsymbol{B}}(\boldsymbol{x})] = \boldsymbol{B} \int_{\mathbb{R}^2} p(\boldsymbol{x})\boldsymbol{\psi}(\boldsymbol{x}) \in \mathbb{R}^D, \tag{3}$$

[1] All the code will be released upon acceptance.

where we used (1). If $\boldsymbol{\psi}(\boldsymbol{x})$ are Gaussian RBFs and $p(\boldsymbol{x})$ is a Gaussian, expression (3) becomes the integral of a product of Gaussians, which has a closed form. The backpropagation step can be done either with automatic differentiation or by using a covariance expression to compute the Jacobians $\partial \boldsymbol{c}/\partial \boldsymbol{\mu}$ and $\partial \boldsymbol{c}/\partial \boldsymbol{\Sigma}$ [14, §3.2].

3. Multimodal continuous attention

In this paper, we extend the continuous attention framework described in §2.2 to **multimodal distributions**. This is done by letting the attention density be a mixture of unimodal distributions $p_k : \mathbb{R}^2 \to \mathbb{R}_+$, for $k \in \{1, \dots, K\}$:

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k p_k(\boldsymbol{x}), \tag{4}$$

where $\boldsymbol{\pi} = [\pi_1, \dots, \pi_K]^\top \in \Delta^K$ are mixing coefficients, defining the weight of each component of the mixture. We let each $p_k(\boldsymbol{x})$ be a Gaussian distribution, so that $p(\boldsymbol{x})$ becomes a mixture of Gaussians; we discuss below possible methods for obtaining the mixing coefficients $\boldsymbol{\pi}$. By doing this, we increase the expressive power of continuous attention mechanisms: by using a sufficient number of Gaussians and adjusting the parameters of the mixture, almost any continuous density can be approximated to arbitrary accuracy [3, §2.3.9].

Forward step. Using (3) and invoking the linearity of expectations, we can compute the output of the multimodal attention mechanism as

$$\boldsymbol{c} = \mathbb{E}_p[\boldsymbol{B}\boldsymbol{\psi}(\boldsymbol{x})] = \sum_{k=1}^{K} \pi_k \underbrace{\mathbb{E}_{p_k}[\boldsymbol{B}\boldsymbol{\psi}(\boldsymbol{x})]}_{\boldsymbol{c}_k} = \sum_{k=1}^{K} \pi_k \boldsymbol{c}_k, \tag{5}$$

where each $\boldsymbol{c}_k$ is the context representation after applying each individual (unimodal) attention mechanism; *i.e.*, **$\boldsymbol{c}$ is a mixture of the context representations for each component**.

Backpropagation step. The backpropagation step for the multimodal case is also simple, since it decomposes into a linear combination of unimodal attention mechanisms, each of which has a simple/closed-form Jacobian.

Relation to multi-head attention. Our multimodal attention has some resemblances with multi-head attention mechanisms [24], if we regard each component of the mixture as if it were a different attention head. Note, however, that our construction differs from multi-head attention, where the projection matrices learned as model parameters are head-specific. On the contrary, we assume that $\boldsymbol{B}$ in (5) is fixed, *i.e.*, it does not depend on k. This avoids head-specific computations and enables a probabilistic interpretation of the resulting density as a mixture of densities.

How can we estimate the parameters of the attention density? To choose the mixing coefficients $\boldsymbol{\pi}$, along with the means $(\boldsymbol{\mu}_k)_{k=1}^K$ and covariance matrices $(\boldsymbol{\Sigma}_k)_{k=1}^K$ of each component of the mixture, we start from a given set of observed image locations along with their importance weights $\{(\boldsymbol{x}_\ell, w_\ell)\}_{\ell=1}^L$. Intuitively, the higher the weight, the more important the contribution of that specific region should be to the network's decision. To parametrize the attention density as a simple unimodal distribution, it is possible to use moment matching. For multimodal distributions, we can think of this problem as that of fitting a mixture model to weighted data. In that context, we have to deal with two different issues: how to estimate the number of components, which we discuss in §5, and how to estimate the parameters defining the mixture model. A popular choice to address the second problem is the EM algorithm, which seeks a maximum likelihood estimate of the mixture parameters and is guaranteed to converge to a local maximum [6, 16]. If p_k is a Gaussian, we can easily adapt the EM algorithm to deal with weighted data (*e.g.*, discrete attention weights and corresponding grid locations), so that we can estimate the full set of parameters of a mixture of Gaussians – defining a multimodal attention density, $p(\boldsymbol{x})$. This is described in detail in the next section.

