How Transferable are Reasoning Patterns in VQA?

Corentin Kervadec[1,2*] Théo Jaunet[2*] Grigory Antipov[1]
Moez Baccouche[1] Romain Vuillemot[2] Christian Wolf[2]
[1]Orange, Cesson-Sévigné, France [2]LIRIS, INSA - École Centrale, Lyon, UMR CNRS 5205, France
corentinkervadec.github.io theo-jaunet.github.io christian.wolf@insa-lyon.fr
{grigory.antipov, moez.baccouche}@orange.com romain.vuillemot@gmail.com

Abstract

Since its inception, Visual Question Answering (VQA) is notoriously known as a task, where models are prone to exploit biases in datasets to find shortcuts instead of performing high-level reasoning. Classical methods address this by removing biases from training data, or adding branches to models to detect and remove biases. In this paper, we argue that uncertainty in vision is a dominating factor preventing the successful learning of reasoning in vision and language problems. We train a visual oracle and in a large scale study provide experimental evidence that it is much less prone to exploiting spurious dataset biases compared to standard models. We propose to study the attention mechanisms at work in the visual oracle and compare them with a SOTA Transformer-based model. We provide an in-depth analysis and visualizations of reasoning patterns obtained with an online visualization tool which we make publicly available[1]. We exploit these insights by transferring reasoning patterns from the oracle to a SOTA Transformer-based VQA model taking standard noisy visual inputs via fine-tuning. In experiments we report higher overall accuracy, as well as accuracy on infrequent answers for each question type, which provides evidence for improved generalization and a decrease of the dependency on dataset biases.

1. Introduction

The high prediction performance obtained by high-capacity deep networks trained on large-scale data has led to questions concerning the nature of these improvements. Visual Question Answering (VQA) in particular has become a testbed for the evaluation of the reasoning and generalization capabilities of trained models, as it combines multiple modalities of heterogeneous nature (images and language) with open questions and large varieties. It has been shown, that current models are prone to exploiting harmful biases in the data, which can provide unwanted shortcuts to learning in the form of "*Clever Hans*" effects [36, 22].

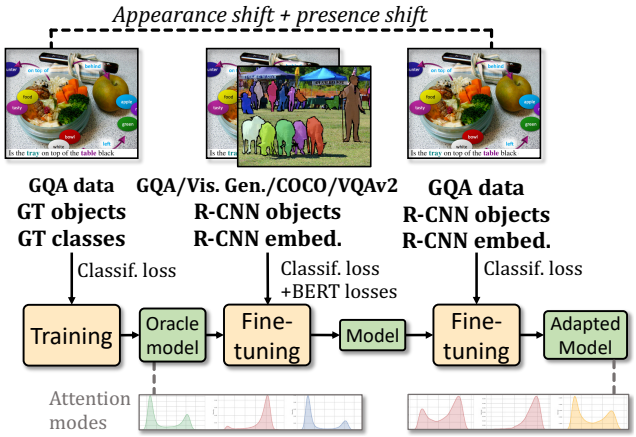


Figure 1. We argue that noise and uncertainties in visual inputs are the main bottleneck in VQA preventing successful learning of reasoning capacities. In a deep analysis, we show that oracle models with perfect sight, trained on noiseless visual data, tend to depend significantly less on bias exploitation. We exploit this by training models on data without visual noise, and then transfer the learned reasoning patterns to real data. We illustrate successful transfer by an analysis and visualization of attention modes.

In this work we study the capabilities of VQA models to "*reason*". An exact definition of this term is difficult, we refer to [7, 22] and define it as "*algebraically manipulating words and visual objects to answer a new question*". In particular, we interpret reasoning as the opposite of exploiting spurious biases in training data. We argue, and in Section 3 will provide evidence for this, that learning to algebraically manipulate words and objects is difficult when visual input is noisy and uncertain compared to learning from perfect information about a scene. When objects are frequently missing, detected multiple times or recognized with ambiguous visual embeddings wrongly overlapping with different categories, relying on statistical shortcuts may be an easy short-

*Both authors contributed equally.
[1]https://reasoningpatterns.github.io

cut for the optimizer [2] We show, that a perfect-sighted oracle model learns to predict answers while significantly less relying on biases in training data. We claim that once any noise has been removed from visual input, replacing object detection output by Ground Truth (GT) object annotations, a deep neural network can more easily learn the reasoning patterns required for prediction and for generalization.

In the line of recent work in AI explainability [28, 32], and data visualization [18, 38, 11], we propose an in-depth analysis of attention mechanisms in Transformer-based models and provide indications of the patterns of reasoning employed by models of different strengths. We visualize different operating modes of attention and link them to different sub tasks ("*functions*") required for solving VQA. In particular, we use this analysis for a comparison between oracle models and standard models processing noisy and uncertain visual input, highlighting the presence of reasoning patterns in the former and less so in the latter.

Drawing conclusions from this analysis, we propose to fine-tune the perfectly-sighted oracle model on the real noisy visual input (see Fig. 1). Using the same analysis and visualization techniques, we show that attention modes absent from noisy models are transferred successfully from oracle models to deployable [3] models, and we report improvements in overall accuracy and generalization.

Contributions — (i) An in-depth analysis of reasoning patterns at work in Transformer-based models, comparing Oracles vs. deployable models, including visualization of attention modes; an analysis of the relationships between attention modes and reasoning, and the impact of attention pruning on reasoning. (ii) We propose to transfer reasoning capabilities, learned by the oracle, to SOTA VQA methods with noisy input and improve overall performance and generalization on the GQA [19] dataset (iii) We show that this transfer is complementary with self-supervised large-scale pre-training (LXMERT [35]/BERT-like).

