Lecture: Deep Learning and Differential Programming

4.4 Self-attention and transformers

https://liris.cnrs.fr/christian.wolf/teaching

Learning language reasoning

Все счастливые семьи похожи друг на друга, каждая несчастливая семья несчастлива по-своему..

Toutes les familles heureuses se ressemblent. Chaque famille malheureuse, au contraire, l'est à sa façon.

Happy families are all alike. Every unhappy family is unhappy in its own way.

Alle glücklichen Familien gleichen einander, jede unglückliche Familie ist auf ihre eigene Weise unglücklich

[L. Tolstoy, 1873]

Examples

VO

If you really want to hear about it, the first thing you'll probably want to know is where I was born, and what my lousy childhood was like, and how my parents were occupied and all before they had me, and all that David Copperfield kind of crap, but I don't feel like going into it, if you want to know the truth. In the first place, that stuff bores me, and in the second place, my parents would have two hemorrhages apiece if I told anything pretty personal about them.

VF Google

Si vous voulez vraiment en entendre parler, la première chose que vous voud probablement savoir est où je suis né, et ce que mon enfance moche était, et comment mes parents étaient occupés et tout ce qu'ils avaient avant moi, et tout ce que David Copperfield, mais je n'ai pas envie d'y aller, si tu veux savoir la vérité. En premier lieu, cela m'ennuie, et en second lieu, mes parents auraient deux hémorragies si je leur racontais quelque chose de très personnel.

VF officielle

Si vous voulez vraiment que je vous dise, alors surement la première chose q aller demander c'est où je suis né, et à quoi ça ressemblait, ma saloperie d'enfance et c'est que faisaient mes parents avant de m'avoir, et toutes ces conneries à la David Copperfield, mais j'ai pas envie de raconter ça et tout. Primo, ce genre de trucs ça me rase, et secundo, mes parents ils auraient chacun une attaque ou même deux chacun, si je me mettais à baratiner sur leur compte quelque chose d'un peu personnel.

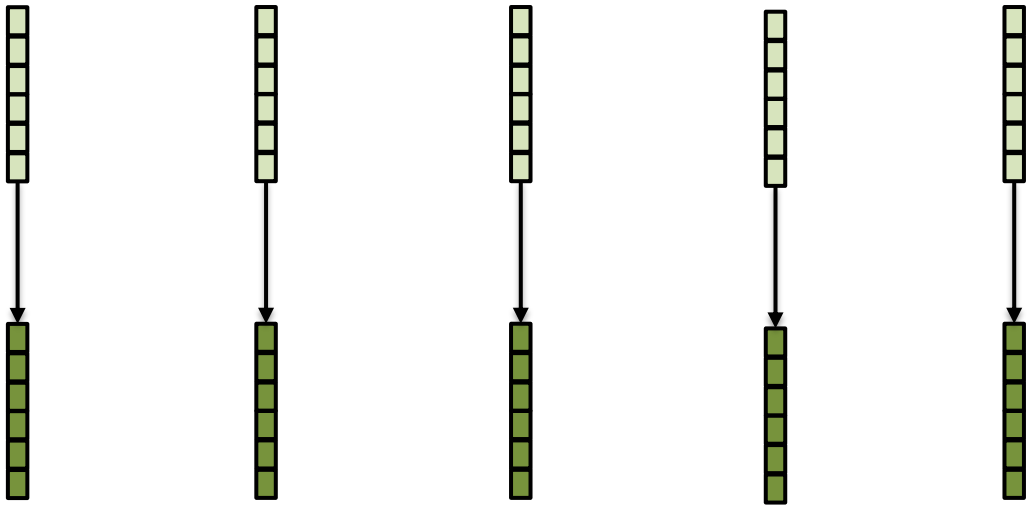
J.D. Salinger, « the catcher in the Rye »

Contextualization

We suppose a set $\mathbf{X} = \{\mathbf{x}_i\}$ of items (vectors).

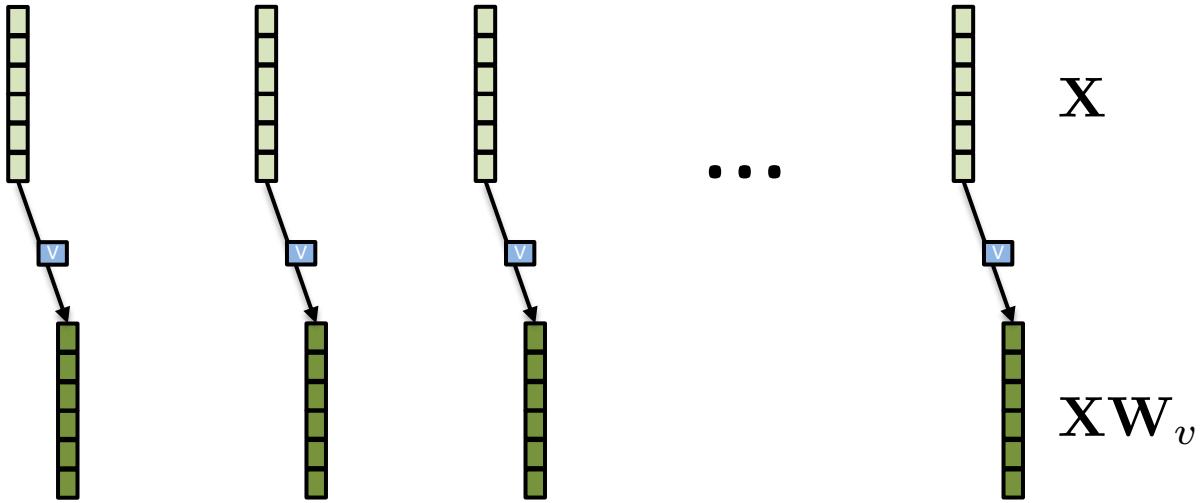
Iteratively "enrich" each item by providing context from the other items

The quick brown fox jumps ....



