

Supplementary material

A. Implementation of underlying VQA model

The VQA model within our method follows the general description of Teney *et al.* [33] as illustrated in Fig. 3 in the main paper. One exception is in the question encoding, where we replace their gated recurrent unit (GRU) with a bag of words, *i.e.* a simple average of word embeddings. The first reason is computational, to avoid the relatively slow evaluation of the unrolled GRU. The second reason is that we encountered instabilities in the training of the adaptation method with the GRU. We suspect this to be due to our first-order approximation of the MAML algorithm.

Most implementation details follow [33]. In particular, the non-linear operations in the network use gated hyperbolic tangent units. We use the “bottom-up attention” features [3] of size 36×2048 , pre-extracted and provided by Anderson *et al.*³ The word embeddings are initialized as GloVe vectors [26] of dimension 300, then optimized with the same learning rate as other weights of the network. All activations except the word embeddings and their average are of dimension 256. The answer candidates are those appearing at least 20 times in the VQA v2 training set, *i.e.* a set of about 2000 answers. The output of the network is passed through a logistic function to produce scores in $[0, 1]$. The final classifier is trained from a random initialization, rather than the visual and text embeddings of [33]. In our ablative and in-depth experiments (Table 1, Fig. 4, and Fig. 6), we use a slightly simplified model, where the “top-down” attention map over the image is uniform. The image features of size 36×2048 are thus averaged uniformly to a vector of size 1×2048 . This significantly reduces the cost of training and evaluating the model since these averages can be precomputed and fit in memory for the whole dataset. The relevance function r_3 (Section 3.4) also uses these global image features.

B. Implementation of adaptation algorithm

We use the AdaDelta algorithm [45] to train the model’s weights (θ_0 and those of the gradient projection) with back-propagation from the loss \mathcal{L}_M . Following this practice, we also found it beneficial to replace the gradient descent step of the adaptation (Eq. 1 and 4) with the AdaDelta weight update (see details in [45]). This effectively determines the size of the gradient step α automatically based on a rolling average of the weights’ and gradients’ magnitudes. This makes the weight updates much more stable, and it eliminates the hyperparameter α .

The gradient projection $g_\psi(\cdot)$ implemented as a simple linear scaling, with no biases, and no cross-talk across di-

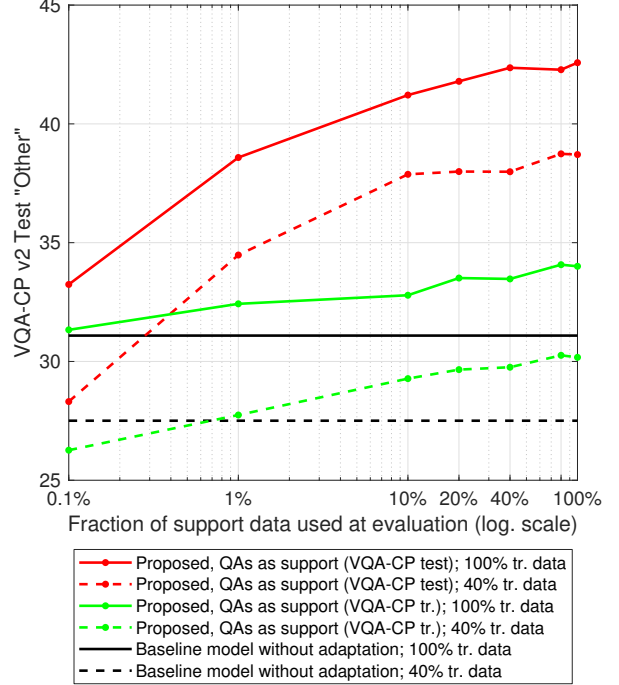


Figure 6. Varying the amount of support data used during evaluation.

mensions. For example, to adapt a linear layer that uses weights $W \in \mathbb{R}^{256 \times 256}$, the gradient $\nabla_W \mathcal{L}_A$ is transformed with

$$g_{\psi_M}(\nabla_W \mathcal{L}_A) = \psi_M \circ \nabla_W \mathcal{L}_A \quad (6)$$

where $\psi_M \in \mathbb{R}^{256 \times 256}$ represents the parameters of the projection and \circ the Hadamard (element-wise) product.