2. Related work

Visual Question Answering (VQA) — as a task was introduced in various datasets, such as VQAv1 [5] and VQAv2 [17] (built from human annotators), or CLEVR [20] and GQA [19] (automatically-generated from fully-synthetic and real-world images, respectively). Additional splits were proposed to evaluate specific reasoning capabilities. For instance, VQA-CP [1] explicitly inverts the answer distribution between train and test splits. Following recent critics and controversies about these evaluations [36, 33], the GQA-OOD dataset [22] introduced a new split of GQA focusing on rare (Out-Of-Distribution / OOD) question-answer pairs, and showed that many VQA models strongly rely on dataset biases. This growing amount of diverse datasets has been accompanied by the development of more sophisticated VQA models. While an exhaustive survey of methods is out of the scope of this paper, one can mention families based on object-level attention [3], bilinear fusion [23], tensor decomposition [6], neural-symbolic reasoning [41], neural [4] and meta [9] module networks.

Transformers and Vision-Language reasoning — In this work, we focus on Transformers [37] due to their wide adoption and their powerful attention mechanism. MCAN [42] and DFAF [15] introduced the use of object-level self-attention and co-attention mechanisms to model intra- and inter-modality interactions in VQA. More recent work [21, 10, 29, 35] suggests that the combination of Transformers with a large-scale BERT [12]-like pretraining can be beneficial for VQA. Self-attention on pixel-level [2, 13] is, but up to our knowledge, not used done for VQA.

Attention and reasoning patterns in Transformers — Analysis of self-attention mechanisms has received considerable attention recently. In [39], a visualization tool for analysing BERT attention layers is introduced. [11] studies how training strategies and fine-tuning impact attention in BERT-like models. Voita et al. [40] classifies BERT's attention heads according to their functionality, reporting a significant simplification of the model's complexity via pruning. Finally, [30, Appendix A.5] measures energy distributions and classifies them based on their meta-stable states.

Following this work in NLP, similar studies have appeared in vision and language. [21] explores the emergence of word-object alignment in the attention maps when adding a weakly supervised objective. [8, 26] study to what extent the attention maps in BERT-like pretrained VQA Transfomers encode various vision-language information. While these methods provide a better understanding of the amount of information captured by VQA models, they do not shed light on how this information is used. In our work, we analyze how various VQA tasks are encoded in different attention heads. To this end, we apply an energy-based analysis inspired by [30]. In addition, we study attention in perfect-sighted oracle Transformers in order to identify which patterns lead to better reasoning. Our findings lead to an *Oracle Transfer* strategy, which allows to improve performance and generalization in standard transformer models. Finally, our work is related to [14], which found evidence for relationships between questions and the modulation of (non transformer) model parameters on the synthetic CLEVR [20] dataset.

3. Analysis of Reasoning Patterns

In this section we will analyze reasoning behavior in Transformer-based VQA models and make the case for the impact of training on visual GT data. First, to make the paper self-contained, we provide a short introduction into

[2] See [16] for an interesting review of *shortcut learning*.

[3] *Deployable*: the model **does not** use ground-truth visual inputs.

Vision-Language (VL)-Transformers as proposed in standard literature [42, 15, 35, 29, 25, 10, 34].

Given two different modalities, Vision (V) and Language (L), VL-Transformers are composed of the succession of intra-modality $T_{-}^{V}(\cdot)$, $T_{-}^{L}(\cdot)$ and inter-modality $T_{\times}^{V \leftarrow L}(\cdot,\cdot)$, $T_{\times}^{L \leftarrow V}(\cdot,\cdot)$ multi-head attention layers [37]. As defined in the seminal paper [37], multi-head attention layers $T(\cdot)$ (both intra- and inter-modality ones) can be expressed as a set of self-attention layers $t(\cdot)$ which are performed in parallel on several "heads". For example, given an input sequence $\boldsymbol{v}$ of the visual embeddings, a visual intra-modality n-head attention layer $T_{-}^{V}(\cdot)$ performs as a set of h visual intra-modality self-attention layers $\{t_{-}^{V(1)}(\cdot), ..., t_{-}^{V(h)}(\cdot)\}$, the outputs of which are concatenated and then combined:

$$T_{-}^{V}(\boldsymbol{v}) = \left[t_{-}^{V(1)}(\boldsymbol{v}),\ ...,\ t_{-}^{V(h)}(\boldsymbol{v}) \right] W^{\boldsymbol{O}} \qquad (1)$$

where $W^{\boldsymbol{O}}$ is a trainable matrix which is particular for each multi-head attention layer. Each layer $t_{-}(\cdot)$ is defined on a set of input (vision or language) embeddings $\boldsymbol{x}$ of the same dimension d as

$$t_{-}(\boldsymbol{x}) = \sum_{j} \boldsymbol{\alpha}_{ij} \boldsymbol{x}_{j}^{v}, \qquad (2)$$

where the query $\boldsymbol{x}^{q}$, key $\boldsymbol{x}^{k}$ and value $\boldsymbol{x}^{v}$ matrices are given as follows: $\boldsymbol{x}^{q}{=}\boldsymbol{W}^{q}\boldsymbol{x}$, $\boldsymbol{x}^{k}{=}\boldsymbol{W}^{k}\boldsymbol{x}$ and $\boldsymbol{x}^{v}{=}\boldsymbol{W}^{v}\boldsymbol{x}$. All $\boldsymbol{W}^{\cdot}$ are trainable parameters. In particular, $\boldsymbol{x}^{q}$ and $\boldsymbol{x}^{k}$ are used to calculate the self-attention weights $\boldsymbol{\alpha}_{\cdot j}$ as follows:

$$\boldsymbol{\alpha}_{\cdot j}{=}(\boldsymbol{\alpha}_{1j}, ..., \boldsymbol{\alpha}_{ij}, ..., \boldsymbol{\alpha}_{nj}){=}\sigma(..., \frac{\boldsymbol{x}_{i}^{q^{T}} \boldsymbol{x}_{j}^{k}}{\sqrt{d}}, ...), \qquad (3)$$

with σ being the softmax operator. In this paper, we mainly focus on the attention maps $\{\alpha_{ij}\}$ which are composed of these self-attention weights.

