Appendix: Combining Knowledge and Reasoning through Probabilistic Soft Logic for Image Puzzle Solving

Somak Aditya, Yezhou Yang, Chitta Baral
School of Computing, Informatics and Decision Systems Engineering
Arizona State University
{saditya1,yz.yang,chitta}@asu.edu

Yiannis Aloimonos
UMIACS, Computer Science
University of Maryland, College Park
yiannis@cs.umd.edu

Appendices

A Additional Ablation varying Top Detections

			2.8k (WN)	
			K=1	K=5
VQA	VB	UR †	5.0	7.8
		GUR	6.4	10.6
Clarifai	VB	UR †	7.4	16.7
		GUR	7.2	16.6
	RR	UR	11.2	22.4
		GUR	12.2	23.3
	All	UR	**13.18**	28.9
		GUR	**13.2**	**29.9***
Resnet	VB	UR †	13.1	23.5
		GUR	14.8	23.2
	RR	UR	12.8	26.8
		GUR	14.9	26.2
	All	UR	16.1	28.2
		GUR	**16.5***	28.9

Table 1: Additional Ablation by Varying top K: Accuracy (in percentage) on the Image Riddle Dataset. Pipeline variants (VB, RR and All) are combined with Bias-Correction stage variants (GUR, UR). We show only wordnet-based accuracies by varying the top detections chosen. (*- Best, † - Baselines).

In this experiment, we vary the number (K) of top detections that we choose to calculate the similarity. We show our results for the 2.8K riddles (barring the 500 riddles kept for validation set). As the results show, the GUR variant (Clarifai+All+GUR and ResNet+All+GUR) achieves the best results. The WordNet based accuracy shows clear improvements (**13**% increase for Clarifai and **5**% increase over ResNet baseline, for top 5). This experiment also suggests, ResNet top K performance is really impressive for K=1.

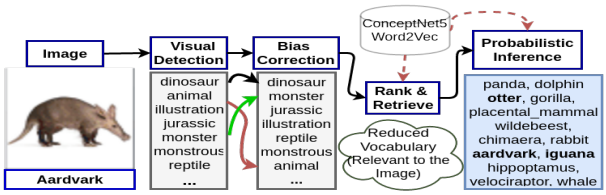

Figure 1: Clarifai detections and results from different stages for the aardvark image (for **BUR** variant).

B BiasedUnRiddler Variation (BUR)

In Figure 1: *dinosaur, animal* and *reptile* all provide evidence that the image has an animal. The word *dinosaur* provides some specific information. The other words do not add any additional information. Some high-confidence detections such as *monstrous, monster* provide erroneous abstract information. Hence, our next objective is to re-weight the seeds so that: i) the more specific seed-words should have higher weight than the ones which provide *similar* but more general information; ii) the seeds that are too frequently used or detected in corpus, should be given lower weights.

Specificity and Popularity: We compute eigenvector centrality score (ECS) for each word in the context of ConceptNet. Higher ECS indicates higher connectivity of a word in the graph. This yields a higher similarity score to many words and might give an unfair bias to this *seed* (and words implied by this *seed*) in the inference model. Hence, the higher the ECS, the word provides less specific information for an image. Additionally, we use the concreteness rating (CR) from Brysbaert et al. (2014). In this paper, the top 39955 frequent English words are rated from the scale of 1 (very abstract) to 5 (very concrete). For example, the mean ratings for *monster, animal* and *dinosaur* are $3.72, 4.61$ and 4.87 respectively.

Problem Formulation: We formulate the problem as a resource flow problem on a graph. The directed graph G

is constructed in the following way: we order the *seeds* based on decreasing centrality scores (CS). We compute CS as:

$$CS = (ECS + (-CR))/2, \quad (1)$$

where we normalize ECS and $-CR$ to the scale of 0 to 1. For each seed u, we check the immediate next node v and add an edge (u, v) if the (ConceptNet-based) similarity between u and v is greater than $\boldsymbol{\theta}_{sim,ss}$[1]. If in this iteration, a node v is not added in G, we get the most recent predecessor u for which the similarity exceed $\boldsymbol{\theta}_{sim,ss}$ and add (u, v).

