# Learning Visual Question Answering by Bootstrapping Hard Attention

Mateusz Malinowski, Carl Doersch, Adam Santoro, and Peter Battaglia

DeepMind, London, United Kingdom

**Abstract.** Attention mechanisms in biological perception are thought to select subsets of perceptual information for more sophisticated processing which would be prohibitive to perform on all sensory inputs. In computer vision, however, there has been relatively little exploration of *hard* attention, where some information is selectively ignored, in spite of the success of *soft* attention, where information is re-weighted and aggregated, but never filtered out. Here, we introduce a new approach for hard attention and find it achieves very competitive performance on a recently-released visual question answering datasets, equalling and in some cases surpassing similar soft attention architectures while entirely ignoring some features. Even though the hard attention mechanism is thought to be non-differentiable, we found that the feature magnitudes correlate with semantic relevance, and provide a useful signal for our mechanism's attentional selection criterion. Because hard attention selects important features of the input information, it can also be more efficient than analogous soft attention mechanisms. This is especially important for recent approaches that use *non-local pairwise* operations, whereby computational and memory costs are quadratic in the size of the set of features.

**Keywords:** Visual Question Answering, Visual Turing Test, Attention

## 1 Introduction

Visual attention is instrumental to many aspects of complex visual reasoning in humans [1,2]. For example, when asked to identify a dog's owner among a group of people, the human visual system adaptively allocates greater computational resources to processing visual information associated with the dog and potential owners, versus other aspects of the scene. The perceptual effects can be so dramatic that prominent entities may not even rise to the level of awareness when the viewer is attending to other things in the scene [3,4,5]. Yet attention has not been a transformative force in computer vision, possibly because many standard computer vision tasks like detection, segmentation, and classification do not involve the sort of complex reasoning which attention is thought to facilitate.

Answering detailed questions about an image is a type of task which requires more sophisticated patterns of reasoning, and there has been a rapid recent proliferation of computer vision approaches for tackling the *visual question answering* (visual QA) task [6,7]. Successful visual QA architectures must be able
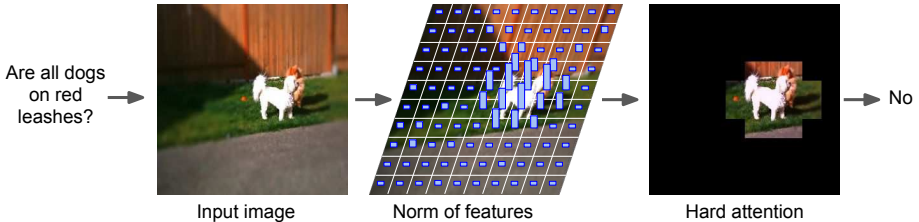
Fig. 1: Given a natural image and a textual question as input, our visual QA architecture outputs an answer. It uses a hard attention mechanism that selects only the important visual features for the task for further processing. We base our architecture on the premise that the norm of the visual features correlates with their relevance, and that those feature vectors with high magnitudes correspond to image regions which contain important semantic content.

to handle many objects and their complex relations while also integrating rich background knowledge, and attention has emerged as a promising strategy for achieving good performance [7,8,9,10,11,12,13,14].

We recognize a broad distinction between types of attention in computer vision and machine learning – *soft* versus *hard* attention. Existing attention models [7,8,9,10] are predominantly based on soft attention, in which all information is adaptively re-weighted before being aggregated. This can improve accuracy by isolating important information and avoiding interference from unimportant information. Learning becomes more data efficient as the complexity of the interactions among different pieces of information reduces; this, loosely speaking, allows for more unambiguous credit assignment.

By contrast, hard attention, in which only a subset of information is selected for further processing, is much less widely used. Like soft attention, it has the potential to improve accuracy and learning efficiency by focusing computation on the important parts of an image. But beyond this, it offers better computational efficiency because it only fully processes the information deemed most relevant. However, there is a key downside of hard attention within a gradient-based learning framework, such as deep learning: because the choice of which information to process is discrete and thus non-differentiable, gradients cannot be backpropagated into the selection mechanism to support gradient-based optimization. There have been various efforts to address this shortcoming in visual attention [15], attention to text [16], and more general machine learning domains [17,18,19], but this is still a very active area of research.

Here we explore a simple approach to hard attention that bootstraps on an interesting phenomenon [20] in the feature representations of convolutional neural networks (CNNs): learned features often carry an easily accessible signal for hard attentional selection. In particular, selecting those feature vectors with the greatest $L2$-norm values proves to be a heuristic that can facilitate hard attention – and provide the performance and efficiency benefits associated with – without

requiring specialized learning procedures (see Figure 1). This attentional signal results indirectly from a standard supervised task loss, and does not require explicit supervision to incentivize norms to be proportional to object presence, salience, or other potentially meaningful measures [20,21].

We rely on a canonical visual QA pipeline [7,9,22,23,24,25] augmented with a hard attention mechanism that uses the $L2$-norms of the feature vectors to select subsets of the information for further processing. The first version, called the Hard Attention Network (HAN), selects a fixed number of feature vectors by choosing those with the top norms. The second version, called the Adaptive Hard Attention Network (AdaHAN), selects a variable number of feature vectors that depends on the input. Our results show that our algorithm can actually outperform comparable soft attention architectures on a challenging visual QA task. This approach also produces interpretable hard attention masks, where the image regions which correspond to the selected features often contain semantically meaningful information, such as coherent objects. We also show strong performance when combined with a form of non-local pairwise model [26,25,27,28]. This algorithm computes features over pairs of input features and thus scale quadratically with number of vectors in the feature map, highlighting the importance of feature selection.