Finally, inter-modality self-attention layers $t_{\times}(\cdot,\cdot)$ are defined in the same way, as the intra-modality ones, but unlike the latter they calculate queries, keys and values on two sets of input embeddings of different modalities. More precisely, for the self-attention layer $t_{\times}^{V \leftarrow L}(\boldsymbol{v}, \boldsymbol{l})$, the query matrix $\boldsymbol{v}^{q} = \boldsymbol{W}^{q}\boldsymbol{v}$ is calculated on vision embeddings, while the key $\boldsymbol{l}^{k} = \boldsymbol{W}^{k}\boldsymbol{l}$ and the value matrices $\boldsymbol{l}^{v} = \boldsymbol{W}^{v}\boldsymbol{l}$ are calculated on the language ones. For the $t_{\times}^{L \leftarrow V}(\boldsymbol{l}, \boldsymbol{v})$ self-attention layer, the matrices are calculated symmetrically ($\boldsymbol{l}^{q} = \boldsymbol{W}^{q}\boldsymbol{l}$, $\boldsymbol{v}^{k} = \boldsymbol{W}^{k}\boldsymbol{v}$ and $\boldsymbol{v}^{v} = \boldsymbol{W}^{v}\boldsymbol{v}$, respectively). In addition, each $t_{\times}^{V \leftarrow L}$ (resp. $t_{\times}^{L \leftarrow V}$) is followed by a self-attention t_{-}^{V} (resp. t_{-}^{L}). A general view of the architecture is available in the supp. mat..

Experimental setup — All analyses in this section have been performed with a hidden embedding size $d = 128$ and a number of per-layer heads $h = 4$. This corresponds to a tiny version of the architecture used in LXMERT [35] where $d = 768$ and $h = 12$. Therefore, "tiny-LXMERT" corresponds to the VL-Transformer architecture plus BERT-like (LXMERT) pre-training. Unless specified otherwise, objects have been detected with Faster R-CNN [31]. Visualizations are done on GQA [19] (validation set) as it is particularly well suited for evaluating a large variety of reasoning skills. However, as GQA contains synthetic questions constructed from pre-defined templates, the dataset only offers a constrained VQA environment. Additional experiments might be required to extend our conclusions to more natural setups.

3.1. Visual noise vs. models with perfect-sight

We conjecture that difficulties in the computer vision pipeline are the main cause preventing VQA models in learning to reason well, and which leads them to exploit spurious biases in training data. Most of these methods use pre-trained off-the-shelf object detectors during training and evaluation steps. But in a significant number of cases, the visual objects necessary for reasoning are misclassified, or even not detected at all, as indicated by detection rates of SOTA detectors on the Visual Genome dataset [24], for instance. Under these circumstances, even a perfect VQA model is unable to predict correct answers without relying on statistical shortcuts.

To further explore this working hypothesis, we trained an oracle model with perfect sight, *i.e* a model which receives perfect visual input, and compare it with tiny-LXMERT. Based on the same VL-Transformer, it receives the GT objects from the GQA annotations, encoded as GT bounding boxes and 1-in-K encoded object classes replacing the visual embeddings of the classical model. All GT objects are fed to the model, not only objects required for reasoning. We study the capabilities of both models, the oracle model and the classical one, to "*reason*". Following [22] we measure the reasoning capabilities of a VQA model as the capacity to correctly answer questions, where the GT answer is rare w.r.t. the question group, *i.e* the type of questions being asked. We evaluate the models on the GQA-OOD benchmark [22] designed for OOD evaluation.

Fig. 2 illustrates the model behavior in different situations. At the extreme case (left side of the plot), the model is evaluated on the rarest samples only, while on the right side all samples are considered. We observe that the performance of the classical model taking noisy visual (tiny-LXMERT) drops sharply for (image, question) pairs with rare GT answers, which is an indication for a strong dependency on dataset biases. We would like to insist that in this benchmark the rarity of a GT answer is determined w.r.t. the question type, which allows to measure biases taking into account language. The oracle model, on the other hand, obtains performances which are far less dependent on answer rarity, providing evidence for its ability to overcome

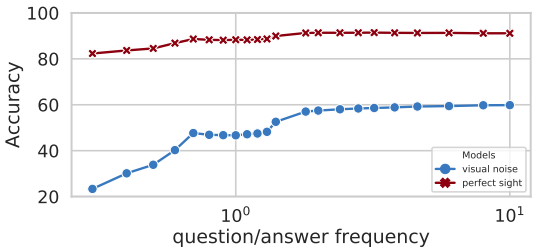
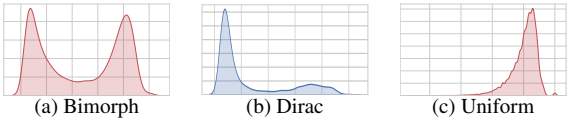


Figure 2. Uncertainties and noise in visual input dominate the difficulties in learning reasoning: comparison of the out-of-distribution generalization between two different VQA Models. A perfectly-sighted oracle model and a standard noisy vision based model trained on the GQA-OOD benchmark [22]. For the classical model, accuracy drops for questions where the GT answer is rare (left side) compared to frequent answers (right side), indicating probable bias exploitation. In contrast, the oracle obtains high performance also on rare answers. Both models are "tiny-LXMERT".