If a word u is more abstract than v and if they are quite similar in terms of conceptual similarity, then word v provides similar but more specific information than word u. Each node has a resource $\tilde{P}(u|\mathcal{I}_k)$, the confidence assigned by the Neural Network. If there is an edge from the node, some of this resource should be sent along this edge until for all edges $(u, v) \in G$, w_v becomes greater than w_u. We formulate the problem as a Linear Optimization problem:

$$\begin{aligned} \underset{\boldsymbol{w}=(w_1,\ldots w_{|\boldsymbol{S_k}|})}{\text{minimize}} \quad & \sum_{(u,v)\in G} max\{w_u - w_v, 0\} \\ \text{subject to} \quad & \sum_{s\in \boldsymbol{S_k}} w_s = \sum_{s_k \in \boldsymbol{S_k}} \tilde{P}(\boldsymbol{s_k}|\mathcal{I}_k) \\ & w_u = \tilde{P}(u|\mathcal{I}_k), u \notin G \\ & w_u \geq 0.5\tilde{P}(u|\mathcal{I}_k), \forall u \in G \end{aligned}$$

To limit the resource a node u can send, we limit the final minimum value by $0.5\ \tilde{P}(u|\mathcal{I}_k)$. The solution provides us with the necessary weights for the set of seeds $\boldsymbol{S_k}$ in $\mathcal{I}_k$. We normalize these weights and get $\tilde{W}(\boldsymbol{S_k})$. These weights are then passed to the next stage.

C Intermediate Results for the "Aardvark" Riddle

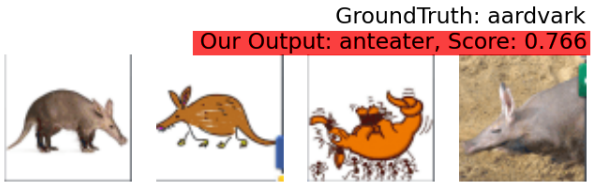

Figure 2: The four different Images for the "aardvark" riddle.

From the four figures in Figure 2, we get the top 20 Clarifai detections as given in the Table 2.

Based on the GUR approach (**GUR+All** in paper), our PSL Stage I outputs probable concepts (words or

[1]We use 0.9 as the value in our experiments.

Image1	Image2	Image3	Image4
monster	**food**	fun	rock
jurassic	small	retro	**nobody**
monstrous	vector	clip	travel
primitive	dinosaur	**halloween**	**water**
lizard	wildlife	**set**	**sea**
paleontology	cartoon	border	**aquatic**
vertebrate	nature	messy	outdoors
dinosaur	**evolution**	ink	**sand**
creature	reptile	design	**beach**
wildlife	outline	ornate	bird
nature	cute	decoration	wildlife
evolution	sketch	**ornament**	**biology**
reptile	painting	vector	**zoology**
wild	silhouette	**contour**	carnivora
horizontal	horizontal	cartoon	nature
illustration	art	cute	horizontal
animal	illustration	silhouette	animal
side view	graphic	art	side view
panoramic	animal	illustration	panoramic
mammal	panoramic	graphic	mammal

Table 2: Top 20 detections from Calrifai API. The detections that are completely noisy is colored using red. It can be observed that the third image does not give any evidence of an animal present.

