How Transferable are Reasoning Patterns in Visual Question Answering?

CVPR VIRTUAL JUNE 19-25

C. Kervadec[1,2] T. Jaunet[2] G. Antipov[1] M. Baccouche[1] R. Vuillemot[3] C. Wolf[2]

[1] Orange Innovation [2] LIRIS, INSA Lyon
[3] LIRIS, EC Lyon

Check-out our interactive demonstration (online)

https://reasoningpatterns.github.io

Reasoning *vs.* Biases in VQA

VQA model are notorious for their tendency to rely on **shortcuts** [2,3], preventing them to learn to **reason**.

We claim that shortcut learning in VQA is in part due to the visual uncertainty (image representation imperfect).

Our contributions

In-depth analysis of **reasoning patterns** at work in VQA

- Analysing attention mechanisms at work in a **VL-Transformer**
- Comparing models with **perfect-sight** *vs.* **noisy visual inputs**

An **oracle transfer** method:

- Transfer reasoning capabilities, learned by the oracle, to a **standard VQA model with noisy input**
- Improve overall performance and generalisation on GQA[1]

Oracle vs. Standard model

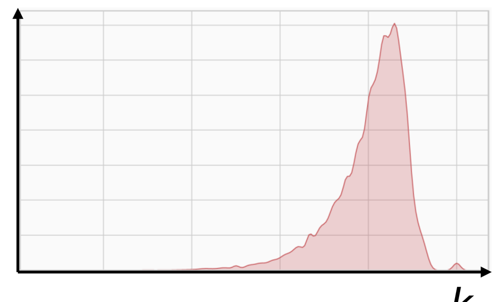
We propose to compare two settings:

- **Standard** *Vision is uncertain* *deployable*
  - Image is represented as a set of objects extracted using a pre-trained object detector
- **Oracle** *Vision is **perfect*** ***not** deployable*
  - Image is represented using human annotations

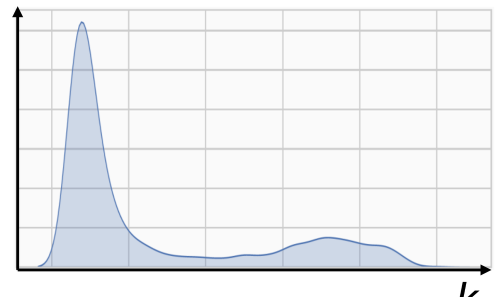
Our experiments are based on a **Vision-Langage (VL)-Transformer**

Attention modes in VL-Transformers

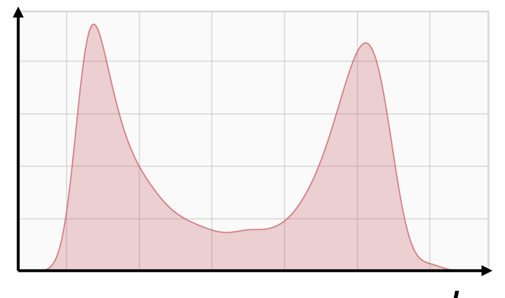
We identify and plot three main attention modes in Oracle's attention heads:



1 **Uniform**
- Average attention



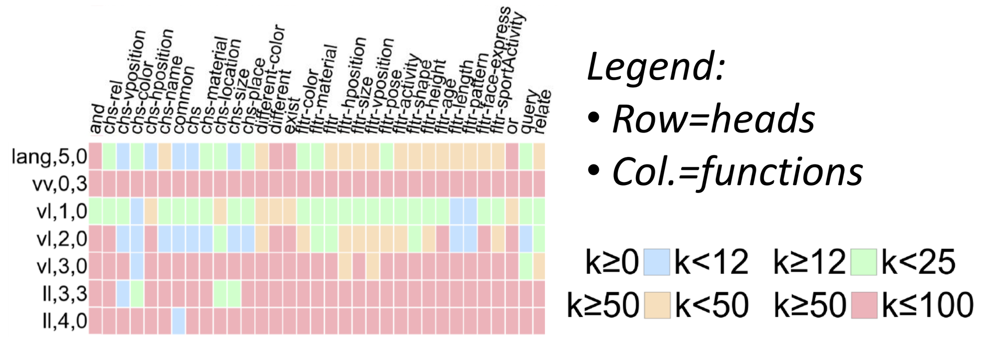
2 **Dirac**
- Peaky attention



3 **Bimorph**
- Depends on the question

Oracle learned functions

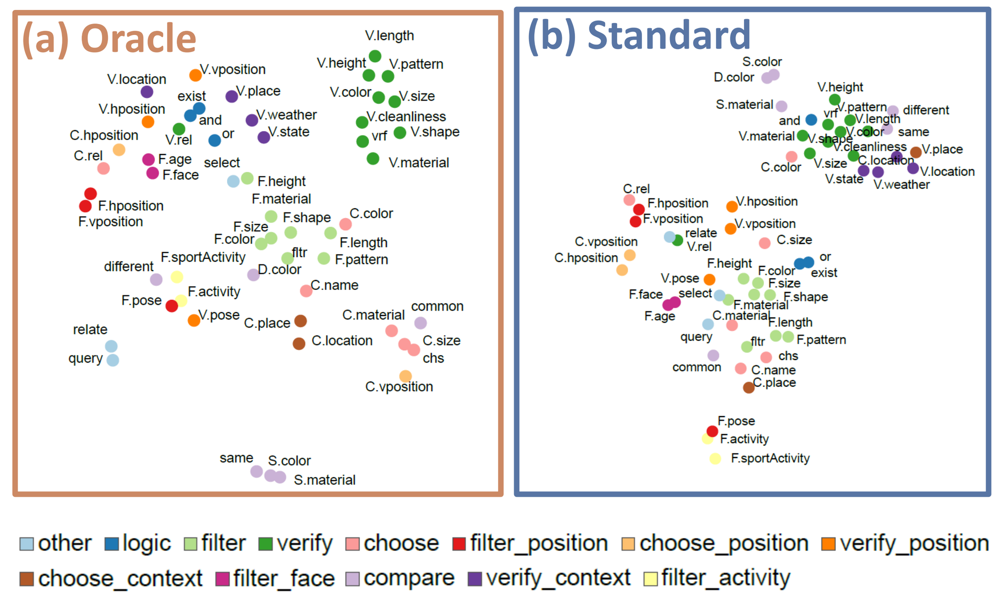
We link functions (*e.g.* `choose color`) to attention modes. Here for Oracle:



- Oracle adapts its attention to the task

But Standard did not

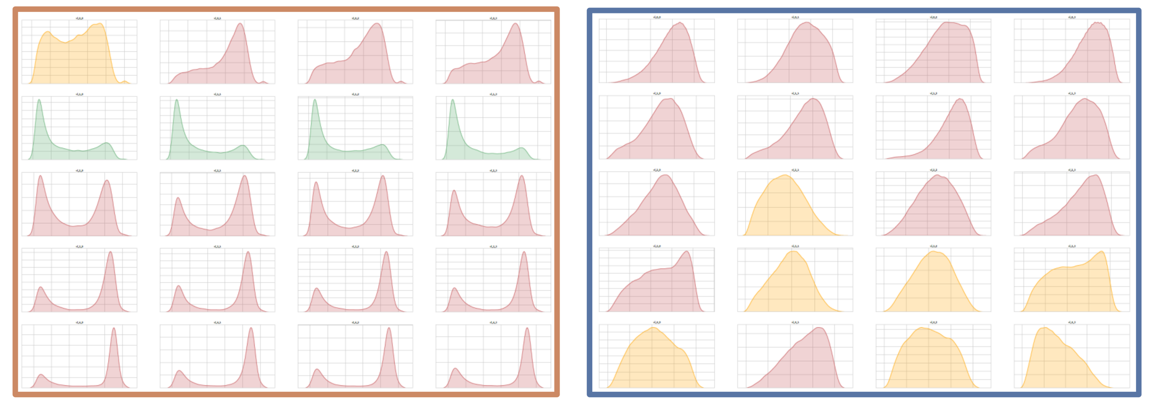
No relationships between attention modes and function in Standard*:



* t-SNE visualisation. See paper to get more details

Higher diversity in Oracle

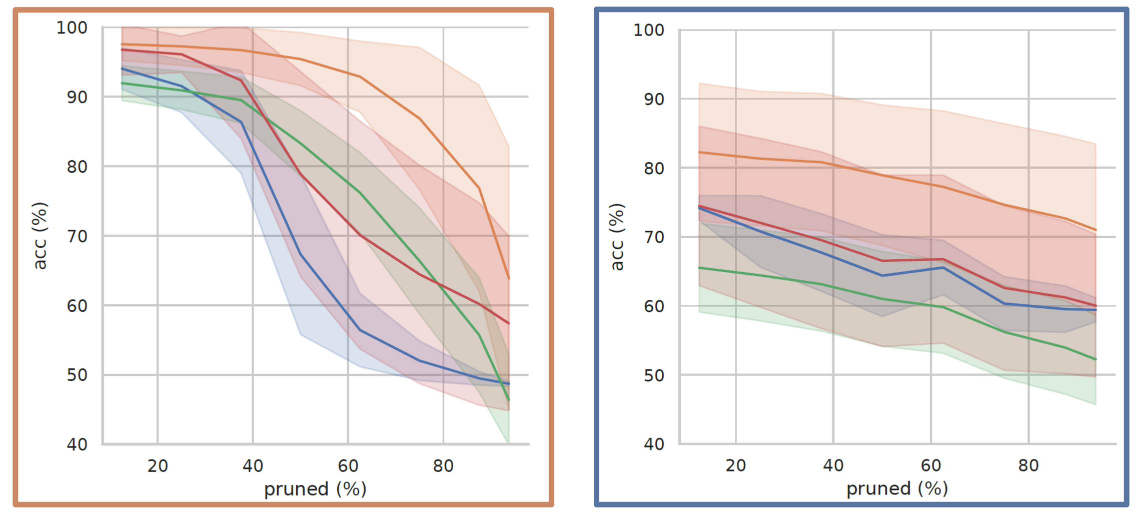
Oracle's heads have more diverse attention modes:
- Standard ones are mostly uniform



(a) Oracle (b) Standard

Head pruning

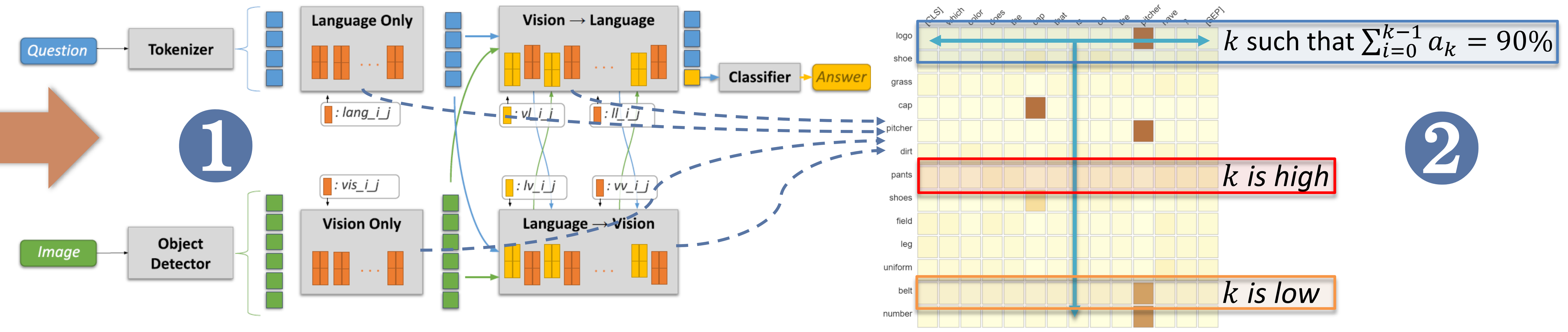
Impact of pruning varying numbers of attention heads in cross-modal layers on accuracy. For Oracle, the impact is related to function. Result are different for Standard.