The adaptation algorithm uses a number $T=3$ updates during training and evaluation. This value was selected in 1–5 by cross-validation.

The whole method is trained with mini-batches of size 128. The evaluation also uses mini-batches (of the same size) in a transductive manner, *i.e.* sharing information across multiple test instances, as done in existing implementations of MAML [12, 24]. This means that the adaptation algorithm effectively uses support data retrieved for 128 questions at a time. The primary reason for mini-batches during evaluation is computational, but we did not observe improvements in accuracy with smaller batch sizes (down to processing one single instance at a time), whether for training and/or evaluation.

C. Additional experiments

C.1. Varying the amount of support data

We performed additional experiments in which we varied the amount of support data available during the evaluation

³<https://github.com/peteanderson80/bottom-up-attention>

of the model (Fig. 6). This serves to verify that the model makes actual use of information from the support data. We indeed observe that the performance increases as more data is made available. We repeated the experiment with a model initially trained with only 40% of the data (dashed lines in Fig. 6). The trend of the accuracy versus the amount of support data remains similar. The overall performance is however lower. This indicates room for improvement for the adaptation algorithm. Ideally, a model trained with less data should approach the performance of a model trained with more data, when provided with this data (as support) at test time.

C.2. Generalization to support from a different distribution

We evaluated the proposed model by providing it with support data from a different distribution than the data it is originally trained with (Tables 3–4). For these experiments, we use the VQA-CP in a “leave-one-out” setting: we use the test set itself as the support data, and masking the intersection of the support data with a test instance currently evaluated. More precisely, all QAs relating to the same image as the current test question are left out of the utilized support. The results of this experiment show that the model can very effectively adapt to this novel support data, as the accuracy gets a significant jump, approaching the performance of the validation set (which is of the same distribution as the initial training data). We suspected the increase in performance might be simply due to the larger amount of data (the original training data plus the additional test set provided as support). We disproved this hypothesis by repeating the experiment with a model trained with less initial training data and less support data, such as to match the same total amount of data provided to the baseline (details in the supplementary material). This experiment gave a similarly high accuracy, which demonstrates that the model is indeed capable of adapting on-the-fly to the provided support data, even when it significantly differs from the data it was originally trained with.

D. Qualitative results





























We provide additional qualitative results in the following pages. A first set of results uses support data made of QAs. A second set uses support data made of captioned images (as indicated in column headings).

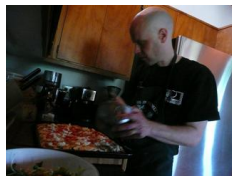
VQA-CP v2 Test split, “Other” questions		
Ours with adaptation and, as support data:	QAs Tr.	QAs Test
Uniform sampling $r=r_0$	31.33	32.83
Relevance function $r=r_1$	31.79	37.19
Relevance function $r=r_2$	31.76	36.28
Relevance function $r=r_3$	31.68	33.52
Relevance function $r=r_2r_3$	31.09	37.78
Relevance function $r=r_1r_2r_3$	34.25	43.52

Table 3. Complement to Table 1. We evaluate the different versions of our model with, as support data, QAs from the training set (first column, identical to Table 1) and QAs from the test set (second column, in a leave-one-out protocol). These results are not comparable to competing models since they use more data, but the clear improvement in the second column demonstrates that the model clearly adapts to support data from a distribution different from the one it was trained with (since the support QAs now reflect the distribution of the test questions). We envision this capability to allow a pretrained VQA model to be applied to various domains by simply providing it, at test time, with domain-specific support data.

	VQA-CP v2 Test split			
	Overall	Yes/no	Numbers	Other
Ours with adaptation and, as support data:				
QAs (VQA-CP tr.), $r=r_1r_2r_3$	46.00	58.24	29.49	44.33
QAs (VQA-CP test), $r=r_1r_2r_3$	52.09	62.02	47.66	48.21

Table 4. Complement to Table 2 (first row is identical to Table 2). This demonstrates the same effect as explained for Table 3.