4. The EM algorithm for mixtures of Gaussians

The EM algorithm is the standard method to estimate the parameters defining a mixture model. It starts with an initial estimate for the parameters of the mixture and iteratively updates them until convergence or up to a predefined number of iterations (see [16] for a detailed exposition). In this paper, we assume that each component of the mixture is a Gaussian, *i.e.*, the multimodal attention density $p(\boldsymbol{x})$ takes the form of a mixture of Gaussians.

4.1. EM with weighted data

Let $\mathcal{X} = \{(\boldsymbol{x}_1, w_1), \dots, (\boldsymbol{x}_L, w_L)\}$ be the observed data along with their weights. In our approach, each $\boldsymbol{x}_\ell \in \mathbb{R}^2$ is the center of a grid region, and $w_\ell \in [0, 1]$ is the corresponding discrete attention weight. Our goal is to maximize the likelihood function

$$\mathcal{L}(\Theta) = \sum_{\ell=1}^{L} w_\ell \log p(\boldsymbol{x}_\ell \mid \Theta), \tag{6}$$

where $\Theta = \{(\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}_{k=1}^K$. We adapt the EM algorithm for mixtures of Gaussians to handle weighted data, by changing the way the parameters are re-estimated at each iteration. The algorithm goes as follows:

1. Initialize the parameters $\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$ and evaluate the initial value of the weighted log-likelihood function:

$$\mathcal{L}(\Theta) = \sum_{\ell=1}^{L} w_\ell \log \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_\ell; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}, \quad (7)$$

   where the log-likelihood of each point is multiplied by the correspondent weight.

2. **E step.** Evaluate the responsibilities using the current parameter values:

$$\gamma_{\ell k} = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_\ell | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_\ell | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}. \quad (8)$$

3. **M step.** Re-estimate the parameters using the current responsibilities:

$$\pi_k^{\text{new}} = \sum_{\ell=1}^{L} w_\ell \gamma_{\ell k}, \quad (9)$$

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{\pi_k^{\text{new}}} \sum_{\ell=1}^{L} w_\ell \gamma_{\ell k} \boldsymbol{x}_\ell, \quad (10)$$

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{\pi_k^{\text{new}}} \sum_{\ell=1}^{L} w_\ell \gamma_{\ell k} (\boldsymbol{x}_\ell - \boldsymbol{\mu}_k^{\text{new}})(\boldsymbol{x}_\ell - \boldsymbol{\mu}_k^{\text{new}})^\top. \quad (11)$$

4. Re-evaluate the weighted log-likelihood (7) using the current parameter values and check for convergence of either the parameters or the log likelihood. Return to step 2 if the convergence criterion is not satisfied.

If the weight associated with each observation is the same, *i.e.*, $w_\ell = 1/L$, we recover the usual expressions for the EM algorithm.

4.2. Initialization

The EM algorithm requires an initial choice for the set of parameters $\Theta = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$. This is a relevant issue because EM is not guaranteed to converge to a global maximizer of the log-likelihood function, but rather a local one, meaning that the final estimate depends on the initialization. An effective strategy is to run EM multiple times with different random initializations and choose the final estimate that leads to the highest likelihood [16].

5. Estimating the number of components

The maximum likelihood criterion cannot be used to estimate the number of components K in a mixture density: If $\mathcal{M}_k$ is a class composed by all Gaussian mixtures with K components, it is trivial to show that $\mathcal{M}_K \subseteq \mathcal{M}_{K+1}$ and thus the maximized likelihood is a non decreasing function of K, useless as a criterion to estimate K [7]. For this reason, several **model selection** methods have been proposed to estimate the number of components of a mixture [17, Chapter 6]. We focus on penalized likelihood methods such as the Bayesian Information Criterion (BIC, [21]) or the Minimum Description Length (MDL, [20]), where the EM algorithm is used to obtain different parameter estimates for a range of values of k, $\{\hat{\Theta}_k,\ k = k_{\min}, \dots, k_{\max}\}$, and the number of components is chosen according to

$$k^\star = \underset{k \in \{k_{\min}, \dots, k_{\max}\}}{\arg\min} \ \mathcal{C}(\hat{\Theta}_k, k), \quad (12)$$

where $\mathcal{C}(\hat{\Theta}_k, k)$ is a model selection criterion. We use a criterion of the form