Image1	Image2	Image3	Image4
dolphin	graph_toughness	decorative	bison
rhinoceros	cartography	graph_toughness	american_bison
komodo_dragon	color_paint	graph	**marsupial**
african_elephant	graph	artwork	gibbon
lizard	spectrograph	spectrograph	monotreme
gorilla	revue	kesho_mawashi	moose
crocodile	linear_functional	tapestry	mole
indian_elephant	simulacrum	map	**wildebeest**
wildebeest	pen_and_ink	arabesque	**echidna**
elephant	luck_of_draw	sgraffito	turtle
echidna	**cartoon**	linear_functional	mule_deer
chimaera	camera_lucida	hamiltonian_graph	mongoose
chimpanzee	explode_view	emblazon	tamarin
liger	micrographics	pretty_as_picture	chimpanzee
gecko	hamiltonian_graph	art_deco	wolverine
rabbit	crowd_art	dazzle_camouflage	prairie_dog
iguana	**depiction**	ecce_homo	western_gorilla
hippopotamus	echocardiogram	pointillist	**anteater**
mountain_goat	scenography	pyrography	okapi
loch_ness_monster	linear_perspective	echocardiogram	skunk

Table 3: Top 20 detections per each image from PSL Stage I (GUR).

Image1	Image2	Image3	Image4
panda	**like_paint**	hamiltonian_graph	giraffe
dolphin	projective_geometry	graph_toughness	waterbuck
african_forest_elephant	**diagram**	lacquer	sandy_beach
placental_mammal	line_of_sight	figuration	moose
otter	venn_diagram	war_paint	**wildebeest**
gorilla	hippocratic_face	graph	skunk
wildebeest	real_number_line	spectrograph	**anteater**
chimaera	sight_draft	map	**echidna**
african_savannah_elephant	x_axis	arabesque	bobcat
florida_panther	simulacrum	fall_off_analysis	mule_deer
liger	cartoon	art_collection	bison
rabbit	diagrammatic	statue	pygmy_marmoset
aardvark	camera_lucida	delineate	mongoose
iguana	explode_view	jack_o_lantern	sea_otter
hippopotamus	crowd_art	gussie_up	**squirrel_monkey**
hadrosaur	lottery	ecce_homo	wolverine
mountain_goat	depiction	pointillist	okapi
panda_bear	conecept_design	art_deco	cane_rat
velociraptor	infinity_symbol	pyrography	whale
whale	scenography	scenography	american_bison

Table 4: Top 20 detections per each image from PSL Stage I (BUR).

phrases) depending on the initial set of detected class-labels (*seeds*). They are provided in Table 3. Note that, these are the top *targets* detected from almost 0.2 million possible candidates. Observe the following:

i) the highlighted detected animals have a few visual features in common, such as *four short legs, a visible tail, short height* etc.

ii) the detections from the third image does not at all lead us to an animal and the PSL Stage I still thinks that its a cartoon of sort.

iii) the detections from second gets affected because of its close relation to the detections from third image and it infers that the image just depicts cartoon.

In the final PSL Stage II however, the model figures out that there is an animal that is common to all these images. This is mainly because *seeds* from the three images *confidently* predict that some animal is present in the images. That is why most of the top detections correspond to animals and animals having certain characteristics in common.

The top detections from PSL Stage II (GUR) are: *monotreme, gecko, hippopotamus, pyrography, anteater, lizard, mule_deer, chimaera, liger, iguana, komodo_dragon, echidna, turtle, art_deco, sgraffito, gorilla, loch_ness_monster, prairie_dog*.

BUR: For BUR, PSL Stage I outputs probable concepts (words or phrases) depending on the current set of *seeds*. They are provided in the Table 4. Observe that the individual detections are better compared to GUR[2].

Final output from PSL Stage II (for BUR) is comparable to that of the GUR approach. The top detections are: *hadrosaur, sea_otter, diagrammatic, panda, iguana, pyrography, mule_deer, placental_mammal, liger, panda_bear, art_deco, squirrel_monkey, giraffe, echidna, otter, anteater, pygmy_marmoset, hippopotamus*.

Here, the set of output mainly contains the concepts (words or phrases) that either represents "animals with some similar visual characteristics to aardvark" or it pertains to "cartoon or art".

[2]The output from the PSL Stage I for BUR, is completely independent of the other images. In essence, for each image, we are predicting all relevant concepts from a large vocabulary given a few detections from a small set of class-labels.