(a) Bimorph (b) Dirac (c) Uniform

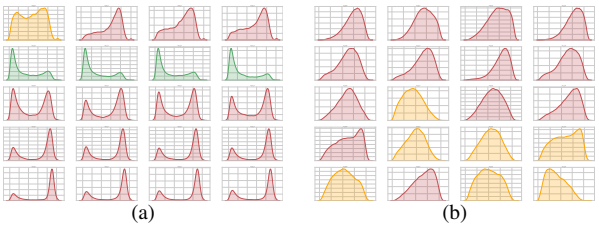
Figure 3. Attention modes learned by the oracle model. Following [30], for each head we plot the distribution of the number k of tokens required to reach 90% of the attention energy (GQA-val). X-axis (from 0 to 100%): ratio of the tokens k w.r.t. the total number of tokens. Plots are not attention distributions, but distributions of indicators of attention distributions. We observe three major modes: (a) "*bimorph*" attention[5], unveil two different types of attention distribution for the same head; (b) dirac attention with high k-median, *i.e* small meta stable state; (c) uniform attention, with low k-median, *i.e* very large meta stable state.

statistical biases. As a consequence, we conjecture that the visual oracle is closer to a real "reasoning process", by predicting answer resulting from a manipulation of words and objects, rather than by having captured statistical shortcuts. In the absence of GT on reasoning, we admit that there is no formal proof to this statement, but we believe that the evidence above is sufficient.

3.2. Attention modes in VL-Transformers

Attention distributions are at the heart of the VL-Transformer. They are not directly supervised during training, their behavior emerges from training the different VQA objectives, *i.e* the discriminative loss as well as the eventual additional BERT-like objectives [35]. Their definition as a strength of association between different items makes them a prime candidate for visualization of inner workings of deep models. We analyze attention, and in particular we observe different attention modes in trained VQA models.

Following [30] , we visualize the distribution of attention energy associated with each Transformer head in multi-headed attention. For each attention map, associated with a given head for a given sample, we calculate the number k of tokens required to reach a total sum of 90% of the distribution energy. A low k-number is caused by peaky attention, called *small meta-stable state* in [30], while a high k-number indicates uniform attention, close to an average operation (*very large meta-stable state*). For each head, and over a subset of validation samples, we plot the distribution of k-numbers, and for some experiments we summarize it with a median value taken over samples and over tokens.



(a) (b)

Figure 4. Comparison of k-distribution of $t_{\times}^{L\leftarrow V}$ attention heads for two different models: (a) oracle; (b) noisy visual input. Rows indicates different $T_{\times}^{L\leftarrow V}$ layers. Heads are colored according to the median of the k-number.

Diversity in attention modes — In this experiment we focus on the oracle VL-Transformer, where we observed a high diversity in attention modes. We also observed that some layers' heads, especially those processing the visual modality (t_{-}^{V} or $t_{\times}^{V\leftarrow L}$) are mainly working with close-to-average attention distributions (*very large meta-stable states* [30]). On the other hand, we observed smaller meta-stable states in the language layers (t_{-}^{L} or $t_{\times}^{L\leftarrow V}$). This indicates that the reasoning process in the oracle VL-Transformer is in large part executed by the model as a transformation of the language features, which are successively contextualized (i.e. influenced) by the visual features (and not the opposite).

In contrast to the attention modes reported in [30], we also observed bi-modal k-number distributions, shown in Fig. 3-a, which are a combination of a dirac (Fig. 3-b) and uniform (cf Fig. 3-c) attention modes. We call these modes "*bimorph*" attention, since they reveal the existence of two different shapes of attention distribution: for some samples, a dirac activation is generated, while other samples lead to uniform attention (averaging over tokens)[5]. Besides, in Fig. 4, we compare attention mode diversity between the noisy visual model and the oracle $t_{\times}^{L\leftarrow V}$ heads, where we observe higher diversity for the oracle. In particular, "*bimorph*" attention are mostly performed by the oracle.

3.3. Attention modes and task functions

In this experiment, we study the relationships between attention modes and question types, which correspond to dif-

[5]We remind that these plots are distributions of indicators of distributions: uniform behavior does not show up as a flat plot, but as plot with a peak on the right side — it may in these plots look like a Dirac.

ferent functions of reasoning required to solve the problem instance. In other words, we explore to what extent the neural model adapts its attention distribution to the question at hand. We group the set of questions according to functions using the GQA [19] annotation, using 54 different functions such as *e.g* "filter color", "verify size", *etc*.[6].

We link functions to the attention modes introduced in Section 3.2. In Fig. 5 we show functions in columns and a selection of attention heads in rows, while the color encodes the median k-number for the oracle model. We observe a strong dependency. Certain functions, *e.g* the majority of the "choose X" functions, tend to cause the emergence of small meta-stable states. In these modes, the attention mechanism is fundamental as it allows the model to attend to specific token combinations by detecting specific patterns. On the other hand, some functions requiring to attend to very general image properties, such as "choose location" or "verify weather", seem to be connected to very large meta-stable states. We conjecture, that to find general scene properties, a large context is needed. In this modes, the attention mechanism is less important, and replacing it with a simple averaging operation is likely to keep performance — an experiment we explore in Section 3.4. Similarly, when focusing on heads instead of functions, we observe that a majority of heads typed as $t_{\times}^{V\leftarrow L}(\cdot)$ or $t_{-}^{V}(\cdot)$ tends to behave independently of the question functions and they generally show close-to-uniform attention. On the other hand, the $t_{-}^{L}(\cdot)$ and $t_{\times}^{L\leftarrow V}(\cdot)$ heads are highly dependant on the question functions. As shown in Fig.s 5 and 6, these heads does not behave in the same way and are not "activated" (*i.e* have a smaller metastable-state) for the same combination of functions. This provides some evidence for modularity of the oracle VL-Transformer, each attention head learning to specialize to one or more functions. In addition, in Fig. 6, we visualize the difference in oracle attention modes between two different function configurations: Fig. 6-a is the distribution of median k-numbers over *all* samples, *i.e* involving all functions, whereas Fig. 6-b shows the distribution over samples involving the "choose color" function. We show the 3rd $T_{\times}^{V\leftarrow L}$ Transformer layer heads. Over all functions, these heads show "*bimorph*" behavior, whereas on questions requiring to choose a color, these same heads show either dirac or uniform behavior.