Input question	Samples of retrieved support data (QAs)			Predicted scores
 <p>What season might this be ? Correct answer: winter.</p>	 <p>What season does it appear to be ? fall.</p>	 <p>What season might this be ? summer.</p>	 <p>What season is it ? spring.</p>	<p>Without adaptation:</p> <ul style="list-style-type: none"> ■ winter ■ fall ■ spring ■ summer ■ snow <p>After adaptation:</p> <ul style="list-style-type: none"> ■ winter ■ fall ■ summer ■ spring ■ snow
	 <p>What season is it ? fall.</p>	 <p>What season is it ? fall.</p>	 <p>What season is this ? fall.</p>	
 <p>What is teddy bear made of ? Correct answer: fur.</p>	 <p>What is bear standing on ? concrete.</p>	 <p>What is bear made out of ? concrete.</p>	 <p>What material is bear made of ? cloth.</p>	<p>Without adaptation:</p> <ul style="list-style-type: none"> ■ fur ■ fabric ■ cloth ■ paper ■ concrete <p>After adaptation:</p> <ul style="list-style-type: none"> ■ cotton ■ teddy bear ■ fabric ■ fur ■ none
	 <p>What is on bear is face ? fur.</p>	 <p>What is behind bear ? concrete.</p>	 <p>What material is polar bear walking on ? concrete.</p>	
 <p>What sport is being played ? Correct answer: baseball.</p>	 <p>What sport are these kids getting ready to play ? baseball.</p>	 <p>What sport are people playing ? baseball.</p>	 <p>What sport are they playing ? baseball.</p>	<p>Without adaptation:</p> <ul style="list-style-type: none"> ■ baseball ■ softball ■ yes ■ playing baseball ■ baseball bat <p>After adaptation:</p> <ul style="list-style-type: none"> ■ baseball ■ baseball field ■ softball ■ soccer ■ tennis
	 <p>What sport are these guys playing ? baseball.</p>	 <p>What sport are they playing ? baseball.</p>	 <p>What sport are they playing ? baseball.</p>	
 <p>What is this man doing ? Correct answer: painting.</p>	 <p>What sport is man doing ? fishing.</p>	 <p>What is hanging behind man ? painting.</p>	 <p>What is man doing ? painting.</p>	<p>Without adaptation:</p> <ul style="list-style-type: none"> ■ fishing ■ standing ■ boating ■ walking ■ painting <p>After adaptation:</p> <ul style="list-style-type: none"> ■ standing ■ fishing ■ surfing ■ walking ■ boating
	 <p>What is man doing ? standing.</p>	 <p>What is man doing ? standing.</p>	 <p>What is this man doing ? walking.</p>	



What side dish appears in bowl ?
Correct answer: salad.



What is in bowl ? soup.



What is in bowl ? soup.



What is on dish ? soup.



What is in bowl ? soup.



What is in black bowl ? soup.



What is in bowl ? soup.

Without adaptation:

- ☐ pizza
- ☐ none
- ☐ soup
- ☐ salad
- ☐ vegetables

After adaptation:

- ☐ soup
- ☐ salad
- ☐ tomatoes
- ☐ beans
- ☐ none



What kind of flower is white one ?
Correct answer: lily.



What are species of flower represented in this photo ? rose.



What kind of flower is shown ? rose.



What is white plant called ? lily.



What kind of plant is this ? lily.



What is name of flower in vase ? rose.



What type of flower is in vase ? lily.

Without adaptation:

- ☐ tulip
- ☐ lily
- ☐ tulips
- ☐ lilies
- ☐ rose

After adaptation:

- ☐ lily
- ☐ tulip
- ☐ tulips
- ☐ lilies
- ☐ rose



Are bags hard or soft ?
Correct answer: hard.



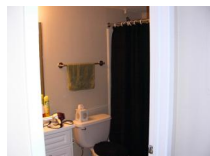
Does sand on beach look soft or coarse ? soft.



Is this ground hard or soft ? soft.



How hard did woman hit ball ? soft.



Is it better to use soft or natural lighting in bathroom ? soft.



Is it better to use soft or natural lighting in bathroom ? soft.



Is chaise lounge in foreground more likely soft or firm ? soft.

Without adaptation:

- ☐ free
- ☐ full
- ☐ soft
- ☐ laptops
- ☐ open

After adaptation:

- ☐ soft
- ☐ clean
- ☐ sunny
- ☐ cold
- ☐ warm



What is this man is name ?
Correct answer: unknown.