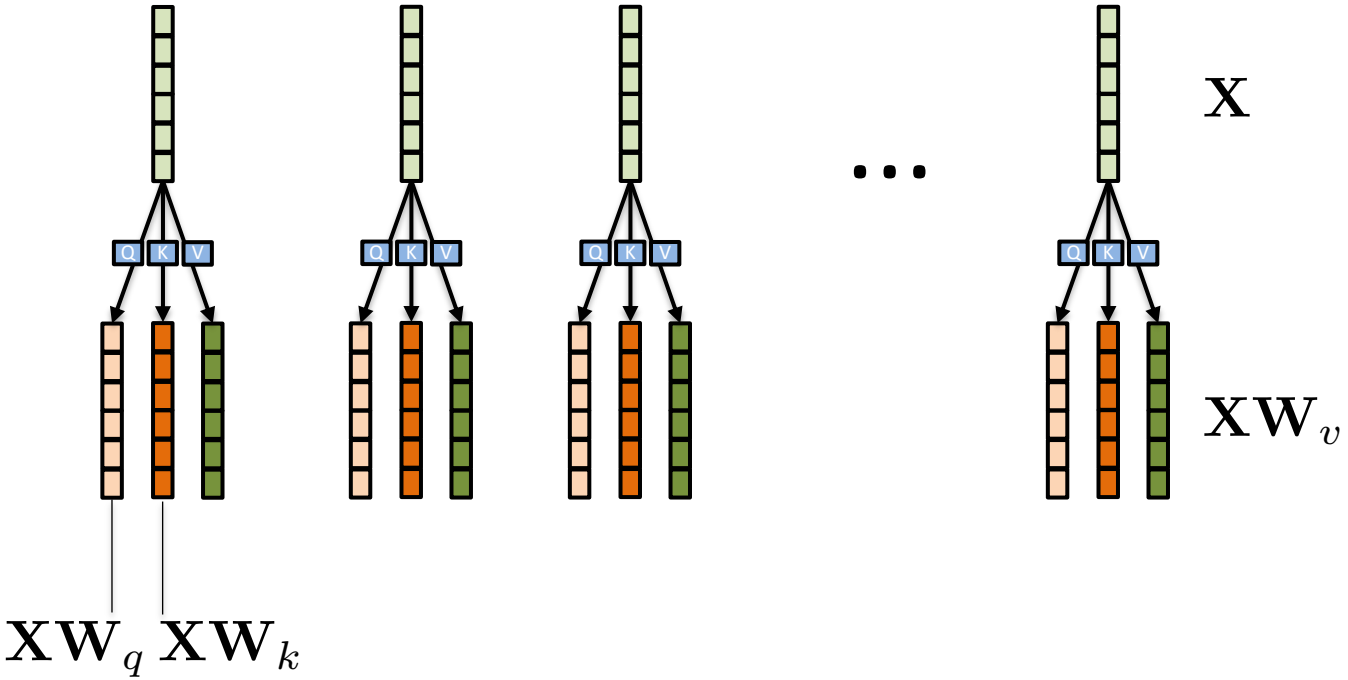
quick brown fox

Transformers: attention is all you need

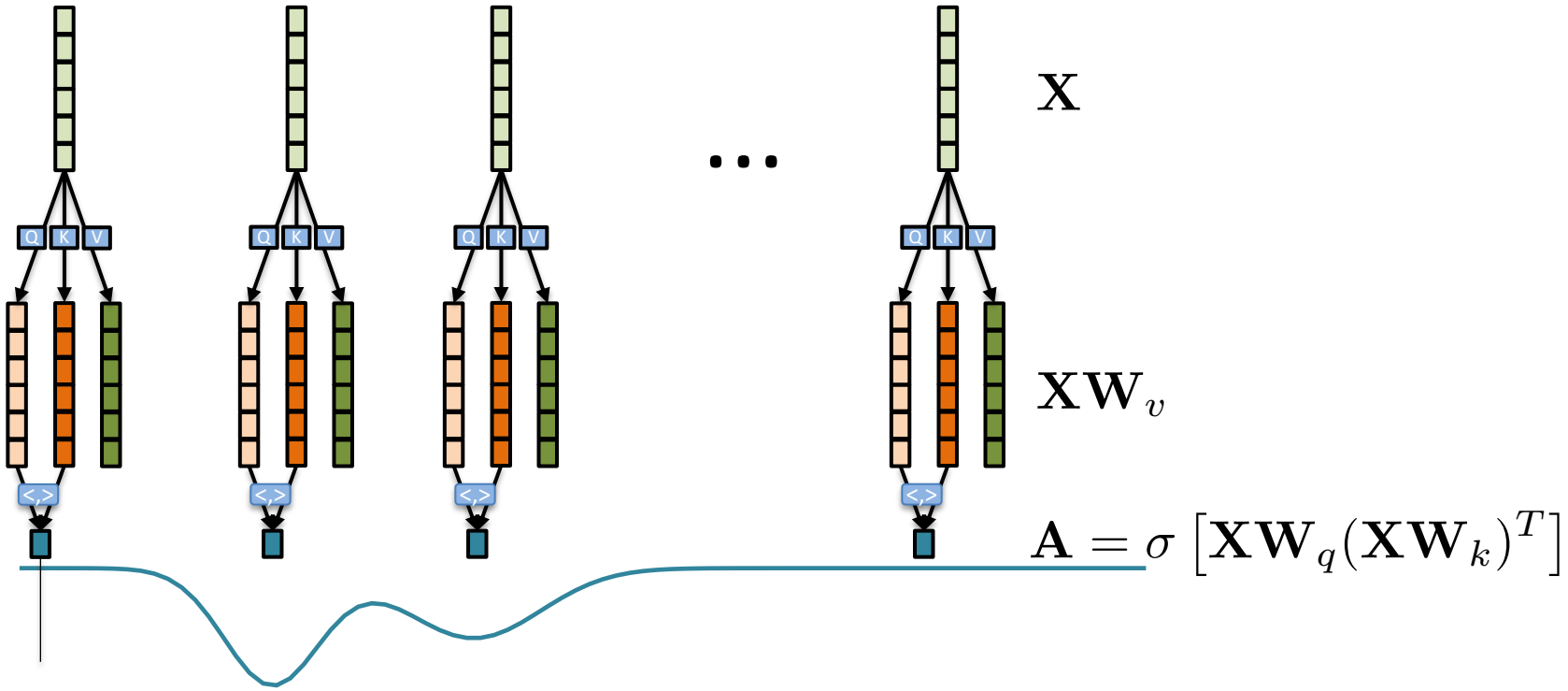


We suppose unordered data (a set) $\mathbf{X} = \{\mathbf{x}_i\}$. Each input item is "transformed" by a linear transform weighted by attention.