$$\mathcal{C}(\hat{\Theta}_k, k) = -2 \log p(\mathcal{X} | \hat{\Theta}_k) + \mathcal{P}(k), \quad (13)$$

where $\mathcal{P}(k)$ is an increasing function penalizing higher values of k (*e.g.*, $\mathcal{P}_{\text{BIC}}(k) = k \log n$, where n is the number of data points). For the weighted data scenario presented in §4.1 we cannot use the number of points; thus, we write

$$\mathcal{P}(k) = \lambda\, k, \quad (14)$$

where $\lambda > 0$ is an hyperparameter obtained using *cross-validation*. The resulting model selection criterion,

$$\mathcal{C}(\hat{\Theta}_k, k) = -2 \log p(\mathcal{X} | \hat{\Theta}_k) + \lambda\, k, \quad (15)$$

will be used in §6 to estimate the number of components in a multimodal continuous attention density.

Attention model. Using the results of the previous section, we model each attention density as 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. This way we expose the model to different numbers of components, maintaining the simplicity of the training procedure without added runtime. At test time, we select the optimum $K^\star$ from a set of possible choices, using the model selection criterion (15). See Algorithm 1 for pseudo-code. (Although we consider multiple random initializations along with the model selection criterion, we omit this step in the algorithm, for simplicity.)

Note that our extension from unimodal to multimodal continuous attention does not increase the number of neural network parameters. Therefore, in practice, it is possible to leverage the learned representations from a pretrained model using either discrete or unimodal continuous attention mechanisms (*e.g.* discrete or continuous softmax) and fine-tune it with our multimodal attention densities. This method allows us to model the attention distribution as an expressive density function that could not be properly modeled using a single Gaussian.

Algorithm 1: Multimodal continuous attention with Gaussian RBFs. During training, we pick the number of components randomly and apply WeightedEM, followed by MultimodalAttention. At test time, we apply ModelSelection in between the previous functions to select the number of components.

```
Parameters: Centers of grid regions and their weights $\mathcal{X} = \{(\boldsymbol{x}_\ell, w_\ell)\}_{\ell=1}^{L}$, initialization $\Theta(K) = \{(\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}_{k=1}^{K}$, number
            of iterations $I$, Gaussian RBFs $\boldsymbol{\psi}(\boldsymbol{x}) = [\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)]_{j=1}^{N}$, value function $\boldsymbol{V_B}(\boldsymbol{x}) = \boldsymbol{B}\boldsymbol{\psi}(\boldsymbol{x})$.

Function 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
                $\gamma_{\ell k} \leftarrow \frac{\pi_k \mathcal{N}(\boldsymbol{x}_\ell | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_\ell | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$                // (8)
        for $k \leftarrow 1$ to $K$ do
            $\pi_k \leftarrow \sum_{\ell=1}^{L} w_l \gamma_{\ell k}$                // (9)
            $\boldsymbol{\mu}_k \leftarrow \frac{1}{\pi_k} \sum_{\ell=1}^{L} w_\ell \gamma_{\ell k} \boldsymbol{x}_\ell$,  $\boldsymbol{\Sigma}_k \leftarrow \frac{1}{\pi_k} \sum_{\ell=1}^{L} w_\ell \gamma_{\ell k} (\boldsymbol{x}_\ell - \boldsymbol{\mu}_k)(\boldsymbol{x}_\ell - \boldsymbol{\mu}_k)^\top$                // (10), (11)
    return $\Theta = \{(\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}_{k=1}^{K}$

Function ModelSelection($\mathcal{X}, \{\Theta(k)\}_{k=1}^{k_{\max}}, I, \lambda$):
    for $k \leftarrow 1$ to $k_{\max}$ do
        $\hat{\Theta}_k \leftarrow$ WeightedEM($\mathcal{X}, \Theta(k), I$)
        $\log p(\mathcal{X} | \hat{\Theta}_k) \leftarrow \sum_{\ell=1}^{L} w_\ell \log \left\{ \sum_{k=1}^{K} \hat{\pi}_k \mathcal{N}(\boldsymbol{x}_\ell | \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k) \right\}$,  $\mathcal{C}(\hat{\Theta}_k, k) \leftarrow -2 \log p(\mathcal{X} | \hat{\Theta}_k) + \lambda\, k$                // (7), (15)
    $k^\star = \arg\min_k \{\mathcal{C}(\hat{\Theta}_k, k)\}$                // (12)
    return $k^\star, \hat{\Theta}_{k^\star}$

Function MultimodalAttention($\boldsymbol{V_B}, \Theta = \{(\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}_{k=1}^{K}$):
    for $k \leftarrow 1$ to $K$ do
        $r_{kj} \leftarrow \mathbb{E}_p[\psi_j(\boldsymbol{x})] = \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_k + \boldsymbol{\Sigma}_j)$,    $\forall j \in [N]$                // [14, §3]
        $\boldsymbol{c}_k \leftarrow \boldsymbol{B}\boldsymbol{r}_k$                // (3)

    $\boldsymbol{c} \leftarrow \sum_{k=1}^{K} \pi_k \boldsymbol{c}_k$                // (5)
    return $\boldsymbol{c}$ (context vector)
```