Why is man on left sleepy ? unknown.



Is this person man or woman ? man.



What sign is near man ? unknown.



Is it man or woman with car ? man.



What street is man on ? unknown.



What color is man is boxers ? unknown.

Without adaptation:

- ☐ unknown
- ☐ man
- ☐ obama
- ☐ not possible
- ☐ don't know

After adaptation:

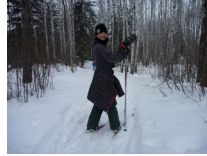
- ☐ unknown
- ☐ none
- ☐ don't know
- ☐ bob
- ☐ nothing



What is strapped to his waist ?
Correct answer: backpack.



What is this man's feet strapped to ? snowboard.



What is tied around their waist ? coat.



What does child have around its waist ? belt.



What is person wearing around his waist ? belt.



What does child have around its waist ? belt.



What is tied around woman's waist ? coat.

Without adaptation:

- leash
- snowboard
- boots
- coat
- belt

After adaptation:

- jacket
- coat
- sweater
- backpack
- dog



What kind of kite is man flying ?
Correct answer: white.



What is flying in air ? kite.



What is flying ? kite.



What is flying ? kite.



What pattern are kites flying in ? none.



What is moving man ? kite.



How is man staying in air ? wind.

Without adaptation:

- sail
- none
- kite
- white
- wind

After adaptation:

- kite
- white
- none
- seagull
- no



Will that fence contain this animal ?
Correct answer: yes.



Is there more than one animal shown ? no.



Does this animal appear to live in zoo ? yes.



Is animal alive ? yes.



Is fence as high as animal when it is standing up ? no.



Is person scared of animal ? yes.



Is this wild animal ? no.

Without adaptation:

- yes
- no
- unknown
- 2
- none

After adaptation:

- elephant
- yes
- no
- elephants
- trunk



What is he holding in his hands ?
Correct answer: pen.



What is person holding ? laptop.



What is man holding ? laptop.



What is man holding on his lap ? laptop.



What is he holding in his hands ? mouse.



What is she holding in her left hand ? laptop.



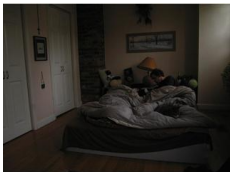


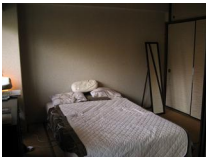

















What is woman holding on her lap ? computer.

Without adaptation:

- computer
- laptop
- mouse
- books
- nothing

After adaptation:

- laptop
- computer
- keyboard
- nothing
- mouse

Input question	Samples of retrieved support data (captions)			Predicted scores
 <p>What color is comforter ? Correct answer: white.</p>	 <p>Large bed covered with a comforter in a bedroom.</p>	 <p>Bedroom shot size bed, white comforter and a lamp.</p>	 <p>Bed with a white comforter.</p>	<p>Without adaptation:</p> <ul style="list-style-type: none"> gray brown black white green
	 <p>Bed with comforter turned down.</p>	 <p>Bed with a white pillow, a white comforter and accessories.</p>	 <p>Bed with comforter turned down and a night table lamp.</p>	<p>After adaptation:</p> <ul style="list-style-type: none"> white black gray brown blue
 <p>What is horse in background doing ? Correct answer: eating.</p>	 <p>Adult black horse and young brown horse interacting.</p>	 <p>Horses in a grassy field with trees in the background.</p>	 <p>Horse running in a grassy field in an enclosed area.</p>	<p>Without adaptation:</p> <ul style="list-style-type: none"> standing grazing walking looking running
	 <p>Horse eating a hay stack.</p>	 <p>Brown horse standing on dirt in a grass field.</p>	 <p>A giraffe in the forefront and a zebra in the background.</p>	<p>After adaptation:</p> <ul style="list-style-type: none"> grazing running standing eating walking
 <p>What is he wearing ? Correct answer: suit.</p>	 <p>Teenager wearing glasses and a tie.</p>	 <p>Man wearing a shirt and a tie making a creepy face.</p>	 <p>Man wearing a black hat and holding an umbrella.</p>	<p>Without adaptation:</p> <ul style="list-style-type: none"> hat fedora jacket cowboy coat
	 <p>Man wearing a vest, a tie, and glasses.</p>	 <p>Bald man with mustache wearing a suit.</p>	 <p>Man standing in a bathroom wearing a shirt.</p>	<p>After adaptation:</p> <ul style="list-style-type: none"> tie suit hat ties clothes



What color are umbrellas ?
Correct answer: green.