## 2   Related Work

Visual question answering, or more broadly the Visual Turing Test, asks "Can machines understand a visual scene only from answering questions?" [6,23,29,30,31,32]. Creating a good visual QA dataset has proved non-trivial: biases in the early datasets [6,22,23,33] rewarded algorithms for exploiting spurious correlations, rather than tackling the reasoning problem head-on [7,34,35]. Thus, we focus on the recently-introduced VQA-CP [7] and CLEVR [34] datasets, which aim to reduce the dataset biases, providing a more difficult challenge for rich visual reasoning.

One of the core challenges of visual QA is the problem of *grounding language*: that is, associating the meaning of a language term with a specific perceptual input [36]. Many works have tackled this problem [37,38,39,40], enforcing that language terms be grounded in the image. In contrast, our algorithm does not directly use correspondence between modalities to enforce such grounding but instead relies on learning to find a discrete representation that captures the required information from the raw visual input, and question-answer pairs.

The most successful visual QA architectures build multimodal representations with a combined CNN+LSTM architecture [22,33,41], and recently have begun including attention mechanisms inspired by soft and hard attention for image captioning [42]. However, only soft attention is used in the majority of visual QA works [7,8,9,10,11,12,43,44,45,46,47,48,49,50,51,52]. In these architectures, a full-frame CNN representation is used to compute a spatial weighting (attention) over the CNN grid cells. The visual representation is then the weighted-sum of the input tensor across space.

The alternative is to select CNN grid cells in a discrete way, but due to many challenges in training non-differentiable architectures, such hard attention alternatives are severely under-explored. Notable exceptions include [6,13,14,53,54,55], but these run state-of-the-art object detectors or proposals to compute the hard attention maps. We argue that relying on such external tools is fundamentally limited: it requires costly annotations, and cannot easily adapt to new visual concepts that aren't previously labeled. Outside visual QA and captioning, some prior work in vision has explored limited forms of hard attention. One line of work on discriminative patches builds a representation by selecting some patches and ignoring others, which has proved useful for object detection and classification [56,57,58], and especially visualization [59]. However, such methods have recently been largely supplanted by end-to-end feature learning for practical vision problems. In deep learning, spatial transformers [60] are one method for selecting an image regions while ignoring the rest, although these have proved challenging to train in practice. Recent work on compressing neural networks (e.g. [61]) uses magnitudes to remove weights of neural networks. However it prunes permanently based on weight magnitudes, not dynamically based on activation norms, and has no direct connection to hard-attention or visual QA.

Attention has also been studied outside of vision. While the focus on soft attention predominates these works as well, there are a few examples of hard attention mechanisms and other forms of discrete gating [15,16,17,18,19]. In such works the decision of *where to look* is seen as a discrete variable that had been optimized either by reinforce loss or various other approximations (e.g. straight-through). However, due to the high variance of these gradients, learning can be inefficient, and soft attention mechanisms usually perform better.

## 3   Method

Answering questions about images is often formulated in terms of predictive models [24]. These architectures maximize a conditional distribution over answers $a$, given questions $q$ and images $x$:

$$\hat{a} = \arg\max_{a \in \mathcal{A}} p(a|x, q) \tag{1}$$

where $\mathcal{A}$ is a countable set of all possible answers. As is common in question answering [7,9,22,23,24], the question is a sequence of words $q = [q_1, ..., q_n]$, while the output is reduced to a classification problem between a set of common answers (this is limited compared to approaches that generate answers [41], but works better in practice). Our architecture for learning a mapping from image and question, to answer, is shown in Figure 2. We encode the image with a CNN [62] (in our case, a pre-trained ResNet-101 [63], or a small CNN trained from scratch), and encode the question to a fixed-length vector representation with an LSTM [64]. We compute a combined representation by copying the question representation to every spatial location in the CNN, and concatenating it with (or simply adding it to) the visual features, like previous
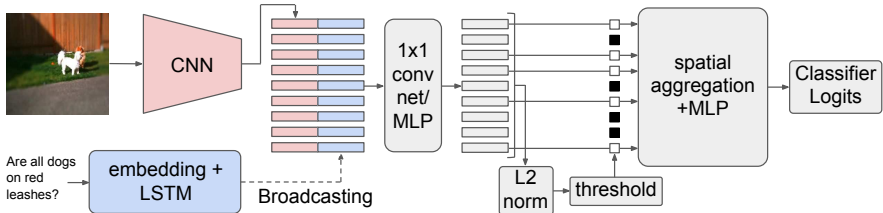
Fig. 2: Our hard attention replaces commonly used soft attention mechanism. Otherwise, we follow the canonical visual QA pipeline [7,9,22,23,24,25]. Questions and images are encoded into their vector representations. Next, the spatial encoding of the visual features is unraveled, and the question embedding is broadcasted and concatenated (or added) accordingly to form a multimodal representation of the inputs. Our attention mechanism selectively chooses a subset of the multimodal vectors that are next aggregated and processed by the answering module.

work [7,9,22,23,24,25]. After a few layers of combined processing, we apply attention over spatial locations, following previous works which often apply soft attention mechanisms [7,9,8,10] at this point in the architecture. Finally, we aggregate features, using either sum-pooling, or relational [25,27] modules. We train the whole network end-to-end with a standard logistic regression loss over answer categories.

### 3.1 Attention Mechanisms

Here, we describe prior work on soft attention, and our approach to hard attention.