6. Experiments

6.1. Visual Question Answering

Dataset and metrics. We use the VQA-v2 dataset [10] with the standard splits (443K, 214K, and 453K question-image pairs for train/dev/test, the latter subdivided into test-dev, test-standard, test-challenge and test-reserve). We report results in terms of accuracy in the test-dev and test-standard splits. All the models we experiment with are trained only on the train split, without data augmentation.

Architecture. We adapt the implementation of the encoder-decoder version of the Modular Co-Attention Network (MCAN, [27]),[2] and represent the image input with grid features generated by Jiang *et al*. [11], using a ResNet pretrained on Visual Genome [13] that outputs a feature map of size $L \times 2048$, where L is the number of features ($\overline{L} = 506$ and $L_{\max} = 608$). To represent the question words we use 300-dimensional GloVe word embeddings [18], yielding a feature matrix representation.

Attention models. We consider three different attention models: discrete attention, unimodal continuous attention, and multimodal continuous attention (ours). The discrete attention model attends over a grid and uses the softmax transformation to map scores into probabilities. For the continuous attention models, we normalize the image size into the unit square $[0, 1]^2$. Then, we transform the image into a continuous function $\boldsymbol{V_B} : \mathbb{R}^2 \to \mathbb{R}^D$ using ridge regression, and fit a Gaussian (unimodal continuous attention) or a mixture of Gaussians (multimodal continuous attention) as the attention density. In the first case, we obtain $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ with moment matching; in the second case, we use the method described in §3 – we set the maximum number of components to $k_{\max} = 4$ and, during training, we pick the number of components randomly from a uniform distribution; at test time, we use 3 random initializations for each k and apply the model selection criterion (15) to choose the optimum number of components. In both cases, we use $N = 100 \ll 506$ Gaussian RBFs $\mathcal{N}(\boldsymbol{x}; \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$, with $\tilde{\boldsymbol{\mu}}$ linearly spaced in $[0, 1]^2$ and $\tilde{\boldsymbol{\Sigma}} = 0.001 \cdot \mathbf{I}$. The number of neural network parameters is the same in all attention models, both discrete and continuous.

[2] https://github.com/MILVLG/mcan-vqa

ATTENTION	Test-Dev				Test-Standard			
	Yes/No	Number	Other	Overall	Yes/No	Number	Other	Overall
Discrete softmax	86.76	52.90	60.78	70.59	86.91	53.22	61.10	70.94
Unimodal continuous	86.57	53.69	60.38	70.41	86.73	53.55	60.75	70.73
Multimodal continuous	86.62	53.23	60.46	70.42	86.88	53.31	60.79	70.79

Table 1. Accuracies of different models on the *test-dev* and *test-standard* splits of VQA-v2.