A group of people at a metal table with umbrellas.



People enjoying a meal with wine under white umbrellas.



Elderly women stand in a large room with colorful umbrellas.

Without adaptation:

- white
- purple
- green
- yellow
- blue



Crowd of adults holding red umbrellas in a march.



Adult and child holding umbrellas in a park.



Group of people walking with red umbrellas.

After adaptation:

- green
- blue
- black
- orange
- white



What color is nose of plane ?
Correct answer: red.



Older air plane parked under a bridge.



Red, yellow, blue, and white plane parked on concrete.



Big blue air plane parked with people.

Without adaptation:

- red
- white
- black
- gray
- pink



Plane sitting on a runway at an airport.



Air force plane sitting on tarmac with propellers.



White plane sitting on a runway.

After adaptation:

- white
- red
- black
- gray
- silver



What color is tablecloth ?
Correct answer: green and red.



A bowl of broccoli and pasta sit on a checkered tablecloth.



Half-eaten food and beer on a patterned tablecloth.



Plates of food on a red tablecloth.



Colorful plate of appetizers on a white linen tablecloth.



Sandwich for halloween on a tablecloth covered table.



Restaurant sandwich platter on a plaid tablecloth.

Without adaptation:

- plaid
- red and white
- green
- checkered
- green and white

After adaptation:

- red and white
- checkered
- plaid
- green
- black and white



Are all of these people friends ?
Correct answer: yes.



People riding skateboards down the street.



Large group of smiling people raising their hands.



Group people riding skis on snow.

Without adaptation:

- yes
- no
- unknown
- family
- can't tell



Bunch of people with skis ride on snow.



People gathered in front of a government building flying kites.



Group of people skiing on snow.

After adaptation:

- lot
- many
- 100
- all
- 50



What utensil is in girl is hand ?
Correct answer: fork.



Girl pulling up a spoonful of cheesy casserole stands.



Girl standing at the kitchen counter holding a spoon.



Woman and girl with plates of cakes and rolls.



Boy and girl sitting at a dinner table and both pointing.

Without adaptation:

- ☐ fork
- ☐ spoon
- ☐ knife
- ☐ right
- ☐ fork and knife

After adaptation:

- ☐ pizza
- ☐ knife
- ☐ fork
- ☐ fork and knife
- ☐ plate



What is child running on top of ?
Correct answer: leaves.



Man and child flying a kite in a field.



Dog running in a park with a frisbee in his mouth.



Small dog running up truck.



Small child on a street with a stop sign.



Person in black uniform running with a soccer ball.



Children running and playing with kites in a park area.

Without adaptation:

- ☐ umbrella
- ☐ nothing
- ☐ grass
- ☐ ground
- ☐ rain

After adaptation:

- ☐ leaves
- ☐ grass
- ☐ frisbee
- ☐ ground
- ☐ umbrella



Why do majority of people have on same color ?
Correct answer: blue.



Couple of people standing with ski on snow.



Group of people standing on skis.



Man stands with child wearing skis and people sitting.



Group of people on snow with skis.



Group of people skiing down a snow covered slope.



Many people with ski on a mountain dressed for ski.

Without adaptation:

- ☐ 0
- ☐ no
- ☐ racing
- ☐ 1
- ☐ yes

After adaptation:

- ☐ skiing
- ☐ blue
- ☐ white
- ☐ yellow
- ☐ safety



Does this cake have healthy element ?
Correct answer: no.



Group of children standing at a table eating cake.



Red cake with white frosting displayed with vase and sunflowers.



Bride and groom cutting a wedding cake.



Large white multi layered cake sitting on a table.



Table decorated with flowers, utensils, and a marriage cake.



Birthday cake sitting on a kitchen counter.

Without adaptation:

- ☐ yes
- ☐ no
- ☐ n
- ☐ flowers
- ☐ don't know

After adaptation:

- ☐ wedding
- ☐ yes
- ☐ no
- ☐ fruit
- ☐ none