**Soft Attention.** In most prior work, soft attention is implemented as a weighted mask over the spatial cells of the CNN representation. Let $\boldsymbol{x} := CNN(x), \boldsymbol{q} := LSTM(q)$ for image $x$ and question $q$. We compute a weight $w_{ij}$ for every $\boldsymbol{x}_{ij}$ (where $i$ and $j$ index spatial locations), using a neural network that takes $\boldsymbol{x}_{ij}$ and $\boldsymbol{q}$ as input. Intuitively, weight $w_{ij}$ measures the "relevance" of the cell to the input question. $w$ is nonnegative and normalized to sum to 1 across the image (generally with softmax). Thus, $w$ is applied to the visual input via $\hat{\boldsymbol{h}}_{ij} := w_{ij}\boldsymbol{x}_{ij}$ to build the multi-modal representation. This approach has some advantages, including conceptual simplicity and differentiability. The disadvantage is that the weights, in practice, are never 0. Irrelevant background can affect the output, no features can be dropped from potential further processing, and credit assignment is still challenging.

**Hard Attention.** Our main contribution is a new mechanism for hard attention. It produces a binary mask over spatial locations, which determines which features are passed on to further processing. We call our method the Hard Attention Network (HAN). The key idea is to use the $L2$-norm of the activations at each spatial location as a proxy for relevance at that location. The correlation

between $L2$-norm and relevance is an emergent property of the trained CNN features, which requires no additional constraints or objectives. [20] recently found something related: in an ImageNet-pretrained representation of an image of a cat and a dog, the largest feature norms appear above the cat and dog face, even though the representation was trained purely for classification. Our architecture bootstraps on this phenomenon without explicitly training the network to have it.

As above, let $\boldsymbol{x}_{ij}$ and $\boldsymbol{q}$ be a CNN cell at the spatial position $i, j$, and a question representation respectively. We first embed $\boldsymbol{q} \in \mathbb{R}^q$ and $\boldsymbol{x} \in \mathbb{R}^x$ into two feature spaces that share the same dimensionality $d$, i.e.,

$$\hat{\boldsymbol{x}} := CNN^{1\times1}(\boldsymbol{x}; \theta_x) \in \mathbb{R}^{w \times h \times d} \tag{2}$$

$$\hat{\boldsymbol{q}} := MLP(\boldsymbol{q}; \theta_q) \in \mathbb{R}^d \tag{3}$$

where $CNN^{1\times1}$ stands for a $1 \times 1$ convolutional network and $MLP$ stands for a multilayer perceptron. We then combine both the convolutional image features with the question features into a shared *multimodal* embedding by first broadcasting the question features to match the $w \times d$ shape of the image feature map, and then performing element-wise addition (1x1 conv net/MLP in Figure 2):

$$\boldsymbol{m}_{ij} := \hat{\boldsymbol{x}}_{ij} \oplus \hat{\boldsymbol{q}} \text{ , where } \boldsymbol{m} := [\boldsymbol{m}_{ij}]_{ij} \in \mathbb{R}^{w \times h \times d} \tag{4}$$

Element-wise addition keeps the dimensionality of each input, as opposed to concatenation, yet is still effective [12,24]. Next, we compute the *presence vector*, $\boldsymbol{p} := [p_{ij}]_{ij} \in \mathbb{R}^{w \times h}$ which measures the relevance of entities given the question:

$$p_{ij} := ||\boldsymbol{m}_{ij}||_2 \in \mathbb{R} \tag{5}$$

where $||\cdot||_2$ denotes $L2$-norm. To select $k$ entities from $\boldsymbol{m}$ for further processing, the indices of the top $k$ entries in $\boldsymbol{p}$, denoted $\boldsymbol{l} = [l_1, \ldots, l_k]$ are used to form $\hat{\boldsymbol{m}}^k = [\boldsymbol{m}_{l_1}, ..., \boldsymbol{m}_{l_k}] \in \mathbb{R}^{k \times d}$.

This set of features is passed to the decoder module and gradients will flow back to the weights of the CNN/MLP through the selected features. Our assumption is that important outputs of the CNN/MLP will tend to grow in norm, and therefore are likely to be selected. Intuitively if non-useful features are selected, the gradients will push the norm of these features down, making them less likely to be selected again. But there is nothing in our framework which explicitly incorporates this behavior into a loss. Despite its simplicity, our experiments (Section 4) show the HAN is very competitive with canonical soft attention [9] while also offering interpretability and efficiency.

Thus far, we have assumed that we can fix the number of features $k$ that are passed through the attention mechanism. However, it is likely that different questions require different spatial support within the image. Thus, we also introduce a second approach which adaptively chooses the number of entities to attend to (termed Adaptive-HAN, or AdaHAN) as a function of the inputs, rather than using a fixed $k$. The key idea is to make the presence vector $\boldsymbol{p}$ (the norm of the embedding at each spatial location) "compete" against a threshold

$\tau$. However, since the norm is unbounded from above, to avoid trivial solutions in which the network sets the presence vector very high and selects all entities, we apply a softmax operator to $\boldsymbol{p}$. We put both parts into the competition by only selecting those elements of $\boldsymbol{m}$ whose presence values exceed the threshold,

$$\hat{\boldsymbol{m}}^k = [\boldsymbol{m}_{l_1}, ..., \boldsymbol{m}_{l_k}] \in \mathbb{R}^{k \times d} \text{ , where } \{l_i : \text{softmax}(\boldsymbol{p}_{l_i}) > \tau\} \tag{6}$$

Note that due to the properties of softmax, the competition is encouraged not only between both sides of the inequality, but also between the spatially distributed elements of the presence vector $\boldsymbol{p}$. Although $\tau$ could be chosen through the hyper-parameter selection, we decide to use $\tau := \frac{1}{w \cdot h}$ where $w$ and $h$ are spatial dimensions of the input vector $x_{ij}$. Such value for $\tau$ has an interesting interpretation. If each spatial location of the input were equally important, we would sample the locations from a uniform probability distribution $p(\cdot) := \tau = \frac{1}{w \cdot h}$. This is equivalent to a probability distribution induced by the presence vector of a neural network with uniformly distributed spatial representation, i.e. $\tau = \text{softmax}(\boldsymbol{p}_{\text{uniform}})$, and hence the trained network with the presence vector $\boldsymbol{p}$ has to "win" against the $\boldsymbol{p}_{\text{uniform}}$ of the random network in order to select right input features by shifting the probability mass accordingly. It also naturally encourages higher selectivity as the increase in the probability mass at one location would result in decrease in another location.