D Detailed Accuracy Histograms For Different Variants

In this section, we plot the accuracy histograms for the entire dataset for all the variants (using Clarifai API) of our approach (listed in Table 2 of the paper). We also add the accuracy histograms for variants using **BUR** approach. The plots are shown in the Figure 3. From the plots, the shift towards greater accuracy (increased height in rightmost bins) is evident as we go along the stages of our pipeline.

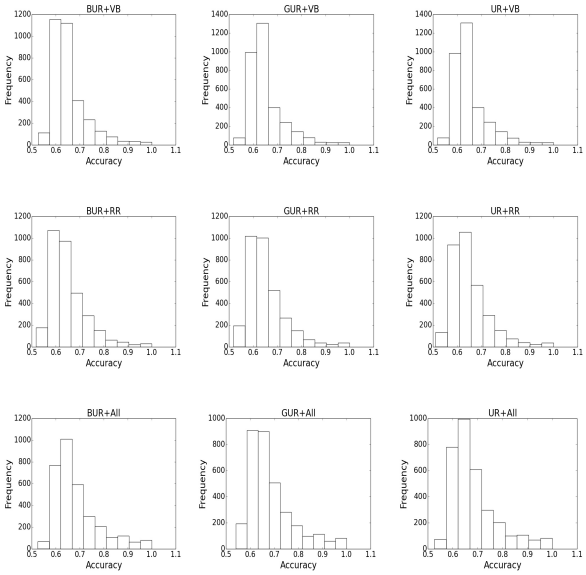

Figure 3: The Word2vec-based accuracy histograms of the BUR, GUR and UR approaches (combined with the VB, RR and All stage variants).

E Visual Similarity: Additional Results

Additional results for Visual Similarity are provided in Tables 5, 6 and 7.

ConceptNet	Visual Similarity	word2vec
man, merby, misandrous, philandry, male_human, dirty_pig, mantyhose, date_woman,guyliner,manslut	priest, uncle, guy, geezer, bloke, pope, bouncer, ecologist, cupid, fella	women, men, males, mens, boys, man, female, teenagers,girls,ladies

Table 5: Similar Words for "Men"

ConceptNet	Visual Similarity	word2vec
saurischian, ornithischian, protobird, elephant bird, sauropsid, cassowary, ibis, nightingale, ceratosaurian, auk, vulture	lambeosaurid, lambeosaur, bird, allosauroid, therapod, stegosaur, triceratops, tyrannosaurus_rex, deinonychosaur,dromaeosaur, brontosaurus	dinosaurs, dino, T._rex, Tyrannosaurus_Rex, T_rex, fossil, triceratops, dinosaur_species, tyrannosaurus,dinos, Tyrannosaurus_rex

Table 6: Similar Words for "Dinosaur"

F VQA Baseline Results

For the images in Figure 2, we show the top 20 answers in Table 8, generated from a state-of-the-art Visual Question Answering system (Lu et al. (2016)), for the questions "what is the image about?". As mentioned in the paper, it can be observed that the answers hardly contain any image-specific information. We believe, this is primarily due to the concept of attention used in the end-to-end learning systems. The words in the questions do not carry any specific information about a region, object or an attribute, for the "image understanding" system to find and hence the system is not able to generate meaningful answers. This shows i) how the problem of "image riddle" differs from traditional Visual Question Answering and ii) the need for systems which recognizes meaning without specific "attention" based on words. Our method, put forward in the paper, provides an example of one such system which utilizes background (ontological) knowledge to solve this puzzle (in other words, to answer this question).

G More Positive and Negative Results

We provide positive and Negative results in Figures 4 and 5 of the "GUR+All" variant of the pipeline. We obtain better results with Clarifai detections rather than Residual Network detections. Based on our observations, one of the key property of the ResidualNetwork confidence score distribution is that there are few detections (1-3) which are given the strongest confidence scores and the other detections have very negligible confidence scores. These top detections are often quite noisy.