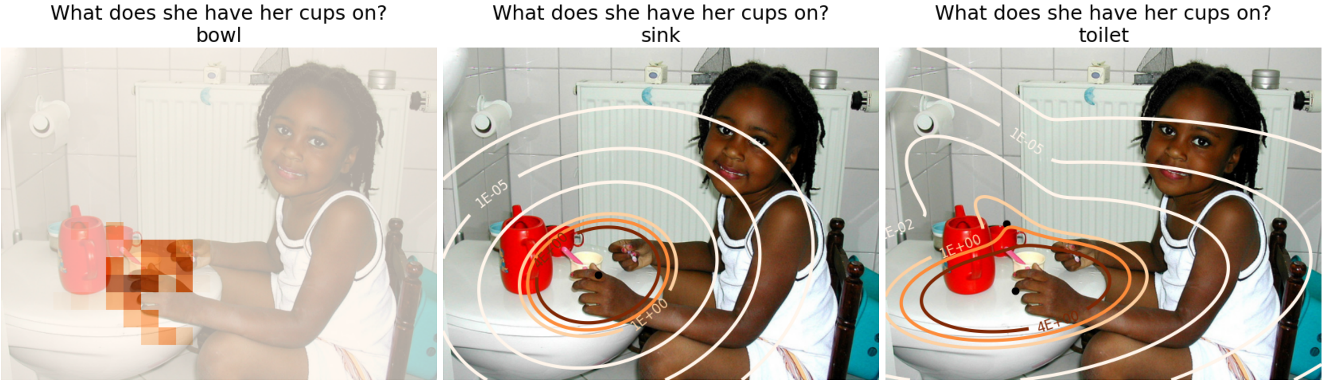
What does she have her cups on?
bowl
What does she have her cups on?
sink
1E-05
1E+00
What does she have her cups on?
toilet
1E-05
1E-02
1E+00

Figure 2. Examples of attention maps in the VQA-v2 dataset. Left: discrete softmax attention. Middle: unimodal continuous attention. Right: Multimodal continuous attention (ours).

Settings. All models are trained for a maximum of 15 epochs using the Adam optimizer [12] with a learning rate of $\min(2.5t \cdot 10^{-5}, 5 \cdot 10^{-4})$, where t is the epoch number. After 10 epochs, the learning rate is multiplied by 0.2 every 2 epochs. For continuous attention models, we use a penalty of 0.01 in the ridge regression step. For multimodal continuous attention, we perform 5 and 10 iterations of the EM algorithm during training and testing, respectively. We set $\lambda = 5$, which leads to the selection of $K^\star = 1$ in 80.8% of the examples, $K^\star = 2$ in 12.4%, $K^\star = 3$ in 4.4%, and $K^\star = 4$ in 2.4%.

Results. The results in Table 1 show similar accuracies for all attention models with a slight overall advantage for the discrete attention model. Note however that the multimodal and unimodal continuous attentions use much fewer basis functions than image regions ($N \ll \overline{L} = 506$).

Attention visualization. We identify two main strengths of multimodal continuous attention when compared to discrete or unimodal continuous attention. First, previously proposed continuous attention models face difficulties in complex scenes (*e.g.*, if there are multiple regions of interest that are far from each other), due to being limited to a single mode. In those cases, unimodal attention ellipses become wide and less interpretable, assigning a high probability mass to a region that is not the most relevant one; or they focus on a single region and completely disregard the others. As suggested by Figure 1, multimodal attention densities tend to perform considerably better in such situations, by increasing the number of components in the attention mixture and adequately setting the mean and covariance matrix of each Gaussian component.

Another interesting case is illustrated by the example in Figure 2. Although there is a single region of interest in the image, its complex shape confuses the non-structured and scattered discrete attention model. Besides, as a result of being overly focused, a simple Gaussian distribution is not enough to fully encompass all the relevant objects in the scene. By increasing the number of components in the mixture, continuous attention models become capable of more accurately segregate objects from ground, encompassing their actual shapes.

Comparing multimodal attention maps. Figures 3 and 4 illustrate how the model selection criterion (15) is used to estimate the number of components in the attention mixture. In the first example, when asked how many zebras are facing left, our attention model chooses $K^\star = 3$, aligning the ellipses properly. It is interesting to see that by decreasing or increasing the number of components in the mixture, the attention map becomes less interpretable (see, for instance, that for $K = 2$ there is a distribution *peak* between two zebras, and for $K = 4$, we can clearly identify one extra component with its mean located on the ground). A similar analysis can be done for the example in Figure 4, where our model opts to use only two components.

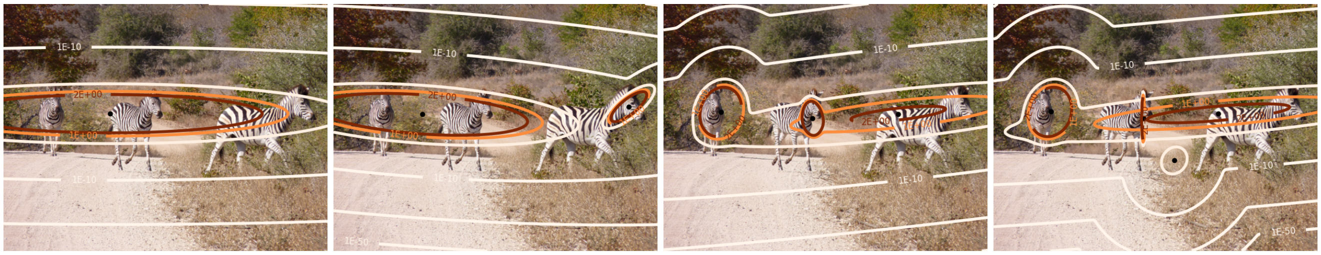

Figure 3. Attention maps generated when answering the question: **How many zebras are facing in the left direction?** Our model selection criterion chooses $K^\star = 3$.