In contrast to the commonly used soft-attention mechanism, our approaches do not require extra learnable parameters. HAN requires a single extra but interpretable hyper-parameter: a fraction of input cells to use, which trades off speed for accuracy. AdaHAN requires no extra hyper-parameters.

## 3.2   Feature Aggregation

**Sum Pooling.** A simple way to reduce the set of feature vectors after attention is to sum pool them into a constant length vector. In the case of a soft attention module with an attention weight vector $w$, it is straightforward to compute a pooled vector as $\sum_{ij} w_{ij} \boldsymbol{x}_{ij}$. Given features selected with hard attention, an analogous pooling can be written as $\sum_{\kappa=1}^{k} \boldsymbol{m}_{l_\kappa}$.

**Non-local Pairwise Operator.** To improve on sum pooling, we explore an approach which performs reasoning through non-local and pairwise computations, one of a family of similar architectures which has shown promising results for question-answering and video understanding [25,26,27]. An important aspect of these non-local pairwise methods is that the computation is quadratic in the number of features, and thus hard attention can provide significant computational savings. Given some set of embedding vectors (such as the spatial cells of the output of a convolutional layer) $x_{ij}$, one can use three simple linear projections to produce a matrix of queries, $\boldsymbol{q}_{ij} := \boldsymbol{W}_q \boldsymbol{x}_{ij}$, keys, $\boldsymbol{k}_{ij} := \boldsymbol{W}_k \boldsymbol{x}_{ij}$, and values, $\boldsymbol{v}_{ij} = \boldsymbol{W}_v \boldsymbol{x}_{ij}$ at each spatial location. Then, for each spatial location $i, j$, we compare the query $q_{ij}$ with the keys at all other locations, and sum the

values $v$ weighted by the similarity. Mathematically, we compute

$$\tilde{\boldsymbol{x}}_{lk} = \sum_{ij} \text{softmax}\left(\boldsymbol{q}_{lk}^T \boldsymbol{k}_{ij}\right) \boldsymbol{v}_{ij} \tag{7}$$

Here, the softmax operates over all $i, j$ locations. The final representation of the input is computed by summarizing all $\tilde{\boldsymbol{x}}_{lk}$ representations, e.g. we use sum-pooling to achieve this goal. Thus, the mechanism computes non-local [26] pairwise relations between embeddings, independent of spatial or temporal proximity. The separation between keys, queries, and values allows semantic information about each object to remain separated from the information that binds objects together across space. The result is an effective, if somewhat expensive, spatial reasoning mechanism. Although expensive, similar mechanism has been shown useful in various tasks, from synthetic visual question [25], to machine translation [27], to video recognition [26]. Hard attention can help to reduce the set of comparisons that must be considered, and thus we aim to test whether the features selected by hard attention are compatible with this operator.

## 4 Results

To show the importance of hard attention for visual QA, we first compare HAN to existing soft-attention (SAN) architectures on VQA-CP v2, and exploring the effect of varying degrees of hard attention by directly controlling the number of attended spatial cells in the convolutional map. We then examine AdaHAN, which adaptively chooses the number of attended cells, and briefly investigate the effect of network depth and pretraining. Finally, we present qualitative results, and also provide results on CLEVR to show the method's generality.

### 4.1 Datasets

**VQA-CP v2.** This dataset [7] consists of about 121K (98K) images, 438K (220K) questions, and 4.4M (2.2M) answers in the train (test) set; and it is created so that the distribution of the answers between train and test splits differ, and hence the models cannot excessively rely on the language prior [7]. As expected, [7] show that performance of all visual QA approaches they tested drops significantly between train to test sets. The dataset provides a standard train-test split, and also breaks questions into different question types: those where the answer is yes/no, those where the answer is a number, and those where the answer is something else. Thus, we report accuracy on each question type as well as the overall accuracy for each network architecture.

**CLEVR.** This synthetic dataset [34] consists of 100K images of 3D rendered objects like spheres and cylinders, and roughly 1m questions that were automatically generated with a procedural engine. While the visual task is relatively simple, solving this dataset requires reasoning over complex relationships between many objects.

## 4.2   Effect of Hard Attention

We begin with the most basic hard attention architecture, which applies hard attention and then does sum pooling over the attended cells, followed by a small MLP. For each experiment, we take the top $k$ cells, out of 100, according to our $L2$-norm criterion, where $k$ ranges from 16 to 100 (with 100, there is no attention, and the whole image is summed). Results are shown in the top of Table 1. Considering that the hard attention selects only a subset of the input cells, we might expect that the algorithm would lose important information and be unable to recover. In fact, however, the performance is almost the same with less than half of the units attended. Even with just 16 units, the performance loss is less than 1%, suggesting that hard attention is quite capable of capturing the important parts of the image.