Oracle vs. Noisy Input — In the next experiment, we explore the difference in behavior between the perfect-sighted oracle and the classical model taking noisy visual input. For each input sample, we create a 80-dimensional representation describing the attention behavior of the model by collecting the k-numbers of the 80 cross-attention heads into a flat vector, taking the median over the tokens for a given head. Fig. 7 shows two different t-

[6]There is limited overlap between functions, *e.g* "filter" contains, among others, the "filter color" and "filter size".

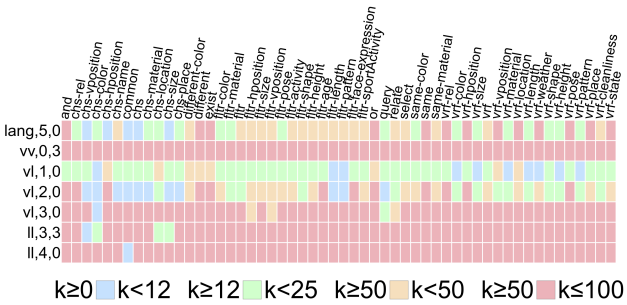


Figure 5. Attention modes for selected attention heads (rows) related to functions required to be solved to answer a question (columns). The head's notation x, i, j refers to the head j of the i-th Transformer layer of type x: 'lang'/'ll'=$t_{-}^{L}(\cdot)$, 'vis'/'vv'=$t_{-}^{V}(\cdot)$, 'vl'=$t_{\times}^{L\leftarrow V}(\cdot)$, 'lv'=$t_{\times}^{V\leftarrow L}(\cdot)$. The VL-Transformer's architecture is presented in supp. mat.. The color encodes the attention mode, *i.e* median of the k-number [30]. We observe (1) attention heads behave differently depending on the function; (2) a given function causes different attention modes for different heads.

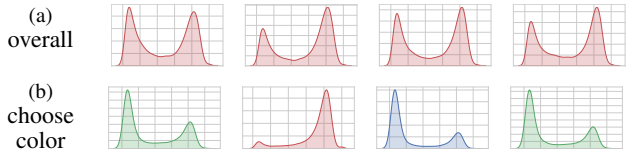
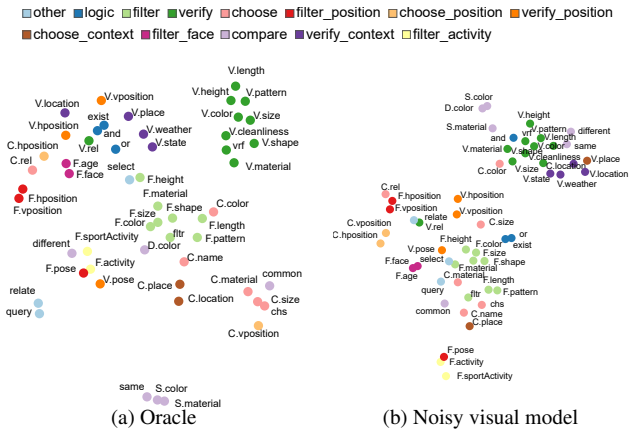


Figure 6. Influence of the question on oracle's "*bimorph*" attention heads. We compare attention modes of the third layer of $T_{\times}^{L\leftarrow V}$ heads as a distribution of the k-numbers [30] over (a) samples of all functions, and (b) samples with questions involving the "choose color" function, and observe a clear difference. The function "choose color" seems to cause the activation (*i.e* emergence of a small meta-stable state) of the 1^{st}, 2^{nd} and 4^{th} head, and the desactivation of the 3^{rd} one, further indicating task dependence of attention head behavior.

SNE projections of these attention behavior space, one for the oracle model and one for the noisy model. While the former produces clusters regrouping functions according to their general type, the function representation of the noisy model is significantly more entangled. We conjecture, that the attention-function relationship provides insights into the reasoning strategies of the model. VQA requires to handle a large variety of reasoning skills and different operations on the input objects and words. Question-specific manipulation of words and objects is essential for correct reasoning. In contrast to the oracle one, the t-SNE plot for the noisy visual model paints a muddier picture, and does not show clear relationships between attention modes and functions.

Caveat — visualizing attention modes does not provide any indication of the attention operation itself, only about the shape of the operation. In particular, an attention head might result in the same low k-number for two different input samples, showing Dirac attention, but could attend do quite different objects or words in both cases.



(a) Oracle (b) Noisy visual model

Figure 7. t-SNE projection of the attention mode space, *i.e* the 80-dim representation median k-numbers, one per head of the model. Colors are functions, also provided as overlaid text. We compare projections of (a) the oracle, and (b) the noisy visual model, and observe a clustering of functions in the attention mode space for the oracle, but significantly less for the noisy input model.

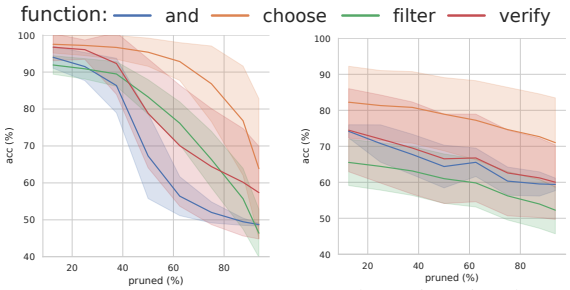
Pruned attentions	n/a	L	V	L←V	V←L
Accuracy	91.5	37.9	91.4	52.8	68.1

Table 1. Impact of pruning different types of attention heads of the trained oracle model. We observe that 'vision' and 'language→vision' Transformers are hardly impacted by pruning, in contrast to 'language' and 'vision→language'. Accuracies (in %) on the GQA validation set.