For example, for the first image in the aardvark riddle (Figure 2), the ResidualNetwork detections are: ***tricer-***

ConceptNet	Visual Similarity	word2vec
snake, marmoset, lemur, sloth marmot, weasel, ferret, beaver, iguana, gecko, monkey, sauria, gazelle	skink, chameleon, iguana, gecko, this_picture, some_reptile, komodo_dragon, virginia, shark, garter_snake, rattlesnake, corn_snake, python	lizards, reptile, toad, snake frog, creature, critter, komodo_dragon, snakes, iguana

Table 7: Similar Words for "lizard"

aardvark_1	aardvark_2
resting, dog, cow, snowboarding, kitchen, black, military, elephant, racing, i don't know, horse, polo, sitting, grazing, running, standing, eating, brown, playing, walking	skis, school, playing game, kite flying, jet, bedroom, working, playing wii, scissors, navy, guitar, polo, snowboarding, plane, apple, orange, baseball, skateboarding, cutting, skiing
aardvark_3	aardvark_4
jumping, playing wii, working, kite flying, parasailing, cutting, traffic light, skateboard, flying, motorcycle, frisbee, navy, halloween, baseball, orange, snowboard, traffic, skateboarding, skiing, snowboarding	playing, nintendo, giraffe, milk, tv, tennis, rock, horse, lion, goat, brushing teeth, baseball, wii, bathroom, surfing, gray, elephant, sheep, standing, frisbee

Table 8: Answers from a Visual Question Answering system for the four images in Figure 2.

atops, *wallaby, armadillo, hog, fox squirrel, wild boar, kit fox, grey fox, Indian elephant, red fox, mongoose, Egyptian cat, wombat, tusker, mink, Arctic fox, toy terrier, dugong, lion*. Only the first detection has 0.84 score and the rest of the scores are very negligible. For the second, third and fourth images, the top detections are respectively:

1. **pick** (0.236), ocarina (0.114), maraca (0.091), chain saw (0.06), whistle (0.03), **can opener** (0.03), **triceratops** (0.02), muzzle, spatula, loupe, hatchet, letter opener, thresher, rock beauty, electric ray, tick, gong, Windsor tie, cleaver, electric guitar
2. **jersey** (0.137), **fire screen** (0.129), **sweatshirt** (0.037), pick (0.035), **comic book** (0.030), book jacket (0.029), plate rack, throne, wall clock, face powder, binder, hair slide,velvet,puck, redbone.
3. **hog** (0.48), wallaby (0.19), wild boar (0.10), Mexican hairless (0.045), gazelle (0.023), wombat (0.017), dhole (0.016), hyena (0.015), **armadillo** (0.009), ibex, hartebeest, water buffalo, bighorn, kit fox, **mongoose**, hare, wood rabbit, warthog, mink, polecat.

These predictions show that for the first and fourth image, there are some animals detected with some distant visual similarities. The second and third image has almost no animal mentions. This also shows some very confident detections (such as **triceratops** for the first image) is quite noisy.

In many cases, due to these high-confidence noisy detections, the PSL-based inference system gets biased towards them. Compared to that, Clarifiai detections provide quite a few (abstract but) correct detections about different aspects of the image (for example, for 2nd Image, predicts labels related to "cartoon/art" and "animal" both). This seems to be one of the reasons, for which the current framework provide better results for Clarifai Detections. Using Residual Network, the final output from the GUR system for the "aardvark" riddle is: *antelope, prairie_dog, volcano_rabbit, marsupial_lion, peccary, raccoon, pouch_mammal, rabbit,* ***otter****, monotreme,*

jackrabbit, hippopotamus, moose, ***tapir****,* ***echidna****, gorilla*.