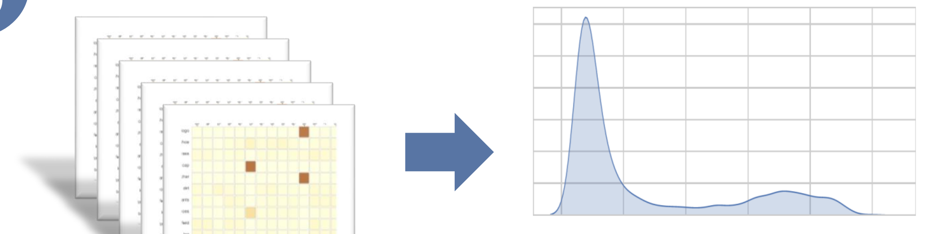
(a) Oracle (b) Standard

function: — and — choose — filter — verify

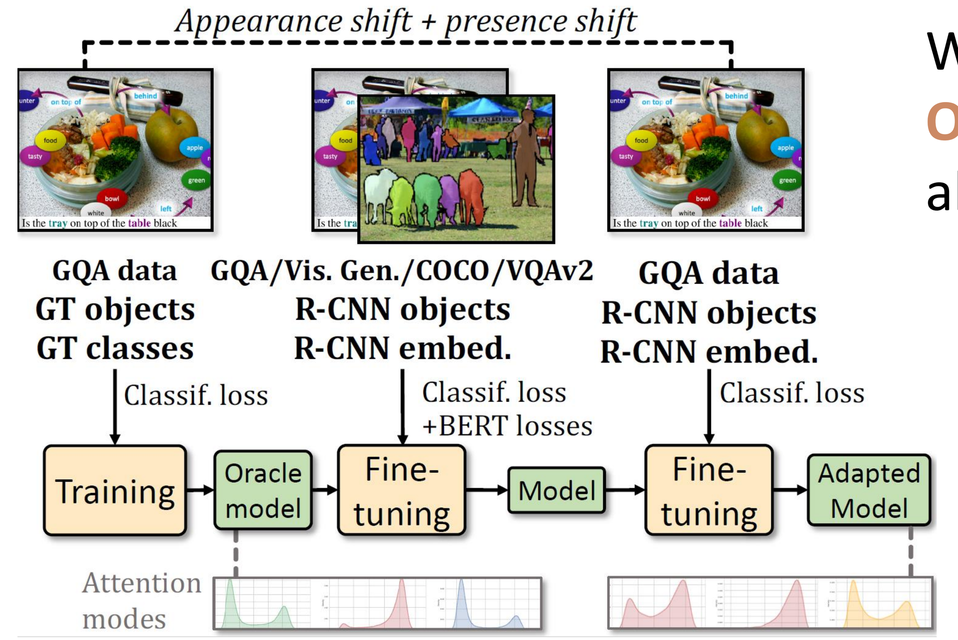
Methodology: attention modes



1 For each head, extract attention maps
2 Measure attention energy ***k***
3 Plot the per-head energy distribution over the dataset



Oracle transfer



We observe significant differences between Oracle and Standard: we highlighted the Oracle ability of adapting reasoning to the task at hand.

We propose to transfer learned reasoning patterns form Oracle to Standard:

1 Train the oracle on perfect vision
2 Optionally, BERT-like pretraining
3 Finetune with standard (noisy) vision

Model	Pretraining Oracle	Pretraining LXMERT/BERT	GQA-OOD [22] acc-tail	GQA-OOD [22] acc-head	GQA [19] overall	VQAv2 [17] 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**

[1] R. Geirhos, et al. Shortcut learning in deep neural networks. In Proc Nature Machine Intelligence 2020
[2] C. Kervadec, et al. Roses are Red, Violets are Blue... But Should VQA expect Them To? In Proc CVPR 2021
[3] D. Hudson, et al. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In Proc CVPR 2019