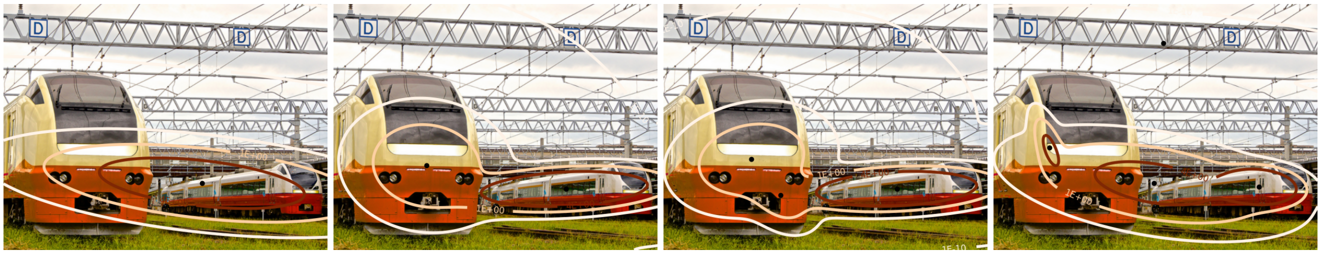

Figure 4. Attention maps generated when answering the question: **How many trains?** Our model selection criterion chooses $K^\star = 2$.

6.2. Human attention

To quantitatively evaluate how interpretable different attention models are, we compare the attention distributions obtained using different models with human attention. For this purpose, we use the VQA-HAT dataset [5] that contains human attention maps obtained through a deblurring procedure: human annotators were presented with a blurred image and a question about it, and were asked to progressively sharpen the regions of the image that help them answer the question correctly. In order to compare the attention distributions with the human attention, we measure the Jensen-Shannon (JS) divergence between them. This metric was proposed in [15] and addresses the limitations of order-based metrics like the Spearman's rank correlation used in [5], by taking into account the magnitude of the attention distributions at a given spacial location.[3]

The results reported in Table 2 show that the attention distributions obtained with multimodal continuous attention mechanisms are more similar to human attention than the ones obtained with discrete or unimodal continuous attention. These results suggest that our method is able to generate more human-interpretable attention maps.

Attention visualization. Figures 5, 6 and 7 illustrate how the attention maps generated by different attention models relate to human attention. To answer the questions, humans sequentially look for regions in the image, until they found all the information they need. Our multimodal attention models replicate this process by identifying multiple regions of interest.

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

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

7. Related work

EM algorithm for weighted data. Gebru *et al.* [9] proposed to incorporate the weights into the model by "*observing* $\boldsymbol{x}$ w *times*" and changing the log-likelihood function accordingly: they raise $\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$ to the power w and notice that $\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})^w \propto \mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}/w)$, deriving a new mixture model where w plays the role of precision. However, they focus on the case where the weights are treated as random variables, which is different from ours.

Sparse continuous attention. Martins *et al.* [14] introduced continuous attention mechanisms for both 1D and 2D applications. In their work, they consider other densities besides Gaussian distributions, in particular densities with sparse support, such as truncated paraboloids, establishing a parallel with Tsallis-regularized prediction maps [23, 4]. In our work, we restrict to Gaussian densities, which are simpler and allow closed-form forward and backpropagation steps. Furthermore, mixtures of Gaussians (the multimodal extension considered in our paper) are more amenable for

[3] The output of all the attention models is strictly dense, assigning a probability mass to every image feature. Since less relevant features are all assigned a very small positive attention value, order-based metrics are less suitable for measuring the similarity between attention distributions.

