Supplementary Material for NEUCORE: Neural Concept Reasoning for Composed Image Retrieval

Shu Zhao
Pennsylvania State University
University Park, USA
smz5505@psu.edu

Huijuan Xu
Pennsylvania State University
University Park, USA
hkx5063@psu.edu

Editors: Marco Fumero, Emanuele Rodolà, Clementine Domine, Francesco Locatello, Gintare Karolina Dziugaite, Mathilde Caron

Table 1: The list of symbols and notations used in this paper. $\star \in \{r, t, rt\}$ denotes the reference, target, or concatenation of reference and target.

Symbol	Description
$I^\star$	raw image
$\mathbf{f}^\star$	visual tokens
$\mathbf{T}$	text modifier
$\mathbf{t}$	contextualized word features
$\mathbf{q}$	sentence feature
$\mathbf{a}$	attention weights
$\mathbf{f}_a^\star$	visual concept feature
$\mathbf{s}$	multi-modal alignment score
c	concept
$C(\cdot)$	concept set
$\mathbf{w}_c$	concept embedding
$\mathbb{M}$	concept vocabulary
$\mathbb{S}$	fusion sequence
$\hat{\mathbf{f}}$	modified feature
$\mathbf{m}$	concept matching score

1 List of Symbols

The list of symbols and notations used in this paper is shown in Table 1.

2 Concept Source

Pseudo concept labels are extracted according to part-of-speech. To evaluate the effectiveness of different concept types, we extract "Noun," "Adj,", "Verb," and "Adv" from the CIRR validation set and combine them as pseudo labels. The results are shown in Table 2. It demonstrates "Noun + Adj + Verb + Adv" achieves the best result, indicating that the number of concepts may affect the performance. More concepts help the model learn richer features.

Proceedings of the I edition of the Workshop on Unifying Representations in Neural Models (UniReps 2023).

Table 2: **Concept Source**. Pseudo concepts labels are extracted by a language parser according to part-of-speech on CIRR dataset.

PoS	$R@5$	$R_s@1$	$\frac{(R@5+R_s@1)}{2}$
Noun	49.12	43.20	46.16
Noun + Adj	49.25	43.28	46.27
Noun + Adj + Verb	50.00	44.01	47.01
Noun + Adj + Verb + Adv	**51.10**	**45.35**	**48.22**

Table 3: **Detailed results on Fashion IQ dataset**.

Method	$\frac{R@10+R@50}{2}$	$R@10$				$R@50$			
		Dress	Shirt	Toptee	Mean	Dress	Shirt	Toptee	Mean
ComposeAE [Anwaar et al., 2021]	20.60	-	-	-	11.80	-	-	-	29.40
TCIR [Chawla et al., 2021]	29.51	19.33	14.47	19.73	17.84	43.52	35.47	44.56	41.18
CIRPLANT [Liu et al., 2021]	25.17	14.38	13.64	16.44	14.82	34.66	33.56	38.34	35.52
TIRG [Vo et al., 2019]	35.32	23.80	19.90	25.82	23.17	48.64	42.14	51.64	47.48
VAL [Chen et al., 2020]	33.82	21.12	21.03	25.64	22.60	42.19	43.44	49.49	45.04
CoSMo [Lee et al., 2021]	31.26	21.39	16.90	21.32	19.87	44.45	37.49	46.02	42.65
ARTEMIS [Delmas et al., 2022]	38.17	**27.16**	21.78	29.20	26.05	52.40	43.64	54.83	50.29
NEUCORE	**39.15**	27.00	**22.84**	**29.63**	**26.45**	**53.79**	**45.00**	**56.65**	**51.75**

3 Detailed Results on Fashion IQ dataset

Table 3 illustrates the detailed results on Fashion IQ validation set. It demonstrates that our proposed NEUCORE model can outperform the SOTA method ARTEMIS in most of the metrics. This also validates that progressive fusion with aligned multi-modal concept alignment can improve the composed image retrieval task.