Transformers: attention is all you need



Transformers: attention is all you need



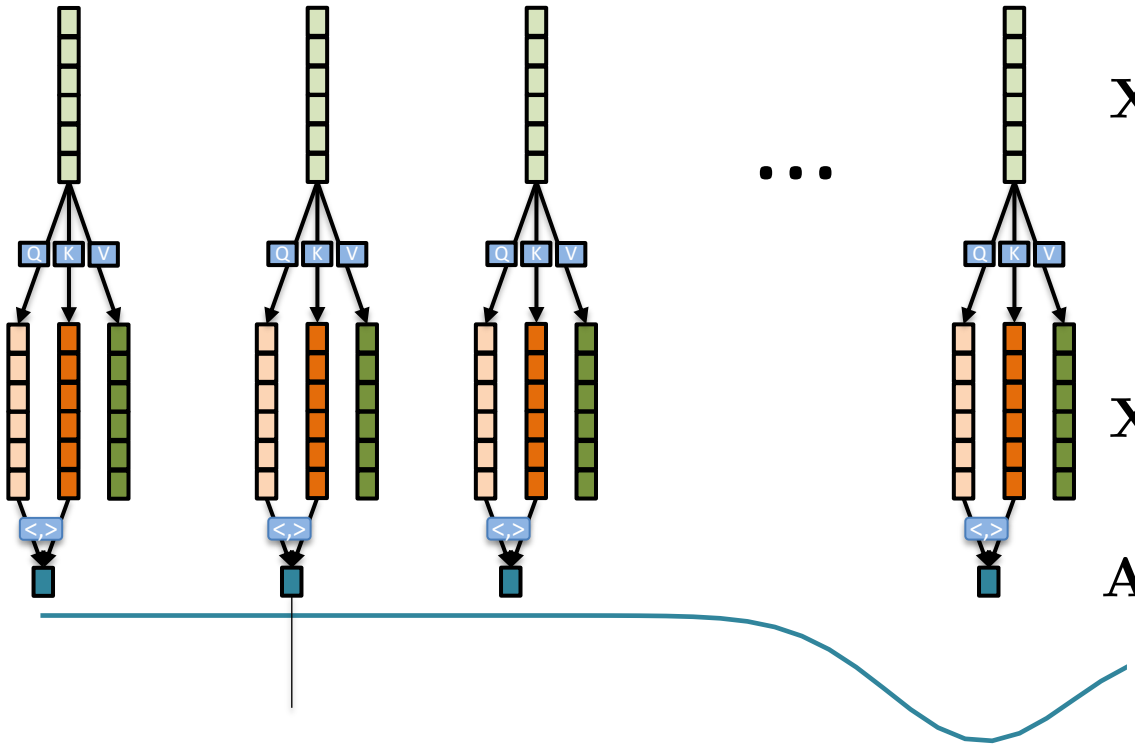
$\mathbf{X}$

$\mathbf{X}\mathbf{W}_v$

$\mathbf{A} = \sigma\left[\mathbf{X}\mathbf{W}_q(\mathbf{X}\mathbf{W}_k)^T\right]$

Each item has its own distribution,
associated with its query vector!