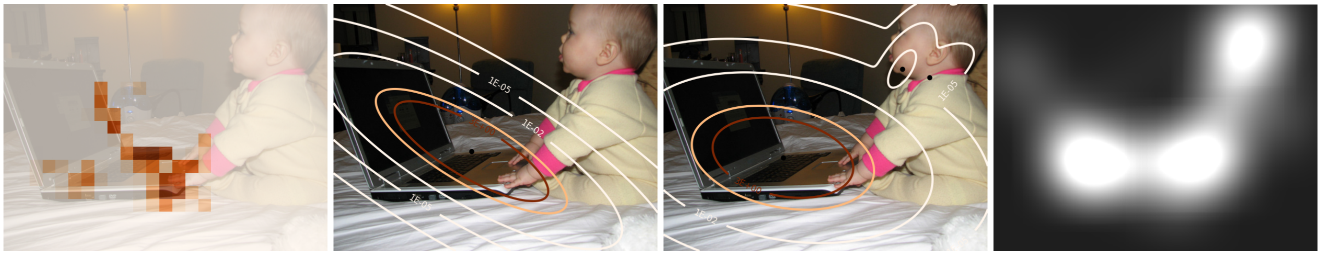

Figure 5. Attention maps generated when answering the question: **Is the baby using the computer?** Discrete attention (JS div. $= 0.66$), unimodal continuous attention (JS div. $= 0.66$), multimodal continuous attention (JS div. $= 0.60$), human attention.

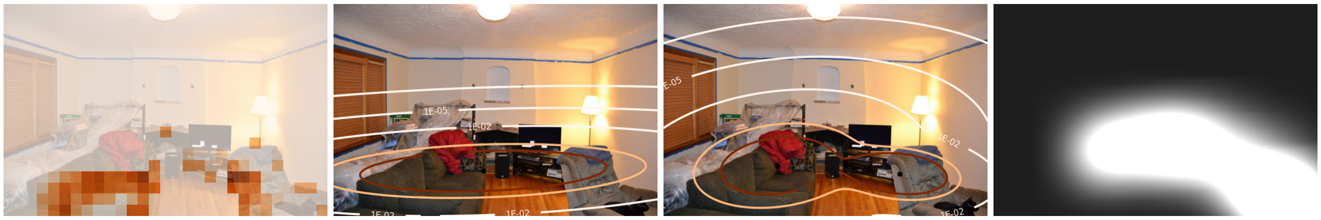

Figure 6. Attention maps generated when answering the question: **What type of furniture is the cat sitting on?** Discrete attention (JS div. $= 0.66$), unimodal continuous attention (JS div. $= 0.56$), multimodal continuous attention (JS div.$= 0.51$), human attention.

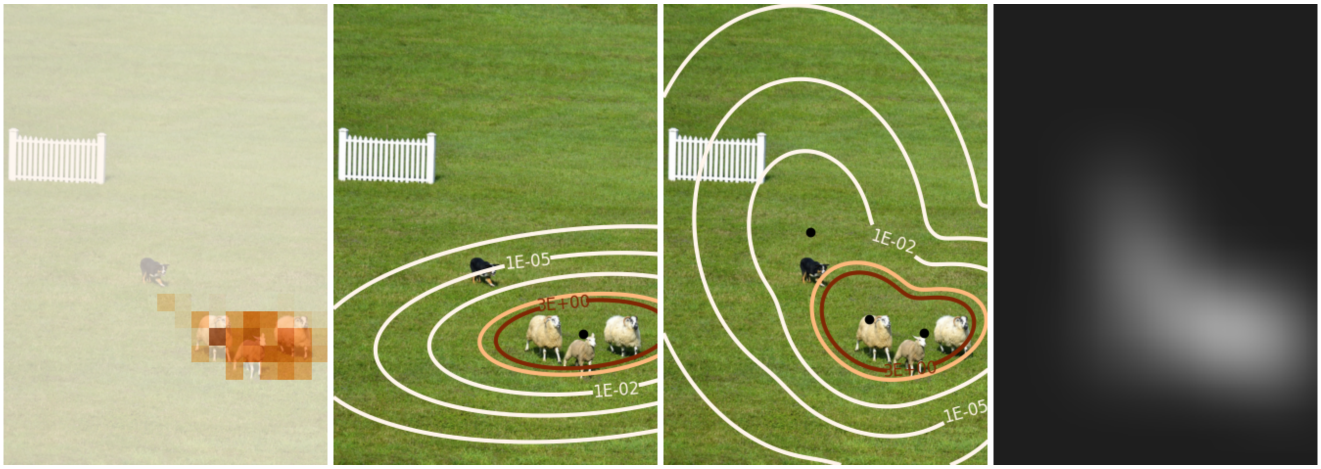

Figure 7. Attention maps generated when answering the question: **How many sheep are there?** Discrete attention (JS div. $= 0.68$), unimodal continuous attention (JS div. $= 0.71$), multimodal continuous attention (JS div. $= 0.68$), human attention.

use in soft clustering with the EM algorithm, since they have tractable and efficient expectation and maximization steps.