| | Percentage of cells | Overall | Yes/No | Number | Other |
|---|---|---|---|---|---|
| HAN+sum | 16% | 26.99 | 40.53 | 11.38 | 24.15 |
| HAN+sum | 32% | 27.43 | 41.05 | 11.38 | 24.68 |
| HAN+sum | 48% | 27.94 | 41.35 | 11.93 | 25.27 |
| HAN+sum | 64% | 27.80 | 40.74 | 11.29 | 25.52 |
| sum | 100% | 27.96 | 43.23 | 12.09 | 24.29 |
| HAN+pairwise | 16% | 26.81 | 41.24 | 10.87 | 23.61 |
| HAN+pairwise | 32% | 27.45 | 40.91 | 11.48 | 24.75 |
| HAN+pairwise | 48% | 28.23 | 41.23 | 11.40 | 25.98 |
| Pairwise | 100% | 28.06 | 44.10 | 13.20 | 23.71 |
| SAN [7,9] | - | 24.96 | 38.35 | 11.14 | 21.74 |
| SAN (ours) | - | 26.60 | 39.69 | 11.25 | 23.92 |
| SAN+pos (ours) | - | 27.77 | 40.73 | 11.31 | 25.47 |
| GVQA [7] | - | 31.30 | 57.99 | 13.68 | 22.14 |

Table 1: Comparison between different number of attended cells (percentage of the whole input), and aggregation operation. We consider a simple summation, and non-local pairwise computations as the aggregation tool.

The fact that hard attention can work is interesting itself, but it should be especially useful for models that devote significant processing to each attended cell. We therefore repeat the above experiment with the non-local pairwise aggregation mechanism described in section 3, which computes activations for every pair of attended cells, and therefore scales quadratically with the number of attended cells. These results are shown in the middle of Table 1, where we can see that hard attention (48 entities) actually boosts performance over an analogous model without hard attention.

Finally, we compare standard soft attention baselines in the bottom of Table 1. In particular, we include previous results using a basic soft attention network [7,9], as well as our own re-implementation of the soft attention pooling algorithm presented in [7,9] with the same features used in other experiments. Surprisingly, soft attention does not outperform basic sum pooling, even with careful implementation that outperforms the previously reported results with the same method on this dataset; in fact, it performs slightly worse. The non-local pairwise aggregation performs better than SAN on its own, although the best result includes hard attention. Our results overall are somewhat worse than the state-of-the-art [7], but this is likely due to several architectural decisions not included here, such as a split pathway for different kinds of questions, special question embeddings, and the use of the question extractor.

|                    | Percentage of cells | Overall | Yes/No | Number | Other |
|--------------------|---------------------|---------|--------|--------|-------|
| AdaHAN+sum         | 25.66%              | 27.40   | 40.70  | 11.13  | 24.86 |
| AdaHAN+pairwise    | 32.63%              | 28.65   | 52.25  | 13.79  | 20.33 |
| HAN+sum            | 32%                 | 27.43   | 41.05  | 11.38  | 24.68 |
| HAN+sum            | 48%                 | 27.94   | 41.35  | 11.93  | 25.27 |
| HAN+pairwise       | 32%                 | 27.45   | 40.91  | 11.48  | 24.75 |
| HAN+pairwise       | 48%                 | 28.23   | 41.23  | 11.40  | 25.98 |

Table 2: Comparison between different adaptive hard-attention techniques with average number of attended parts, and aggregation operation. We consider a simple summation, and the non-local pairwise aggregation. Since AdaHAN adaptively selects relevant features, based on the fixed threshold $\frac{1}{w*h}$, we report here the average number of attended parts.

## 4.3   Adaptive hard attention

Thus far, our experiments have dealt with networks that have a fixed threshold for all images. However, some images and questions may require reasoning about more entities than others. Therefore, we explore a simple adaptive method, where the network chooses how many cells to attend to for each image. Table 2 shows results, where AdaHAN refers to our adaptive mechanism. We can see that on average, the adaptive mechanism uses surprisingly few cells: 25.66 out of 100 when sum pooling is used, and 32.63 whenever the non-local pairwise aggregation mechanism is used. For sum pooling, this is on-par with a non-adaptive network that uses more cells on average (HAN+sum 32); for the non-local pairwise aggregation mechanism, just 32.63 cells are enough to outperform our best non-adaptive model, which uses roughly 50% more cells. This shows that even

| | Percentage of cells | Number of layers | Overall | Yes/No | Number | Other |
|---|---|---|---|---|---|---|
| HAN+sum 25% | | 0 | 26.38 | 43.21 | 13.12 | 21.17 |
| HAN+sum 50% | | 0 | 26.75 | 41.42 | 10.94 | 23.38 |
| HAN+sum 75% | | 0 | 26.82 | 41.30 | 11.48 | 23.42 |
| HAN+sum 25% | | 2 | 26.99 | 40.53 | 11.38 | 24.15 |
| HAN+sum 50% | | 2 | 27.43 | 41.05 | 11.38 | 24.68 |
| HAN+sum 75% | | 2 | 27.94 | 41.35 | 11.93 | 25.27 |

Table 3: Comparison between different number of the attended cells as the percentage of the whole input. The results are reported on VQA-CP v2. The second column denotes the percentage of the attended input. The third column denotes number of layers of the MLP (Equations 2 and 3).

very simple methods of adapting hard attention to the image and the question can lead to both computation and performance gains, suggesting that more sophisticated methods will be an important direction for future work.

### 4.4   Effects of network depth

In this section, we briefly analyze an important architectural choice: the number of layers used on top of the pretrained embeddings. That is, before the question and image representations are combined, we perform a small amount of processing to "align" the information, so that the embedding can easily tell the relevance of the visual information to the question. Table 3 shows the results of removing the two layers which perform this function. We consistently see a drop of about 1% without the layers, suggesting that deciding which cells to attend to requires different information than the classification-tuned ResNet is designed to provide.