Transformers: attention is all you need



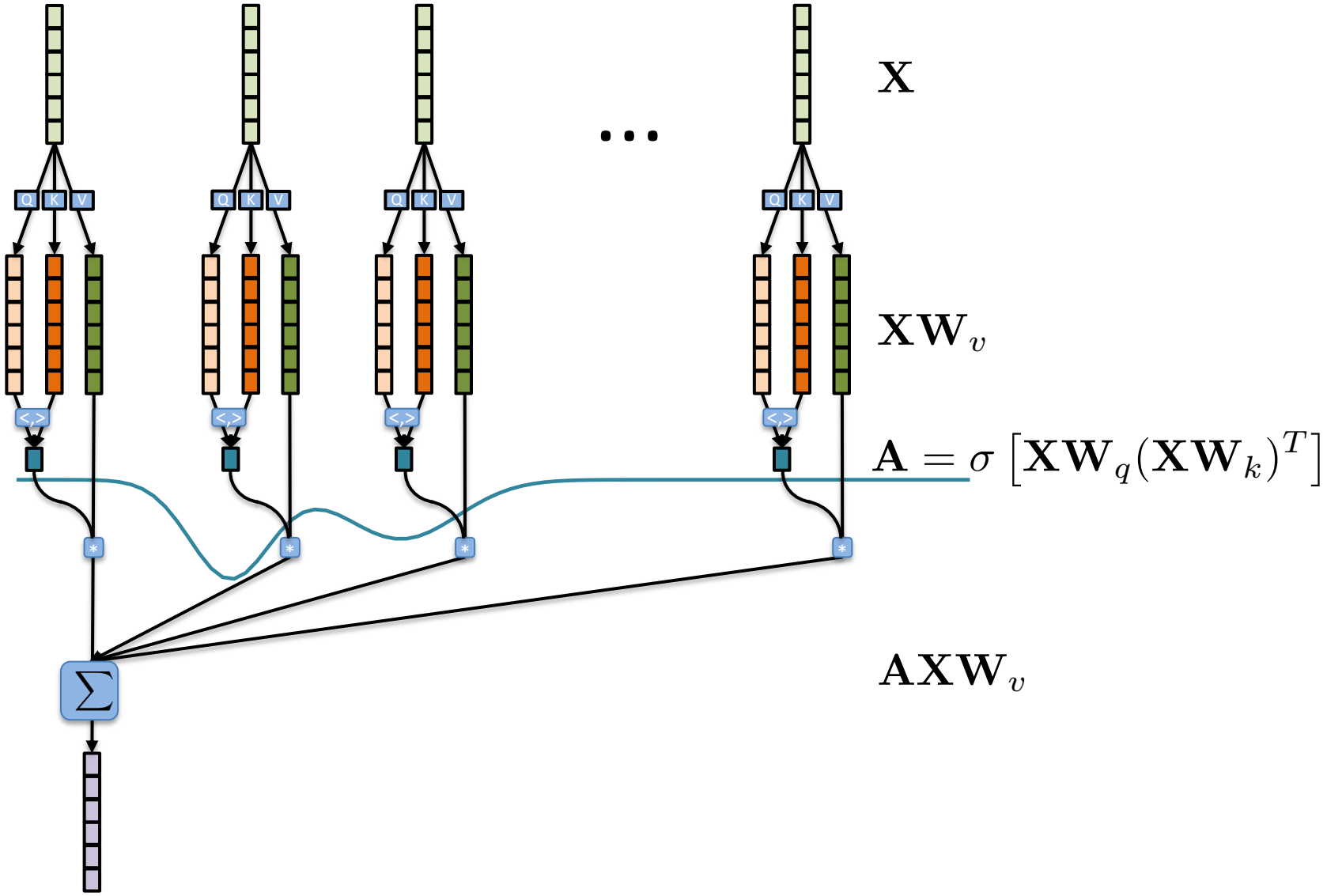
$\mathbf{X}$

$\mathbf{XW}_v$

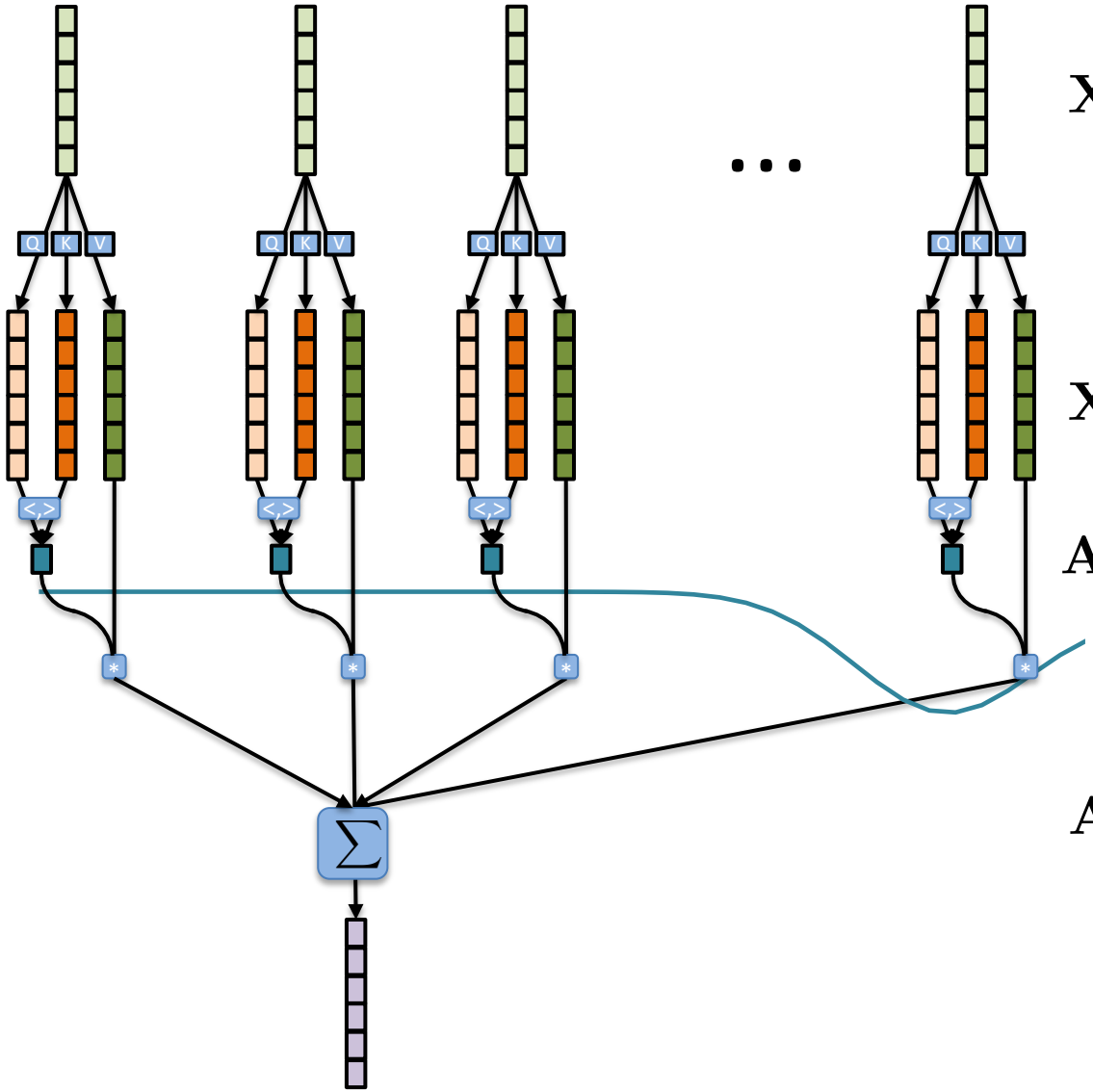
$\mathbf{A} = \sigma\left[\mathbf{XW}_q(\mathbf{XW}_k)^T\right]$

Each item has its own distribution,
associated with its query vector!

Transformers: attention is all you need



Transformers: attention is all you need



$\mathbf{X}$

$\mathbf{X}\mathbf{W}_v$

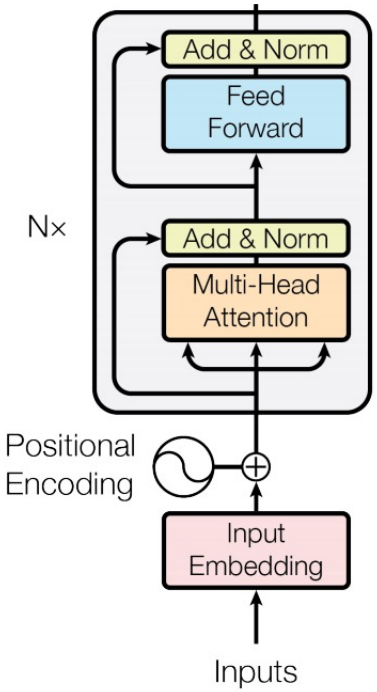
$\mathbf{A} = \sigma\left[\mathbf{X}\mathbf{W}_q(\mathbf{X}\mathbf{W}_k)^T\right]$

$\mathbf{A}\mathbf{X}\mathbf{W}_v$

Integration ...

Toutes les familles heureuses se ressemblent. Chaque famille malheureuse, au contraire, l'est à sa façon.

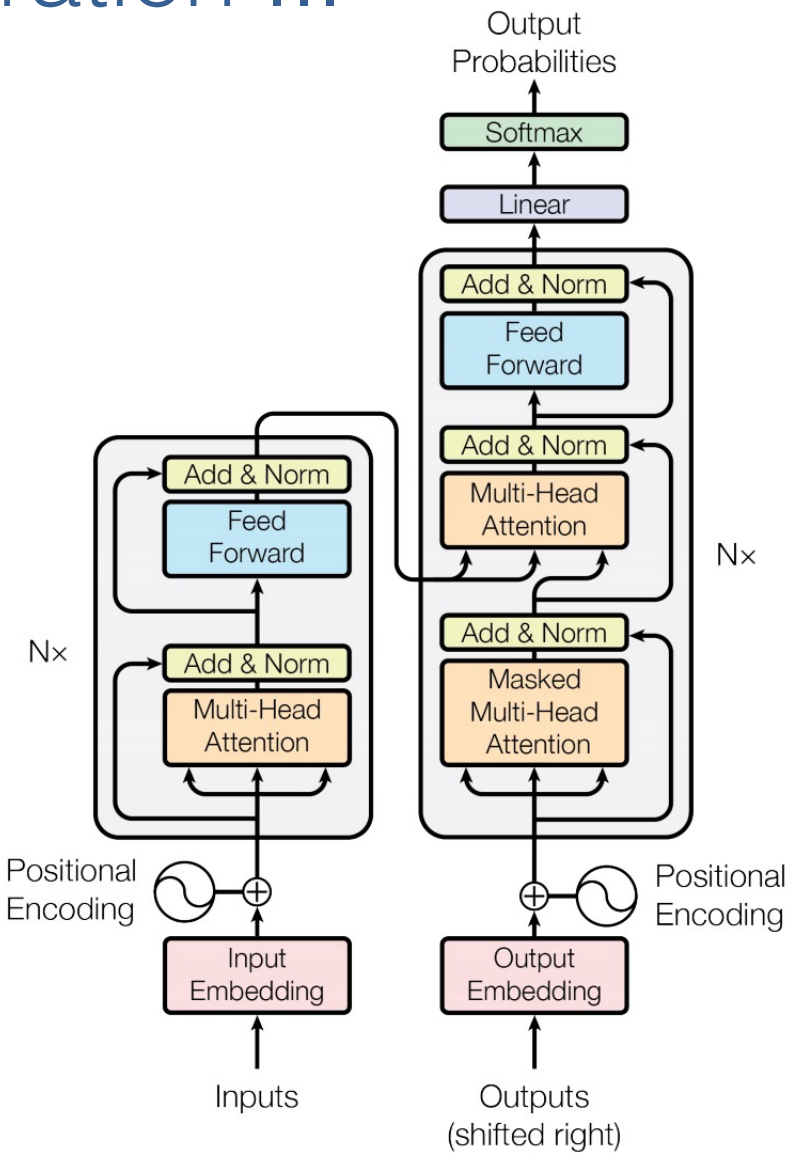
Happy families are all alike. Every unhappy family is unhappy in its own way.