3.4. Attention pruning

We further analyze the role of attention heads by evaluating the effect of pruning heads on model performance. As reported by [40, 30], specific attention heads may be useful during training, but less useful after training. In the same lines, for specific heads we replace the query-key attention map by a uniform one, "*pruned*" heads will therefore simply contextualize each token by an averaged representation of all other tokens, as a head with large meta-stable state would have done. In Table 1 we report the effect of pruning on GQA validation accuracy according to different attention categories and observe that the oracle model is resilient to pruning of the $t^V_-(\cdot)$ and $t^{V\leftarrow L}_\times(\cdot)$ heads, but that pruning of $t^L_-(\cdot)$ and $t^{L\leftarrow V}_\times(\cdot)$ heads results in sharp drops in performance. This indicates that the bulk of reasoning occurs over the language tokens and embeddings, which are contextualized from the visual information through $t^{L\leftarrow V}_\times(\cdot)$ cross-attention. We can only conjecture why this solution emerges after training — we think that among reasons are the deep structure of language and the fact that in current models the answer is predicted from the CLS language token.

Impact on functions — We study the impact of pruning on the different task functions by randomly pruning n cross-attention heads and measuring accuracy for different



(a) Oracle (b) Noisy visual

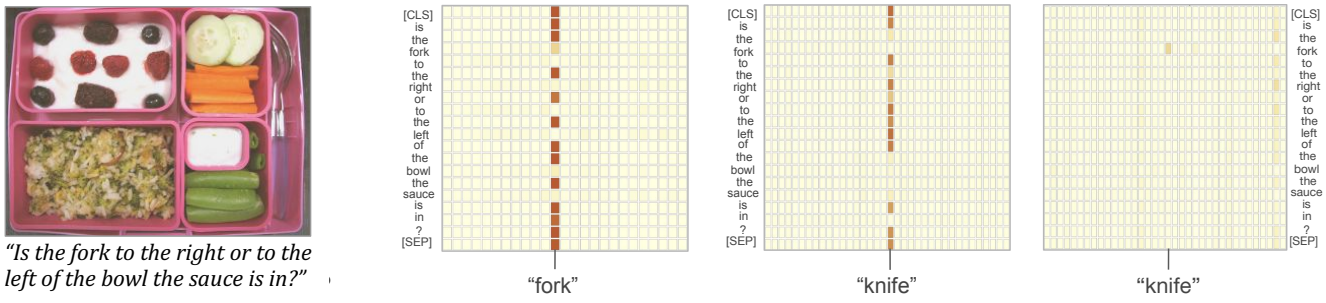
Figure 8. Impact of random pruning of varying numbers of attention heads in cross-modal layers on GQA-validation accuracy. (a) For the oracle, the impact is related to the nature of the function, highlighting its modular property. (b) For the noisy-vision-based model, pruning seems to be unrelated to function types.

function groups, n being varied between 0% (no pruning) to 100% (all heads are pruned), as shown in Fig. 8 for the oracle and noisy vision-based models. For the sake of clarity only 4 different function are shown, additional results are provided in supplementary material. For the perfect-sighted oracle (Fig. 8-a), we first observe that the pruning has a different impact depending on the function. Thereby, while *filter* and *choose* are dominated by negative curvature where performance drops only when a large number of heads are pruned, *verify* and *and*, are characterized by a sharp inflection point and an early steep drop in performance. This indicates that the model has learned to handle functions specifically, resulting in various degrees of reasoning distribution over attention heads. For the noisy vision-based model, on the other hand, the effect of head pruning seems to be unrelated to the function type (Fig. 8-b).

3.5. Interactive visualization

The analysis described above was based on integrating information of various kinds over a full dataset, GQA-validation. Additional insights can be gained by exploring the behavior of individual problem instances and relating them to statistics extracted from the population, in particular attention modes and groups of functions. We have performed this analysis on a large number of samples and we provide a tool, which allows the reader to perform similar experiments online, making it possible to load different oracle or noisy vision based models. This tool is available at [1] (online experience + source code). Fig. 9 gives a simple visualization (see Section 4 for a discussion), a video of its usage is provided in the supplementary material.

Discussion — The experiments of this section have shown a pronounced difference in attention modes between the perfectly-sighed oracle and a noisy vision based model. More importantly, the oracle model shows a strong relationship between attention mode and task function, which we interpret as the capability of adapting reasoning to the task



(a) Oracle (b) Oracle transfer (c) Baseline

Figure 9. Example for the difference in attention in the second $T_{\times}^{L\leftarrow V}$ layer. The oracle drives attention towards a specific object, "fork", also seen after transfer but not in the baseline (we checked for permutations). The transferred model overcame a miss-labelling of the fork as a knife. This analysis was performed with our interactive visualization tool, which also allows to visualize attention models, not shown here (**https://reasoningpatterns.github.io**, online experience + source code; video provided in the supp. mat.).

at hand. The classical model significantly lacks this abilities, suggesting a strategy of transferring patterns of reasoning from an oracle model pre-trained on visual GT to a model taking noisy visual input.

4. Transferring Reasoning Patterns

We propose *Oracle Transfer*, transferring reasoning patterns from a perfectly-sighted model to a deployable model taking noisy visual inputs. We argue, that the first optimization steps are crucial for the emergence of specific attention modes. Training proceeds as follows (see Fig. 1):

1. Training of a perfectly-sighted oracle model on GT visual inputs from the GQA [19] annotations, in particular a *symbolic* representation concatenating the 1-in-K encoded object class and attributes of each object.