References

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46(3):904–911, 2014. 1

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297, 2016. 4

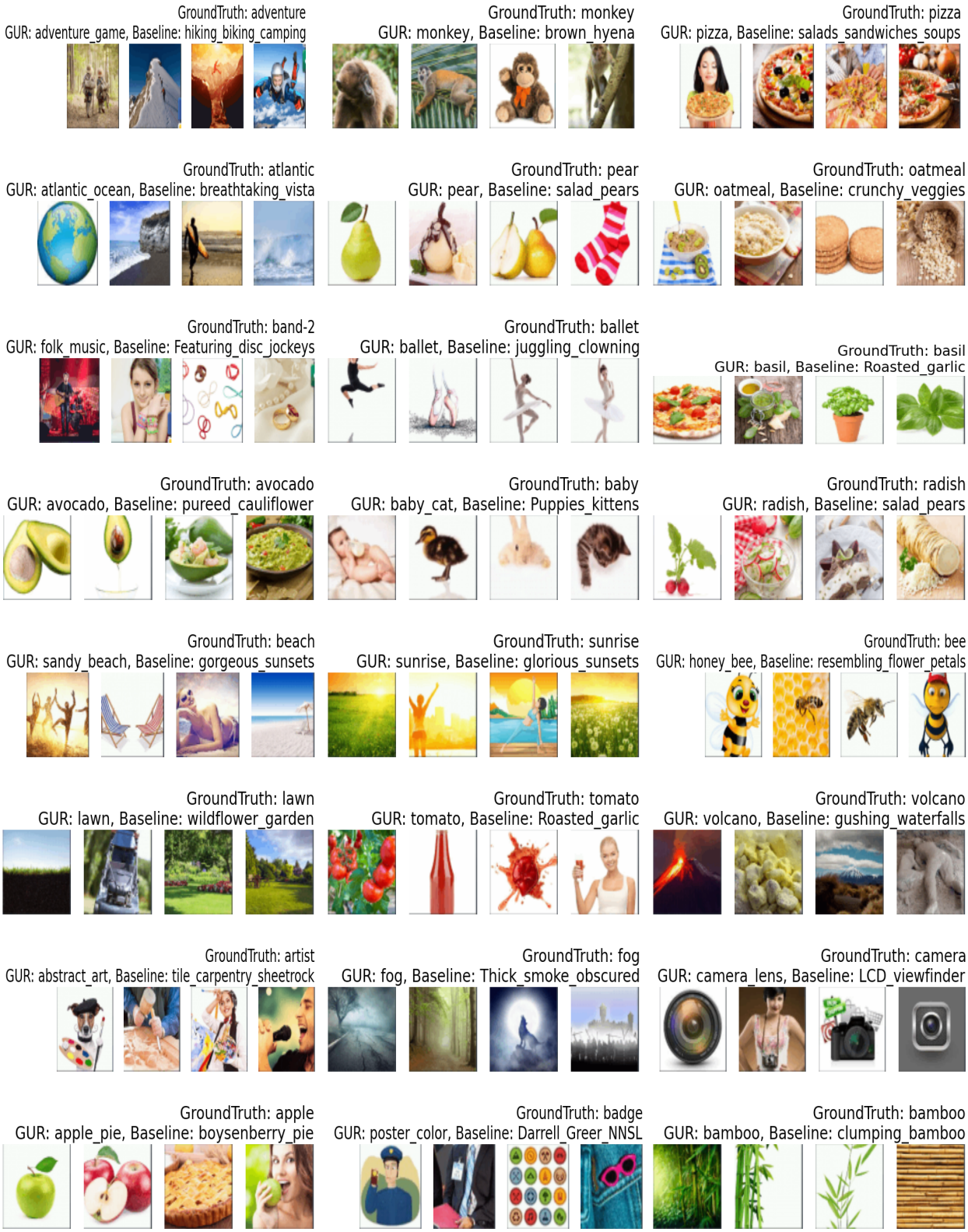

Figure 4: More Positive results from the "GUR" approach on some of the riddles. The groudtruth labels, closest label among top 10 from GUR and the Clarifai baseline are provided for all images. For more results, check http://bit.ly/1Rj4tFc.