Integration ...

Toutes les familles heureuses se ressemblent. Chaque famille malheureuse, au contraire, l'est à sa façon.

Happy families are all alike. Every unhappy family is unhappy in its own way.

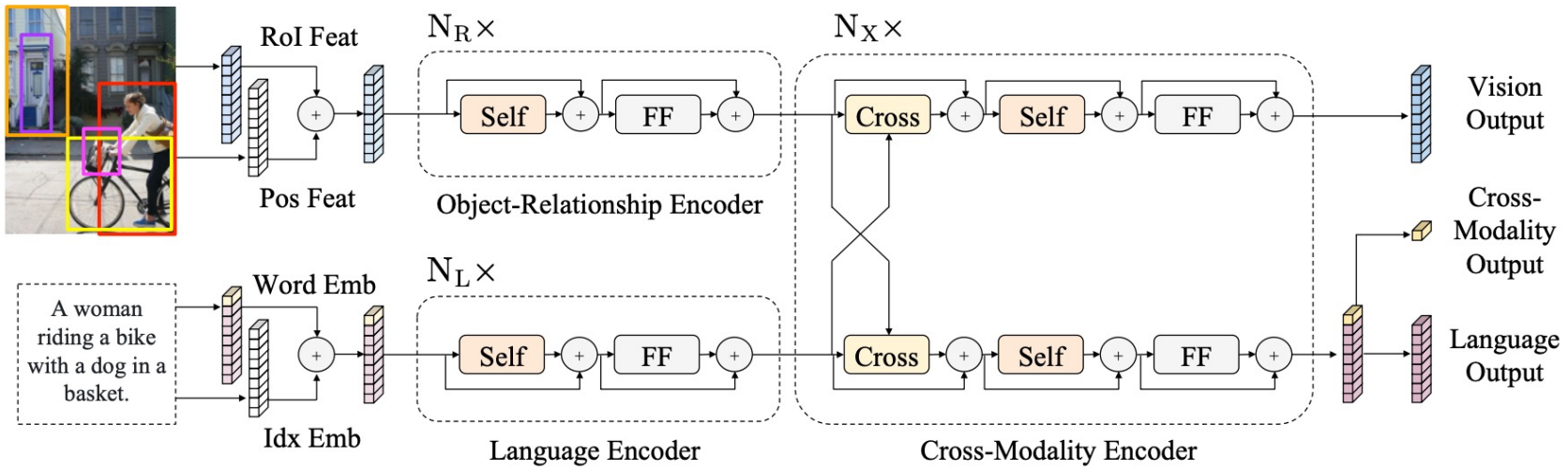


A concrete large-scale application

Vision and Language Reasoning

LXMERT

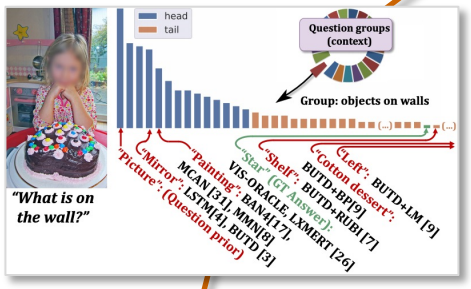
A vision and language encoder with self-attention and cross-attention.



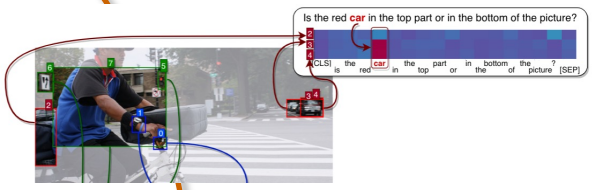
[Tan et Bansal, EMNLP 2019]

VQA

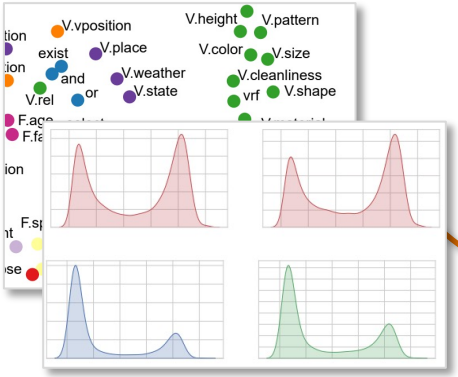
How can we evaluate biases in learning? (CVPR 2021a)



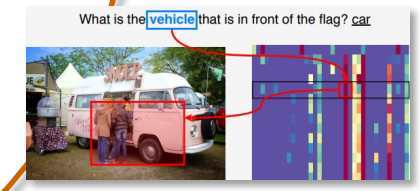
Can we ground object detection through language (in preparation)



How can we visualize and transfer reasoning? (CVPR 2021b)



Can we weakly supervise word-object alignment? (ECAI 2020)

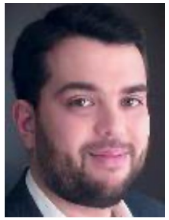


Can we supervise reasoning programs? (arxiv, under review)