### 4.5   Implementation Details.

All our models use the same LSTM size 512 for questions embeddings, and the last convolutional layer of the ImageNet pre-trained ResNet-101 [63] (yielding 10-by-10 spatial representation, each with 2048 dimensional cells) for image embedding. We also use MLP with 3 layers of sizes: $1024, 2048, 1000$, as a classification module. We use ADAM for optimization [65]. We use a distributed setting with two workers computing a gradient over a batch of 128 elements each. We normalize images by dividing them by their norm. We do not perform a hyper-parameter search as there is no separated validation set available. Instead, we rather choose default hyper-parameters based on our prior experience on visual QA datasets. We trained our models until we notice a saturation on the training set. Then we evaluate these models on the test set. Our tables show

the performance of all the methods wrt. the second digits precision obtained by rounding.

Table 1 shows SAN's [9] results reported by [7] together with our in-house implementation (denoted as "ours"). Our implementation has 2 attention hops, 1024 dimensional multimodal embedding size, a fixed learning rate 0.0001, and ResNet-101. In these experiments we pool the attended representations by weighted average with the attention weights. Our in-house implementation of the non-local pairwise mechanism strongly resembles implementations of [26], and [27]. We use 2 heads, with embedding size 512. In Equation 2 and Equation 3, we use $d := 2048$ (the same as dimensionality as the image encoding) and two linear layers with RELU that follows up each layer.

## 4.6   Qualitative Results.

One advantage of our formulation is that it is straightforward to visualize the masks of attended cells given questions and images (which we defer to Figures 1 and 2 in the supplementary material due to space constraints). In general, relevant objects are usually attended, and that significant portions of the irrelevant background is suppressed. Although some background might be kept, we hypothesize the context matters in answering some questions. These masks are occasionally useful for diagnosing behavior: for example, AdaHAN with sum pooling (row 2 in Figure 1) attends incorrectly to the bridge but not the train in the second column, and therefore answers incorrectly. In the tennis court, however, the same method attends incorrectly, but still answers correctly by chance.

We can also see broad differences between the network architectures. For instance, the sum pooling method (row 2) is much more spatially constrained than the pairwise pooling version (row 1), even though the adaptive attention can select an arbitrarily large region. This suggests that sum pooling struggles to integrate across complex scenes. The support is also not always contiguous: non-adaptive hard attention with 16 entities (row 4) in particular distributes its attention widely.

## 4.7   End-to-end Training.

Since our network is not fully differentiable, one might suspect that it will become more difficult to train the lower-level features, or worse, that untrained features might prevent us from bootstrapping the attention mechanism. Therefore, we also trained HAN+sum (with 16% of the input cells) end-to-end together with a relatively small convolutional neural network initialized from scratch. We compare our method against our implementation of the SAN method trained using the same simple convolutional neural network. We call the models: simple-SAN, and simple-HAN.

**Analysis.** In our experiments, simple-SAN achieves about 21% performance on the test set. Surprisingly, simple-HAN+sum achieves about 24% performance on the same split, on-par with the performance of normal SAN that uses more

complex and deeper visual architecture [66]; the results are reported by [7]. This result shows that the hard attention mechanism can indeed be tightly coupled within the training process, and that the whole procedure does not rely heavily on the properties of the ImageNet pre-trained networks. In a sense, we see that a discrete notion of entities also "emerges" through the learning process, leading to efficient training.

**Implementation Details.** In our experiments we use a simple CNN built of: 1 layer with 64 filters and 7-by-7 filter size followed up by 2 layers with 256 filters and 2 layers with 512 filters, all with 3-by-3 filter size. We use strides 2 for all the layers.



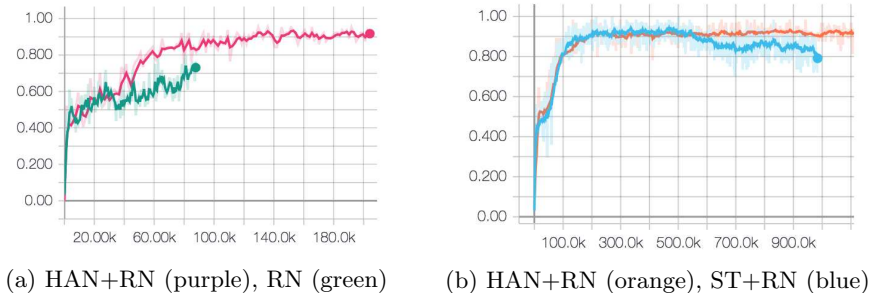(a) HAN+RN (purple), RN (green)        (b) HAN+RN (orange), ST+RN (blue)

Fig. 3: Validation accuracy plots on CLEVR of the methods under the same hyper-parameters setting [25]. (a) HAN+RN (0.25 of the input cells) and standard RN (all input cells) trained for 12 hours to measure the efficiency of the methods. (b) Our approaches to hard attention: the proposed one (orange), and the straight-through estimator (blue).

## 4.8   CLEVR

To demonstrate the generality of our hard attention method, particularly in domains that are visually different from the VQA images, we experiment with a synthetic visual QA dataset termed CLEVR [34], using a setup similar to the one used for VQA-CP and [25]. Due to the visual simplicity of CLEVR, we follow up the work of [25], and instead of relying on the ImageNet pre-trained features, we train our HAN+sum and HAN+RN (hard attention with relation network) architectures end-to-end together with a relatively small CNN (following [25]).