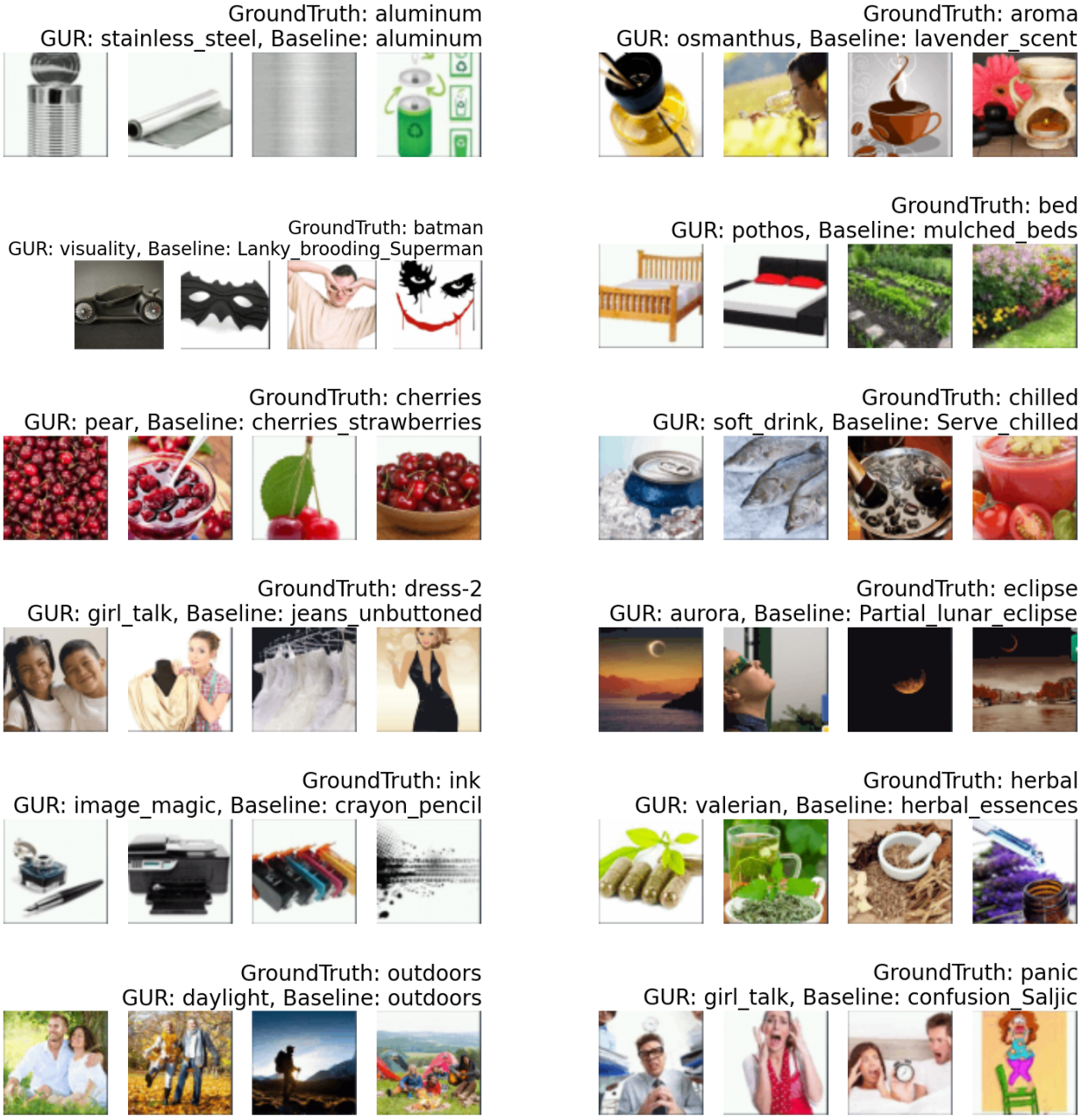

Figure 5: Some Negative results from the "GUR" approach on some of the riddles. The groudtruth labels, closest label among top 10 from GUR and the Clarifai baseline are provided for all images. For more results, check http://bit.ly/1Rj4tFc.