orange™


Corentin Kervadec

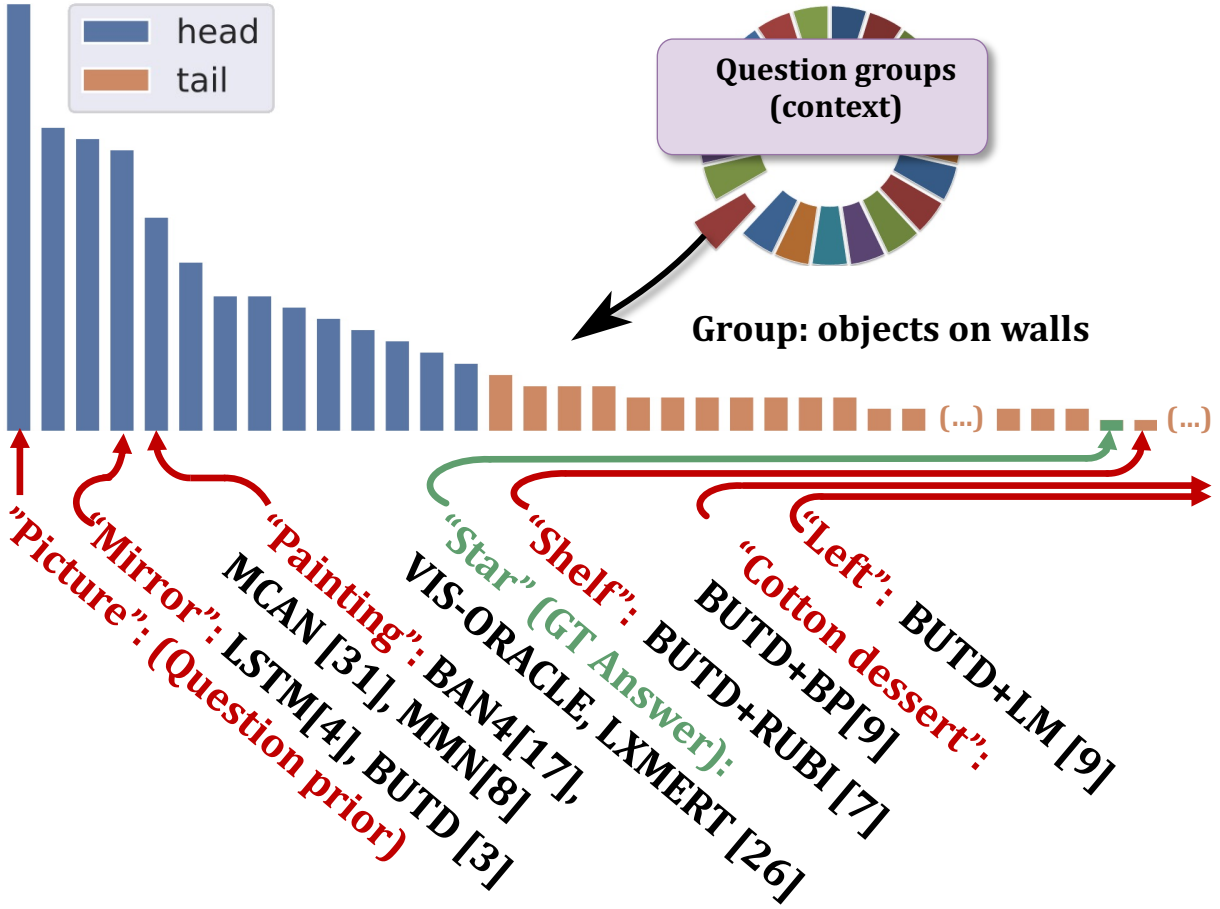

Grigory Antipov


Moez Baccouche


Christian Wolf

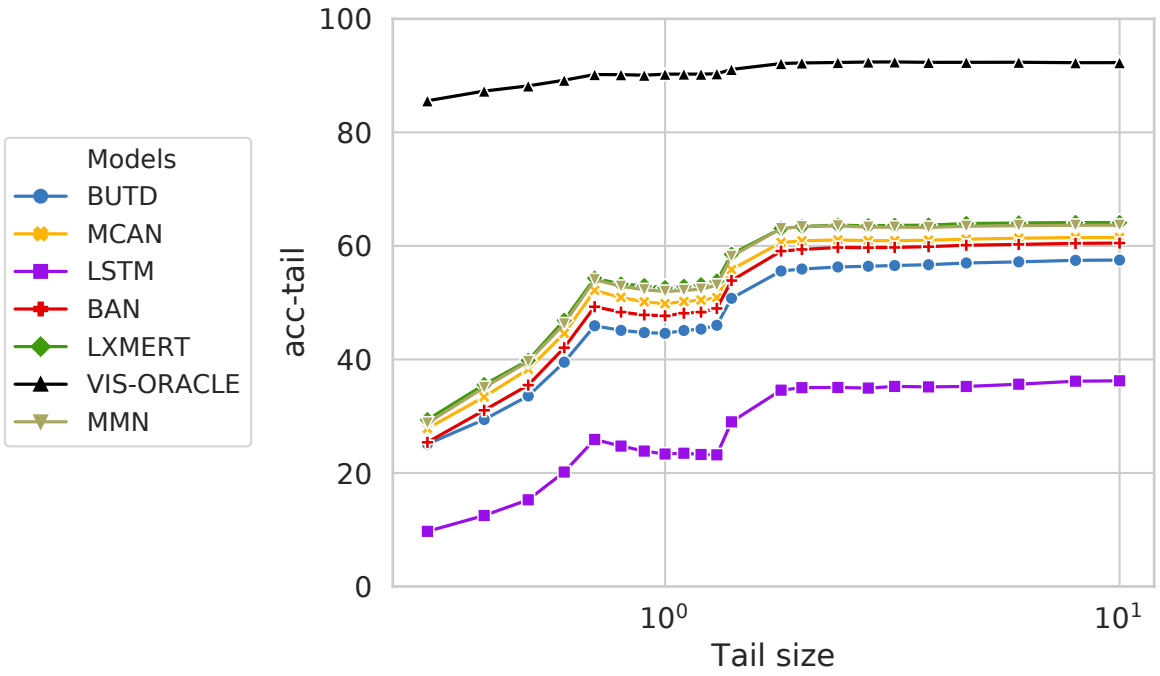
Roses are Red, Violets are blue …
But should VQA expect them to?

"What is on the wall?"



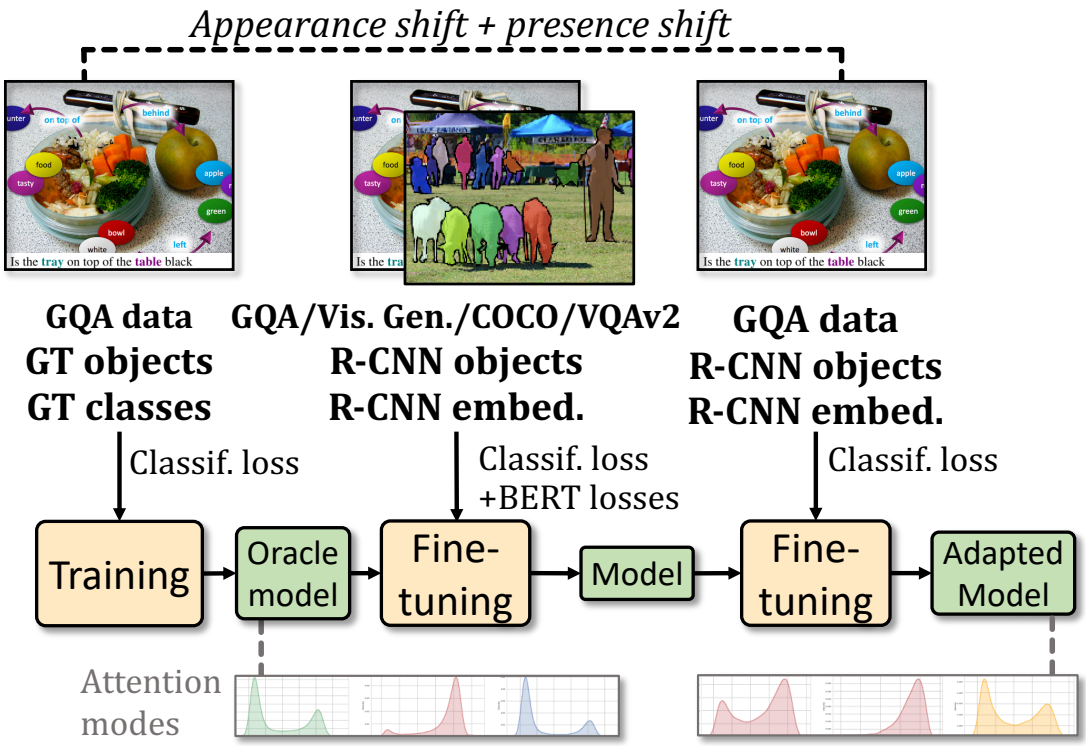
[Kervadec, Antipov, Baccouche, Wolf, CVPR 2001a]