**Analysis.** As reported in prior work [25,34], the soft attention mechanism used in SAN does not perform well on the CLEVR dataset, and achieves only 68.5% [34] (or 76.6% [25]) performance. In contrast, relation network, which also realizes a non-local and pairwise computational model, essentially solves this task, achieving 95.5% performance on the test set. Surprisingly, our HAN+sum

achieves 89.7% performance even without a relation network, and HAN+RN (i.e., relation network is used as an aggregation mechanism) achieves 93.9% on the test set. These results show the mechanism can readily be used with other architectures on another dataset with different visuals. Training with HAN requires far less computation than the original relation network [25], although performance is slightly below relation network's 95.5%. Figure 3a compares computation time: HAN+RN and relation network are trained for 12 hours under the same hyper-parameter set-up. Here, HAN+RN achieves around 90% validation accuracy, whereas RN only 70%. Notably, owing to hard-attention, we are able to train larger models, and achieve 94.7% and 98.8% for HAN+sum and HAN+RN respectively (more details are found in the supplementary material). Although others report slightly better results on CLEVR [49,50], these are not evaluated on real-world datasets such as VQA-CP, or use higher image resolution. We also found the performance to be sensitive to depth, and batch normalization [67], which we present in more detail is the supplementary material.

As an additional baseline, we have experimented with straight-through estimator [17] (supplementary), but we have found it quite unstable (Figure 3b). We also point out that it lacks the training-time computational benefit of our approach: in straight-though, the gradients are still back-propagated through non-selected cells.

## 5   Summary

We have introduced a new approach for hard attention in computer vision that selects a subset of the feature vectors for further processing based on the their magnitudes. We explored two models, one which selects subsets with a pre-specified number of vectors (HAN), and the other one that adaptively chooses the subset size as a function of the inputs (AdaHAN). Hard attention is often avoided in the literature because it poses a challenge for gradient-based methods due to non-differentiability. However, since we found our feature vectors' magnitudes correlate with relevant information, our hard attention mechanism exploits this property to perform the selection. Our results showed our HAN and AdaHAN gave competitive performance on challenging visual QA datasets. Our approaches seem to be at least as good as a more commonly used soft attention mechanism while providing additional computational efficiency benefits. This is especially important for the increasingly popular class of non-local approaches, which often require computations and memory which are quadratic in the number of the input vectors. Finally, our approach also provides interpretable representations, as the spatial locations of the selected features correspond most strongly to those parts of the image which contributed most strongly.

# References

1. Çukur, T., Nishimoto, S., Huth, A.G., Gallant, J.L.: Attention during natural vision warps semantic representation across the human brain. Nature neuroscience **16**(6) (2013) 763

2. Sheinberg, D.L., Logothetis, N.K.: Noticing familiar objects in real world scenes: the role of temporal cortical neurons in natural vision. Journal of Neuroscience **21**(4) (2001) 1340–1350

3. Simons, D.J., Rensink, R.A.: Change blindness: Past, present, and future. Trends in cognitive sciences **9**(1) (2005) 16–20

4. Mack, A., Rock, I.: Inattentional blindness. Volume 33. MIT press Cambridge, MA (1998)

5. Simons, D.J., Chabris, C.F.: Gorillas in our midst: Sustained inattentional blindness for dynamic events. Perception **28**(9) (1999) 1059–1074

6. Malinowski, M., Fritz, M.: A multi-world approach to question answering about real-world scenes based on uncertain input. In: Advances in Neural Information Processing Systems (NIPS). (2014)

7. Agrawal, A., Batra, D., Parikh, D., Kembhavi, A.: Don't just assume; look and answer: Overcoming priors for visual question answering. arXiv preprint arXiv:1712.00377 (2017)

8. Xu, H., Saenko, K.: Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. arXiv:1511.05234 (2015)

9. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016)

10. Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal compact bilinear pooling for visual question answering and visual grounding. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). (2016)

11. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: Proceedings of the Conference on Artificial Intelligence (AAAI). (2018)

12. Kazemi, V., Elqursh, A.: Show, ask, attend, and answer: A strong baseline for visual question answering. arXiv preprint arXiv:1704.03162 (2017)

13. Teney, D., Anderson, P., He, X., Hengel, A.v.d.: Tips and tricks for visual question answering: Learnings from the 2017 challenge. arXiv preprint arXiv:1708.02711 (2017)

14. Ilievski, I., Yan, S., Feng, J.: A focused dynamic attention model for visual question answering. arXiv:1604.01485 (2016)

15. Mnih, V., Heess, N., Graves, A., et al.: Recurrent models of visual attention. In: Advances in neural information processing systems. (2014) 2204–2212

16. Gulcehre, C., Chandar, S., Cho, K., Bengio, Y.: Dynamic neural turing machine with soft and hard addressing schemes. arXiv preprint arXiv:1607.00036 (2016)

17. Bengio, Y., Léonard, N., Courville, A.: Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv preprint arXiv:1308.3432 (2013)

18. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with gumbel-softmax. arXiv preprint arXiv:1611.01144 (2016)

19. Maddison, C.J., Mnih, A., Teh, Y.W.: The concrete distribution: A continuous relaxation of discrete random variables. arXiv preprint arXiv:1611.00712 (2016)

20. Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., Mordvintsev, A.: The building blocks of interpretability. Distill (2018)

21. Oliva, A., Torralba, A., Castelhano, M.S., Henderson, J.M.: Top-down control of visual attention in object detection. In: Image processing, 2003. icip 2003. proceedings. 2003 international conference on. Volume 1., IEEE (2003) I–253

22. Ren, M., Kiros, R., Zemel, R.: Image question answering: A visual semantic embedding model and a new dataset. In: Advances in Neural Information Processing Systems (NIPS). (2015)

23. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). (2015)

24. Malinowski, M., Rohrbach, M., Fritz, M.: Ask your neurons: A deep learning approach to visual question answering. International Journal of Computer Vision (IJCV) **125**(1-3) (2017) 110–135

25. Santoro, A., Raposo, D., Barrett, D.G., Malinowski, M., Pascanu, R., Battaglia, P., Lillicrap, T.: A simple neural network module for relational reasoning. In: Advances in Neural Information Processing Systems (NIPS). (2017) 4974–4983

26. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. arXiv preprint arXiv:1711.07971 (2017)

27. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. (2017) 6000–6010

28. Shaw, P., Uszkoreit, J., Vaswani, A.: Self-attention with relative position representations. arXiv preprint arXiv:1803.02155 (2018)

29. Malinowski, M., Fritz, M.: Towards a visual turing challenge. In: Learning Semantics (NIPS workshop). (2014)

30. Malinowski, M., Fritz, M.: Hard to cheat: A turing test based on answering questions about images. AAAI Workshop: Beyond the Turing Test (2015)

31. Malinowski, M.: Towards holistic machines: From visual recognition to question answering about real-world images. PhD Thesis (2017)

32. Geman, D., Geman, S., Hallonquist, N., Younes, L.: Visual turing test for computer vision systems. In: Proceedings of the National Academy of Sciences, National Academy of Sciences (2015)

33. Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L., Xu, W.: Are you talking to a machine? dataset and methods for multilingual image question answering. In: Advances in Neural Information Processing Systems (NIPS). (2015)

34. Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C.L., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, IEEE (2017) 1988–1997

35. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017)

36. Harnad, S.: The symbol grounding problem. Physica D: Nonlinear Phenomena **42**(1) (1990) 335–346

37. Guadarrama, S., Riano, L., Golland, D., Gouhring, D., Jia, Y., Klein, D., Abbeel, P., Darrell, T.: Grounding spatial relations for human-robot interaction. In: IROS. (2013)

38. Kong, C., Lin, D., Bansal, M., Urtasun, R., Fidler, S.: What are you talking about? text-to-image coreference. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2014)
39. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2015)
40. Rohrbach, A., Rohrbach, M., Hu, R., Darrell, T., Schiele, B.: Grounding of textual phrases in images by reconstruction. In: European Conference on Computer Vision, Springer (2016) 817–834
41. Malinowski, M., Rohrbach, M., Fritz, M.: Ask your neurons: A neural-based approach to answering questions about images. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). (2015) 1–9
42. Xu, K., Ba, J., Kiros, R., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. Proceedings of the International Conference on Machine Learning (ICML) (2015)
43. Xiong, C., Merity, S., Socher, R.: Dynamic memory networks for visual and textual question answering. arXiv preprint arXiv:1603.01417 (2016)
44. De Vries, H., Strub, F., Mary, J., Larochelle, H., Pietquin, O., Courville, A.C.: Modulating early visual processing by language. In: Advances in Neural Information Processing Systems. (2017) 6597–6607
45. Zhu, Y., Groth, O., Bernstein, M., Fei-Fei, L.: Visual7W: Grounded Question Answering in Images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016)
46. Chen, K., Wang, J., Chen, L.C., Gao, H., Xu, W., Nevatia, R.: Abc-cnn: An attention based convolutional neural network for visual question answering. arXiv:1511.05960 (2015)
47. Shih, K.J., Singh, S., Hoiem, D.: Where to look: Focus regions for visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016)
48. Gulcehre, C., Denil, M., Malinowski, M., Razavi, A., Pascanu, R., Hermann, K.M., Battaglia, P., Bapst, V., Raposo, D., Santoro, A., et al.: Hyperbolic attention networks. arXiv preprint arXiv:1805.09786 (2018)
49. Hudson, D.A., Manning, C.D.: Compositional attention networks for machine reasoning. arXiv preprint arXiv:1803.03067 (2018)
50. Mascharka, D., Tran, P., Soklaski, R., Majumdar, A.: Transparency by design: Closing the gap between performance and interpretability in visual reasoning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2018) 4942–4950
51. Kim, Y., Denton, C., Hoang, L., Rush, A.M.: Structured attention networks. arXiv preprint arXiv:1702.00887 (2017)
52. Hu, R., Andreas, J., Darrell, T., Saenko, K.: Explainable Neural Computation via Stack Neural Module Networks. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018)
53. Mokarian, A., Malinowski, M., Fritz, M.: Mean box pooling: A rich image representation and output embedding for the visual madlibs task. In: Proceedings of the British Machine Vision Conference (BMVC). (2016)
54. Tommasi, T., Mallya, A., Plummer, B., Lazebnik, S., Berg, A.C., Berg, T.L.: Solving visual madlibs with multiple cues. In: Proceedings of the British Machine Vision Conference (BMVC). (2016)
55. Desta, M.T., Chen, L., Kornuta, T.: Object-based reasoning in vqa. arXiv preprint arXiv:1801.09718 (2018)

56. Singh, S., Gupta, A., Efros, A.A.: Unsupervised discovery of mid-level discriminative patches. In: Proceedings of the European Conference on Computer Vision (ECCV). Springer (2012) 73–86
57. Doersch, C., Gupta, A., Efros, A.A.: Mid-level visual element discovery as discriminative mode seeking. In: Advances in Neural Information Processing Systems (NIPS). (2013) 494–502
58. Juneja, M., Vedaldi, A., Jawahar, C., Zisserman, A.: Blocks that shout: Distinctive parts for scene classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE (2013) 923–930
59. Doersch, C., Singh, S., Gupta, A., Sivic, J., Efros, A.: What makes Paris look like Paris? SIGGRAPH (2012)
60. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: Advances in Neural Information Processing Systems (NIPS). (2015) 2017–2025
61. Mallya, A., Lazebnik, S.: Packnet: Adding multiple tasks to a single network by iterative pruning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE (2018)
62. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (NIPS). (2012)
63. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv:1512.03385 (2015)
64. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation (1997)
65. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv:1412.6980 (2014)
66. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)
67. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)