4 Qualitative Results

We provide more qualitative retrieval examples from a restricted subset of the CIRR validation set [Liu et al., 2021] where candidate target images are visually similar. It is challenging because the model needs to learn fine-grained vision and language features and their interactions. The retrieval examples are shown in Figure 1. Results demonstrate our NEUCORE model can understand the content of text modifier, find correct correspondence between visual concepts and semantic concepts, and compose the reference image feature and text modifier feature to identify the target image feature.

5 Visualization of Concept Alignment

Our NEUCORE model can mine and align visual concepts with semantic concepts. We employ Grad-CAM [Selvaraju et al., 2017] to identify visual concepts corresponding to semantic concepts on CIRR dataset. The correct visualization results are shown in Figure 2. It demonstrates that our NEUCORE model can align semantic concepts to visual concepts in images under image level weak supervision. We also show some failure visualization cases in Figure 3, where these semantic concepts are describing more abstract visual concepts. We expect that using more advanced vision and language pre-trained models can help alleviate these failure cases.

6 List of Zero-shot Concepts on CIRR_zs Dataset

To demonstrate that our NEUCORE model can deal with novel zero-shot concepts, we create a data split from CIRR validation set, named CIRR_zs. The zero-shot concepts in CIRR_zs are listed as follows:

plus, winglike, unicorn, lounger, camisole, wound, gentle, ibex, hinged, silicone, vary, jajantic, crayon, zone, servicing, rollaway, sorted, description, ended, secondhand, rop, softie, winner, makw, furnished, birn, moblie, hospital, orient, gose, bulk, cum, whote, pilot, hyppopotamus, sharo, simillar, thread,

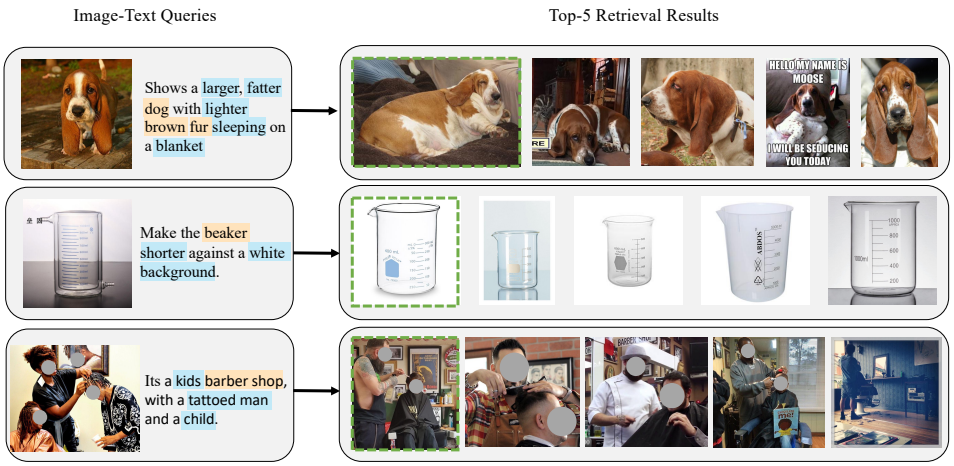

Figure 1: Qualitative examples of image-text queries on CIRR validation set and its Top-5 retrieval results. Green dotted boxes denote the ground-truth target images, and semantic concepts in the input text modifiers are highlighted by different colors. Yellow denotes that a concept appears both in the reference and target images. Blue means that a concept appears only in the target image.

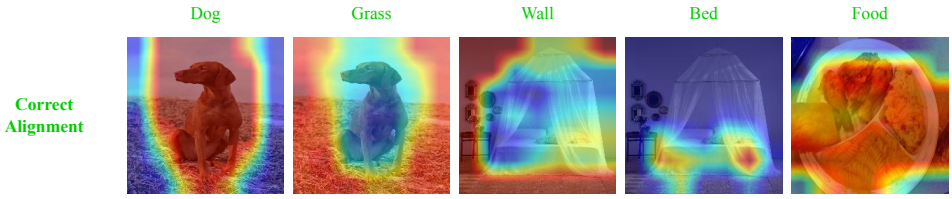

Figure 2: The visualization results of correct concept alignment on CIRR dataset.