Reasoning vs. bias exploitation



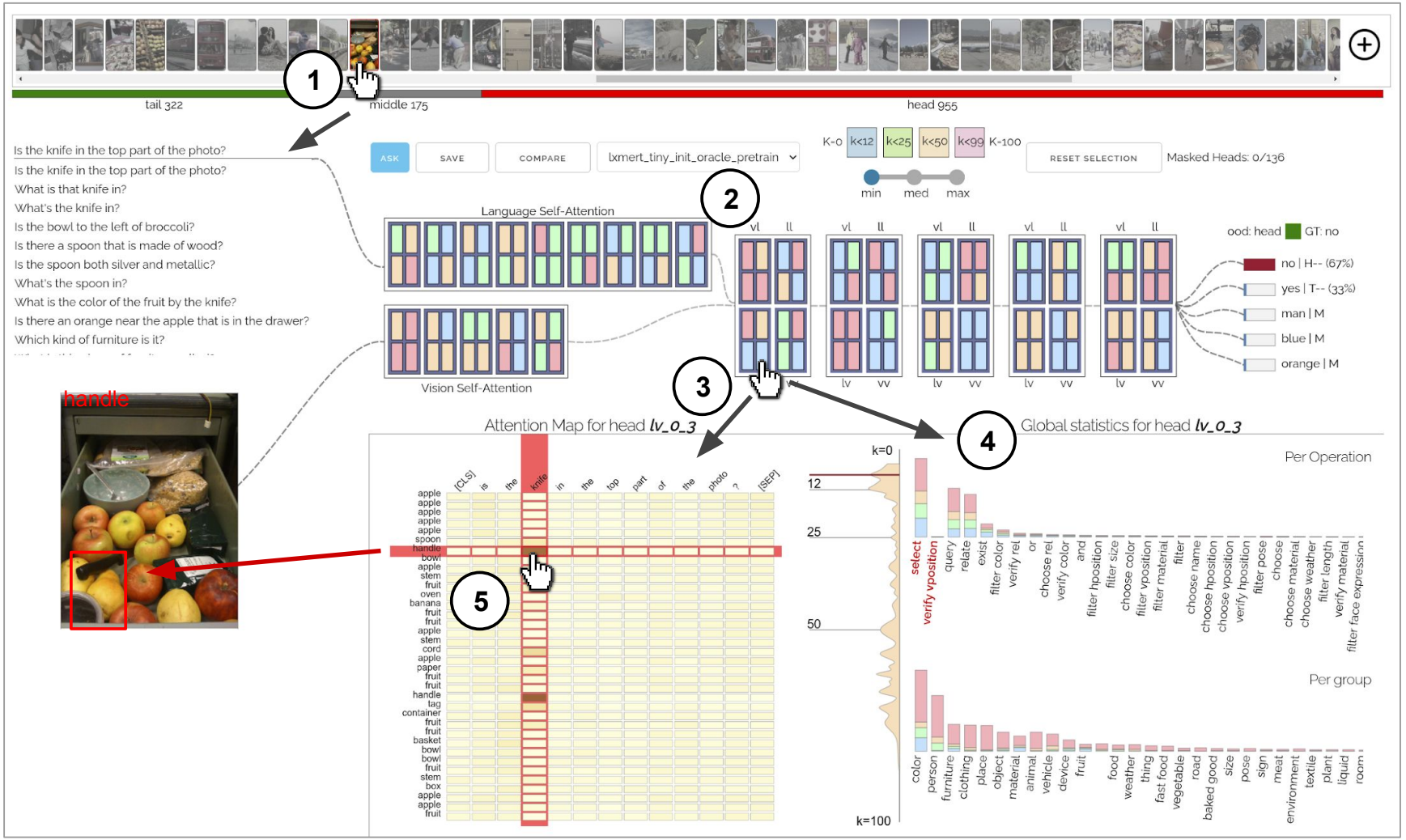
[Kervadec, Antipov, Baccouche, Wolf, CVPR 2021a]

Oracle transfer



[Kervadec, Antipov, Baccouche, Wolf, CVPR 2021b]

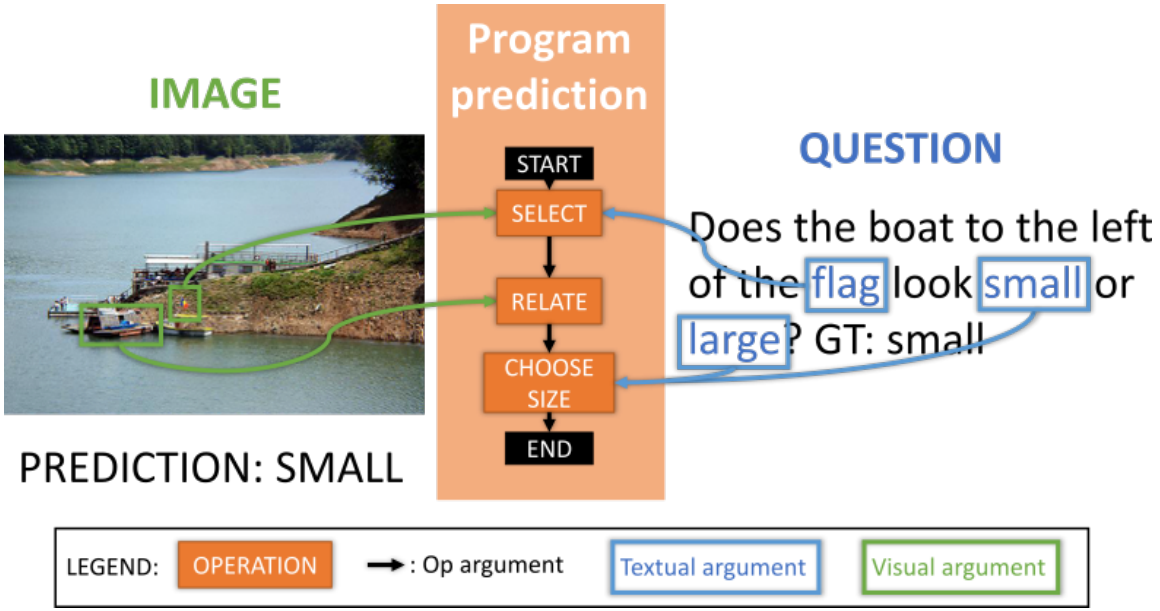
Visualizing Reasoning Patterns



[Jaunet, Kervadec, Vuillemot, Antipov, Baccouche, Wolf, Conditionally accepted to VIS 2021]