2. Initialize a new model *with the oracle parameters*. This new model is taking noisy visual input in a form of the *dense* representation (2048-dim feature vector extracted by Faster-RCNN [31] fused with bounding-boxes). The first visual layers (T_{-}^{V}) are initialized randomly due to the difference in nature between dense and symbolic representations.

3. Optionally and complementary, *continue training* with large-scale self-supervised objectives (LXMERT [35]/BERT-like) on combined data from Visual Genome [24], MS COCO [27], VQAv2 [17].

4. *Fine-tune* with the standard VQA classification objective on the target dataset (GQA [19] or VQAv2 [17]).

Experimental setup — We use the same VL-Transformer architecture defined in Section 3 (more details in supp. mat.), with d=128 and h=4, which corresponds to a tiny version of LXMERT [35] architecture. Following [35], we use 36 objects per image. We evaluate on the GQA [19], GQA-OOD [22] and VQAv2 [17] datasets.

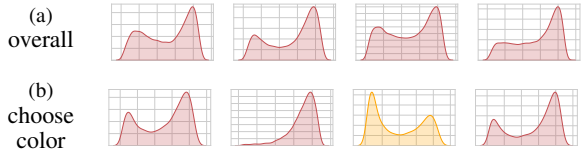


Figure 10. We reproduce Fig. 6 with our VL-Transformer + dense *Oracle Transfer* (same heads/layers). As we can see in (a), the attention heads have retained their "*bimorph*" property, although their shape is distorted by the noisy visual training. In addition, when we measure the attention mode on questions involving the choose color function, in (b), we observe that the attention heads are still function-dependant, although in a lesser extent.

Evaluating transfer — We evaluate the impact of *Oracle Transfer* on three different benchmarks in Table 2, observing that transferring knowledge from the oracle significantly boosts accuracy. We also evaluate the effect of *Oracle Transfer* on bias reduction and benchmark on GQA-OOD [22], reporting gains in Out-Of-Distribution settings — rare samples, "acc-tail" — by a large margin, which suggests improved generalization ability. Our experiments show that *Oracle Transfer* is complementary to large-scale vision-language self-supervised objectives of type LXMERT/BERT-like pretraining as introduced in [35]. An overall gain of about $+1$ accuracy points is observed from models (c) to (d) in Table 2, attributed to *Oracle Transfer*. As a comparison, LXMERT/BERT pretraining alone does not improve "acc-tail" on GQA-OOD.

Cross-dataset training — We explore whether the effects of oracle knowledge generalize beyond the GQA dataset, and evaluate training the oracle on GQA GT annotations, performing LXMERT/BERT pretraining, and transferring to a model trained on VQAv2 dataset [17]. We improve VQAv2 accuracy by a significant margin, suggesting positive transfer beyond GQA (Table 2).

Transfer ablation studies — We evaluate different variants of knowledge transfer, shown in Table 3, on the GQA validation set only. We explore a direct transfer from the

Model	Pretraining		GQA-OOD [22]		GQA [19]	VQAv2 [17]
	Oracle	LXMERT/BERT	acc-tail	acc-head	overall	overall
(a) Baseline			42.9	49.5	52.4	-
(b) Ours	✓		**48.5**	**55.5**	**56.8**	-
(c) Baseline (+LXMERT/BERT)		✓	47.5	54.7	56.8	69.7
(d) Ours (+LXMERT/BERT)	✓	✓	**48.3**	**55.2**	**57.8**	**70.2**

Table 2. Quantitative evaluation of the proposed knowledge transfer from oracle models. All listed models are deployable, no GT input is used for testing. Models: (c)+(d) are pre-trained with LXMERT [35]/BERT-like objectives after *Oracle Transfer*. All scores GQA-OOD-testdev [22]; GQA [19]-testdev; VQAv2-test-std [17]. Training hyperparameters selected on respective validation sets.

Method	Input train	Input test	Acc.
(a) Baseline	Dense	Dense	61.7
(b) Transf. w/o retrain	1-in-K GT	1-in-K pred.	58.8
(c) Transf. w/ T_{-}^{V} retrain	1-in-K GT	Dense	61.7
(d) Transf. w/ retrain	1-in-K GT	Dense	66.3

Table 3. Impact of different types of transfer, GQA [19] val. accuracy. All models are deployable (no GT used for testing).

Method	$\lvert\Theta\rvert$	O	L	OOD	GQA
BUTD [3]	22			42.1	51.6
BAN-4 [23]	50			47.2	54.7
MCAN-6 [42]	52			46.5	56.3
Ours	26	✓		**48.5**	**56.8**
LXMERT-tiny	26		✓	47.5	56.8
LXMERT-tiny + Ours	26	✓	✓	48.3	57.8
LXMERT [35]	212		✓	49.8	59.6

$\lvert\Theta\rvert$ = number of parameters (M); OOD = GQA-OOD [22] Acc-tail.
O = *Oracle Transfer*, L = LXMERT/BERT pretraining.

Table 4. Comparison with SOTA on GQA and GQA-OOD on test-dev. Hyperparameters were optimized on GQA-validation.

oracle to a deployable model without retraining, by making visual input representations comparable. To this end, the deployable model receives 1-in-K encoded class information, albeit not from GT classes but taking classes from the Faster R-CNN detector (Table 3-b). While inferior to the baseline, its performance is surprisingly high, suggesting that the oracle learns knowledge which is applicable in real/noisy settings. Performance gains are, however, only obtained by finetuning the model to the uncertainties in dense visual embeddings. Retraining only the visual block (Table 3-c), performances are on par with the baseline, retraining the full model (Table 3-d) gains $+4.6p$.

Comparison with SOTA — *Oracle Transfer* allows to improve performance of the tiny-LXMERT model both in and out of distribution [22] (Table 4, bottom part). Transfer is parameter efficient and achieves on-par overall accuracy with MCAN-6 [23] while halving capacity.