aniamls, celebration, committee, freestyle, scubadiver, decorated, counch, help, dia, fatter, goldtone, shovel, earing, huddled, florescent, handy, aggressive, sis, undecipherable, attitude, asembled, settle, celebrity, vietnamise, clapping, suv, cologne, wax, prepared, law, unrisen, boklane, screwlike, carcas, treet, monk, handleless, repaired, atheletic, biting, commercial, shrimp, entree, ticker, graze, intense, portait, buddy, swin, engineer, chemistry, unobvious, avenue, seashelle, gummie, misssig, empanada, puzzle, hazelnut, driverside, maze, seller, gnu, another, anmial, gummy, knob, mounted, thermal, furth, seagal, handing, wolflike, clap, barnlike, gape, clover, medusa, under, siringe, vegatation, raelistic, outrigger, cinnamon, brwon, screwtop, rocksurface, snorkeler, creative, wingspan, coastal, innocent, ostrich, volture, eatm, gril, cyclist, on, aless, pinecone, comedy, blueprint, spaghetti, inverted, fried, lea, treeline, unmanned, sweatshirt, swirly, silvertone, lengthy, coverge, attenae, multipack, stocking, hypopotamus, mosque, continental, baggage, stickering, core, silve, hollowed, swimwear, noodle, eyeliner, sphinx, multilayer, morevie, abdoman, hartebe, thumbnail, bouqet, boundary, glide, palican, boklaine, trianglular, steer, orthodox, assembly, canister, terrestrial, variable, spoonful, value, hyppopotomus, liner, wardrove, blackc, ribbontail, diameter, multimeter, sculture, droopy, hi, cachrro, removed, prayer, scent, italian, trimming, coulple, zest, calimari, jogger, frock, healthier, box(es, rabie, supine, sape, mkaing, dungeness, cubicle, last, unrenovated, miror, patterened, dool, pallet, femal, denser, mongoose, pelikan, mottled, mechanic, 50ml, affectionate, wintery, pitchet, cuttingboard, starbuck, thong, churchlike, handler, technology, hyypopotomus, bib, wipe, swollen, tong, hawaiian, lingerie, tupperware, hexagon, flatcap, omlette, makin, direciton, pearl, princess, hooded, façade, playin, skincare, smaler, screwdriver, spill, safe, throny, gopher, furier, embrace, luminous, choir, heard, taupe, bouse, unbrand, tentacled, carrier, gutted, oral, mouthed, squid, active, triney, landscaping, coffe, state, visibility, tortoise, buffet, erotic, bloody, barking, puppeis, account, risen, peper, anima, sandwedge, mold, guide, religious, largre, panty, sipper, differnt, chubbier, rwmove, welcoming, countryard, tattoed, compression, interacting, dove, veggie, descriptive, feminine, highway, orientation, bride, potote, disc, rocklike.

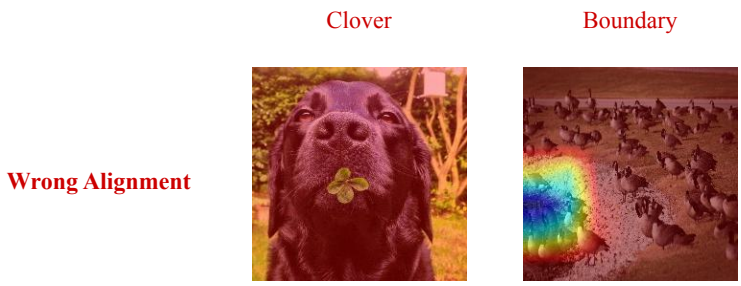

Figure 3: The visualization results of wrong concept alignment on CIRR dataset.

7 Algorithm for Learning the NEUCORE model

The overall learning algorithm for our proposed NEUCORE model is illustrated in Algorithm 1.

Algorithm 1 The overall learning algorithm for the NEUCORE model.

Input: Reference image I_r; Target image I_t; Text modifier T.
Output: Matching score $\mathbf{m}$; Parameters of the NEUCORE model.