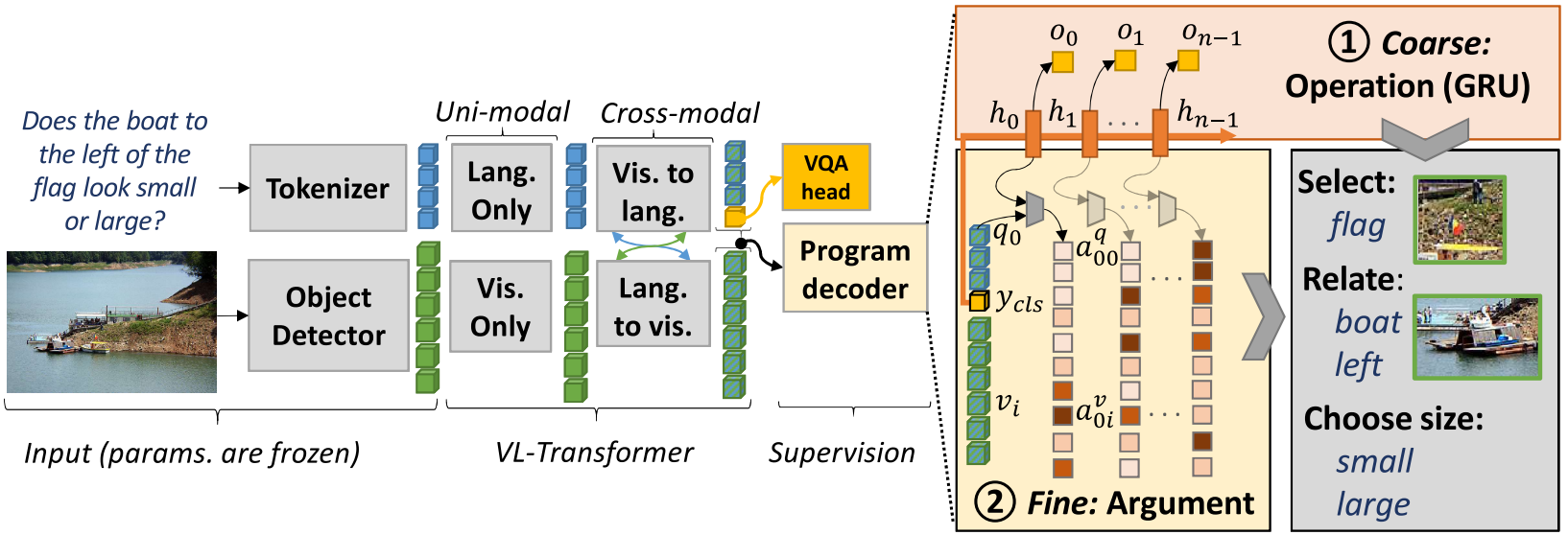
https://visqa.liris.cnrs.fr

GT reasoning programs



[Kervadec, Antipov, Baccouche, Wolf, arxiv / under review]

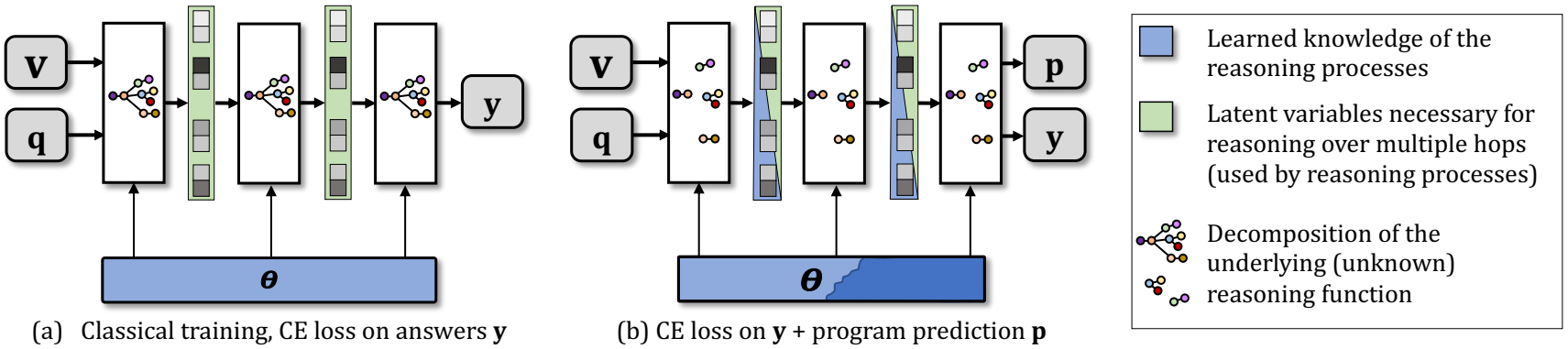
Program prediction



[Kervadec, Antipov, Baccouche, Wolf, arxiv / under review]

Program supervision

We theoretically show through PAC-bounds that program supervision as auxiliary loss can decrease sample complexity under some hypotheses.



(a) Classical training, CE loss on answers **y**

(b) CE loss on **y** + program prediction **p**

[Kervadec, Antipov, Baccouche, Wolf, arxiv / under review]

Oracle + program supervision

	Model	Oracle transf.	Prog. sup.	GQA-OOD [20] acc-tail	GQA-OOD [20] acc-head	GQA [17] test-dev	GQA [17] binary*	GQA [17] open*	GQA [17] test-std	AUC† prog.
scratch	(a) Baseline			42.9	49.5	52.4	-	-	-	/
	(b) Oracle transfer	✓		48.2±0.3	54.6±1.1	57.0±0.3	74.5	42.1	57.3	/
	(c) Ours	✓	✓	**48.8**±0.1	**56.1**±0.3	**57.8**±0.2	**75.4**	**43.0**	**58.2**	97.1
+ lxmert	(d) Baseline			47.5	55.2	58.5	-	-	-	/
	(e) Oracle transfer	✓		47.1	54.8	58.4	77.1	42.6	58.8	/
	(f) **Ours**	✓	✓	**48.0**±0.6	**56.6**±0.6	**59.3**±0.3	**77.3**	**44.1**	**59.7**	96.4

Table 1: Impact of program supervision on *Oracle transfer* [23] for vision-language transformers. LXMERT [36] pre-training is done on the GQA unbalanced training set. We report scores on GQA [17] (*test-dev* and *test-std*) and GQA-OOD (*test*). * binary and open scores are computed on the test-std; † we evaluate visual argument prediction by computing AUC@0.66 on GQA-val.

[Kervadec, Antipov, Baccouche, Wolf, arxiv / under review]