Qualitative analysis & interpretability — Finally, we qualitatively study the effects of *Oracle Transfer* and interpretability of attention heads. As shown in Fig. 10, after transfer, the VL-Transformer preserves the "*bimorph*" property of its attention heads, which was present in the original oracle model (Fig. 4-a), but absent in the baseline (Fig. 4-b). In addition, Fig. 9 shows the attention maps of the $T_{\times}^{L\leftarrow V}$ heads in the second cross-modal layer for an instance. This head, referenced as $VL, 1, 0$ in Fig. 5, is observed to be triggered to questions such as "verify attr" and "verify color" provided as example. We observe that the oracle model draws attention towards the object "fork" in the image, and also, to a lesser extend, in the transferred model, but not in the baseline model. Similar attention patterns were observed on multiple heads in the corresponding cross-modal layer — this analysis took into account possible permutations of heads between models. Interestingly the miss-classification as a "knife" prevents the baseline from drawing attention to it, but not the transferred model.

5. Conclusion

We have provided a deep analysis and visualizations of several aspects of deep VQA models linked to reasoning on the GQA dataset. We have shown, that oracle models produce significantly better results on questions with rare GT answers than models on noisy data, that their attention modes are more diverse and that they are significantly more dependent on questions. We have also performed instance level analysis and we propose a tool available online[1], which allows to visualize attention distributions and modes, and their links to task functions and dataset wide statistics.

Drawing conclusions from this analysis, we have shown that reasoning patterns can be partially transferred from oracle models to SOTA VQA models based on Transformers and BERT-like pre-training. The accuracy gained from the transfer is particularly high on questions with rare GT-answers, suggesting that the knowledge transferred is related to reasoning, as opposed to bias exploitation.

Acknowledgements — C. Wolf acknowledges support from ANR through grant "*Remember*" (ANR-20-CHIA-0018).

References

[1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[2] Alexander Kolesnikov Dirk Weissenborn Xiaohua Zhai Thomas Unterthiner Mostafa Dehghani Matthias Minderer Georg Heigold Sylvain Gelly Jakob Uszkoreit Neil Houlsby Alexey Dosovitskiy, Lucas Beyer. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*, 2021. 2

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 2, 8

[4] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 39–48, 2016. 2

[5] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 2

[6] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2612–2620, 2017. 2

[7] Léon Bottou. From machine learning to machine reasoning. *Machine learning*, 94(2):133–149, 2014. 1

[8] Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *European Conference on Computer Vision*, pages 565–580. Springer, 2020. 2

[9] Wenhu Chen, Zhe Gan, Linjie Li, Yu Cheng, William Wang, and Jingjing Liu. Meta module network for compositional visual reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 655–664, 2021. 2

[10] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120. Springer, 2020. 2, 3

[11] Joseph F DeRose, Jiayao Wang, and Matthew Berger. Attention flows: Analyzing and comparing attention mechanisms in language models. *IEEE Transactions on Visualization and Computer Graphics*, 2020. 2

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019. 2

[13] B. Duke, A. Ahmed, C. Wolf, P. Aarabi, and G.W. Taylor. SSTVOS: Sparse Spatiotemporal Transformers for Video Object Segmentation. In *CVPR*, 2021. 2

[14] Vincent Dumoulin, Ethan Perez, Nathan Schucher, Florian Strub, Harm de Vries, Aaron Courville, and Yoshua Bengio. Feature-wise transformations. *Distill*, 2018. https://distill.pub/2018/feature-wise-transformations. 2

[15] Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven CH Hoi, Xiaogang Wang, and Hongsheng Li. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6639–6648, 2019. 2, 3

[16] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 2

[17] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017. 2, 7, 8

[18] Fred Hohman, Minsuk Kahng, Robert Pienta, and Duen Horng Chau. Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE Transactions on Visualization and Computer Graphics*, 2018. 2

[19] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6700–6709, 2019. 2, 3, 5, 7, 8

[20] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017. 2

[21] Corentin Kervadec, Grigory Antipov, Moez Baccouche, and Christian Wolf. Weak supervision helps emergence of word-object alignment and improves vision-language tasks. In *European Conference on Artificial Intelligence*, 2019. 2

[22] Corentin Kervadec, Grigory Antipov, Moez Baccouche, and Christian Wolf. Roses are red, violets are blue... but should vqa expect them to? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 1, 2, 3, 4, 7, 8

[23] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *Advances in Neural Information Processing Systems*, pages 1564–1574, 2018. 2, 8

[24] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome:

Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 3, 7

[25] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 3

[26] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. What does bert with vision look at? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5265–5275, 2020. 2

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 7

[28] Zachary C Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016. 2

[29] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019. 2, 3

[30] Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, Victor Greiff, et al. Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*, 2020. 2, 4, 5, 6

[31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 3, 7

[32] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. ACM, 2016. 2

[33] Robik Shrestha, Kushal Kafle, and Christopher Kanan. A negative case analysis of visual grounding methods for vqa. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8172–8181, 2020. 2

[34] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2020. 3

[35] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. 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 5103–5114, 2019. 2, 3, 4, 7, 8

[36] Damien Teney, Ehsan Abbasnejad, Kushal Kafle, Robik Shrestha, Christopher Kanan, and Anton van den Hengel. On the value of out-of-distribution testing: An example of goodhart's law. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 407–417. Curran Associates, Inc., 2020. 1, 2

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2, 3

[38] Jesse Vig. A multiscale visualization of attention in the transformer model. *arXiv preprint arXiv:1906.05714*, 2019. 2

[39] Jesse Vig. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, 2019. 2

[40] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, 2019. 2, 6

[41] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. In *Advances in Neural Information Processing Systems*, pages 1031–1042, 2018. 2

[42] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6281–6290, 2019. 2, 3, 8