8. Conclusions and future work

We propose new continuous attention mechanisms that produce multimodal densities in the form of mixtures of unimodal distributions (*e.g.* a Gaussian) and show that they decompose as a linear combination of unimodal attention mechanisms, enabling tractable and efficient forward and gradient backpropagation steps (§3). We use a weighted version of the Expectation-Maximization (EM) algorithm to obtain a selection of relevant regions in the image (§4), and a penalized likelihood method to select the number of components in the mixture (§5). Experiments on visual question answering show that the selected regions mimic human attention more closely than previously proposed models, leading to more interpretable attention maps (§6).

There are several avenues for future research. We used mixture of Gaussians only. However, it seems interesting to consider mixtures of sparse family distributions (*e.g.* mixtures of truncated paraboloids) in which different components may have disjoint supports. Another direction consists in exploring our method in other vision tasks that require learning from images and video, which could equally benefit from focusing on multiple objects simultaneously.

References

[1] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proc. of ICLR*, 2015. 2

[3] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. 3

[4] Mathieu Blondel, André FT Martins, and Vlad Niculae. Learning with fenchel-young losses. *Journal of Machine Learning Research*, 21(35):1–69, 2020. 2, 7

[5] 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 *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016. 1, 2, 7

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39:1–38, 1977. 3

[7] Mario A. T. Figueiredo and Anil K. Jain. Unsupervised learning of finite mixture models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(3):381–396, Mar. 2002. 4

[8] Andrea Galassi, Marco Lippi, and Paolo Torroni. Attention in natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, page 1–18, 2020. 1

[9] Israel D. Gebru, Xavier Alameda-Pineda, F. Forbes, and R. Horaud. Em algorithms for weighted-data clustering with application to audio-visual scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38:2402–2415, 2016. 7

[10] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 5

[11] Huaizu Jiang, I. Misra, Marcus Rohrbach, E. Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10264–10273, 2020. 5

[12] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 6

[13] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, A. David Shamma, S. Michael Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 2017. 5

[14] André Martins, António Farinhas, Marcos Treviso, Vlad Niculae, Pedro Aguiar, and Mario Figueiredo. Sparse and Continuous Attention Mechanisms. In *Advances in Neural Information Processing Systems*, volume 33, pages 20989–21001, 2020. 1, 2, 3, 5, 7

[15] Pedro Henrique Martins, Vlad Niculae, Zita Marinho, and André Martins. Sparse and structured visual attention, 2020. 7

[16] Geoffrey J. McLachlan and T. Krishnan. *The EM algorithm and extensions*. Wiley New York, 1997. 3, 4

[17] Geoffrey J. McLachlan and David Peel. *Finite mixture models*, volume 299 of *Probability and Statistics – Applied Probability and Statistics Section*. Wiley, New York, 2000. 4

[18] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. 5

[19] Ronald A Rensink. The Dynamic Representation of Scenes. *Visual Cognition*, 7(1-3):17–42, 2000. 1

[20] Jorma Rissanen. *Stochastic Complexity in Statistical Inquiry Theory*. World Scientific Publishing Co., Inc., USA, 1989. 4

[21] Gideon Schwarz. Estimating the Dimension of a Model. *Annals Statist.*, 6:461–464, 1978. 4

[22] Shikhar Sharma, Ryan Kiros, and Ruslan Salakhutdinov. Action recognition using visual attention. In *Neural Information Processing Systems (NIPS) Time Series Workshop*, December 2015. 1

[23] Constantino Tsallis. Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, 52:479–487, 1988. 7

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017. 3

[25] Sarah Wiegreffe and Yuval Pinter. Attention is not not Explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China, 2019. Association for Computational Linguistics. 1

[26] 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 Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France, 07–09 Jul 2015. PMLR. 1, 2

[27] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:6274–6283, 2019. 5

[28] Xiaoning Zhang, Tiantian Wang, Jinqing Qi, Huchuan Lu, and Gang Wang. Progressive attention guided recurrent network for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1