Obtain semantic concepts $C(T)$ from a language parser.
repeat
  # Encode vision and language features
  Extract reference visual tokens $\mathbf{f}^r$, target visual tokens $\mathbf{f}^t$, contextualized word features $\mathbf{t}$, and sentence feature $\mathbf{q}$ by Equation (1).

  # Multi-modal concept Alignment. Section 3.1
  Obtain semantic concept word embeddings $\mathbf{w}_c$ by Equation (2).
  Concatenate reference and target visual tokens to get concatenated tokens $[\mathbf{f}^r, \mathbf{f}^r]$ and feed them to transformer layer to exchange their information and obtain $\mathbf{f}^{rt}$ by Equation (3).
  Apply a token-wise softmax operation for concatenated tokens $\mathbf{f}^{rt}$ to get attention weights $\mathbf{a}$ and weighted summary concatenated tokens $\mathbf{f}^{rt}$ according to the attention weights $\mathbf{a}$ by Equation (4).
  Calculate the multi-modal alignment score $\mathbf{s}$ by Equation (5).

  # Progressive multi-modal fusion over concepts. Section 3.2
  Generate fusion sequence $\mathbb{S}$ by Equation (7).
  Progressively fuse the reference image tokens $\mathbf{f}^r$ and the text modifier feature stored in fusion steps $\mathbb{S}$ over concepts $\mathbf{w}_c$ by Equation (8).
  Calculate the concept matching score $\mathbf{m}$ by Equation (9).

  # Loss function
  Calculate the multi-modal concept alignment loss value $\mathcal{L}_c$ by Equation (6).
  Calculate the matching loss value $\mathcal{L}_m$ by Equation (10).
  Calculate the final loss value by Equation (11) and optimize it by BP algorithm.
until Convergence or reach maximum iterations.

8 Limitations and Future Work

Our proposed model, NEUCORE, can mine and align multi-modal concepts without concept-level supervision. However, the improvement in Fashion IQ is relatively small than CIRR. It is mainly because Fashion IQ contains domain-specific concepts, like "suede" and "bely." CIRR is more diverse and covers more concepts. Recently, large vision-language (VL) models have achieved significant progress. Our model does not employ these VL models currently and only focuses on model side, but potentially these VL models could help our model learn more domain-specific concepts. On the other hand, we decompose the text modifier to generate a fusion sequence and progressively fuse the

reference image feature and text modifier feature over aligned multi-modal concepts. Large language models (LLM) can also be utilized to generate a more accurate fusion sequence.

References

Muhammad Umer Anwaar, Egor Labintcev, and Martin Kleinsteuber. Compositional learning of image-text query for image retrieval. In *WACV*, pages 1139–1148. IEEE, 2021.

Pranit Chawla, Surgan Jandial, Pinkesh Badjatiya, Ayush Chopra, Mausoom Sarkar, and Balaji Krishnamurthy. Leveraging style and content features for text conditioned image retrieval. In *CVPR Workshops*, pages 3978–3982. Computer Vision Foundation / IEEE, 2021.

Yanbei Chen, Shaogang Gong, and Loris Bazzani. Image search with text feedback by visiolinguistic attention learning. In *CVPR*, pages 2998–3008. Computer Vision Foundation / IEEE, 2020.

Ginger Delmas, Rafael Sampaio de Rezende, Gabriela Csurka, and Diane Larlus. ARTEMIS: attention-based retrieval with text-explicit matching and implicit similarity. In *ICLR*. OpenReview.net, 2022.

Seungmin Lee, Dongwan Kim, and Bohyung Han. Cosmo: Content-style modulation for image retrieval with text feedback. In *CVPR*, pages 802–812. Computer Vision Foundation / IEEE, 2021.

Zheyuan Liu, Cristian Rodriguez Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *ICCV*, pages 2105–2114. IEEE, 2021.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626. IEEE Computer Society, 2017.

Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval - an empirical odyssey. In *CVPR*, pages 6439–6448. Computer Vision Foundation / IEEE, 2019.