BMVC 2025

Emil Demić and Luka Čehovin Zajc

University of Ljubljana, Faculty of Computer and Information Science — ViCoS Visual Cognitive Systems lab

present

BACK TO THE DRAWING BOARD

(RETHINKING SCENE-LEVEL SKETCH-BASED IMAGE RETRIEVAL)

PROBLEM

Retrieve natural images of complex scenes matching the overall semantics and spatial layout of a free-hand sketch.

CHALLENGES

- Limited training data
- Selective relevance
- Highly ambigiuous

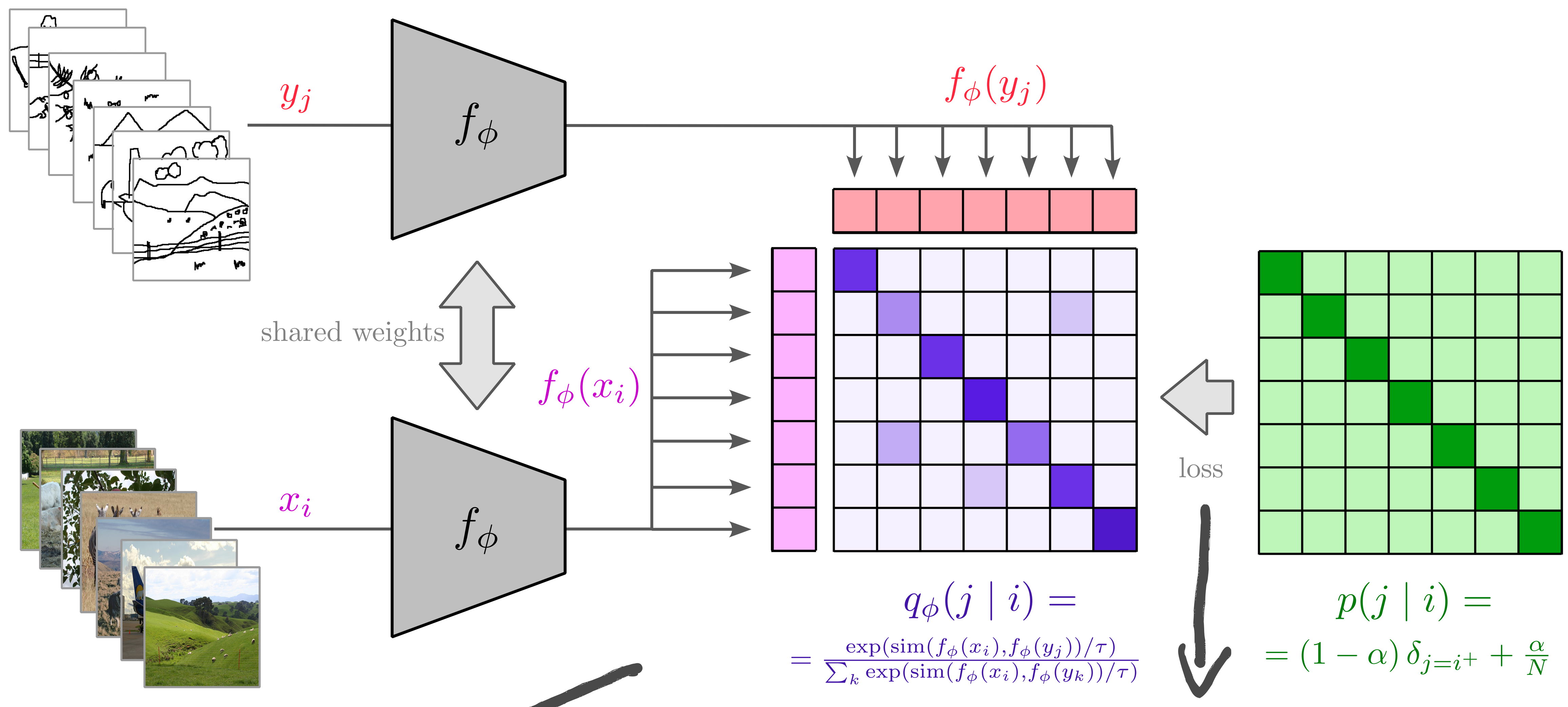
STATE-OF-THE-ART

- Outdated architectures
- Unsuitable pretraining
- Underutilizaton of data

SOLUTION

Use suitable combination of:

1. Appropriate encoder model (ConvNeXt)
2. Cross-modal pre-training (CLIP)
3. Robust loss formulation (de-biased InfoNCE)



$$\mathcal{L} = \sum_i \mathrm{KL}\left(p(\cdot \mid i) \,\|\, q_\phi(\cdot \mid i)\right)$$

RESULTS (FS-COCO)

	Normal			Unseen		
	R@1	R@5	R@10	R@1	R@5	R@10
Siam-VGG	23.3	-	52.6	10.6	-	32.5
HOLEF-VGG	22.8	-	53.1	*10.9*	-	*33.1*
SceneTrilogy	24.1	-	53.9	-	-	-
SceneDiff (w Sketch)	25.2	45.9	55.9	-	-	-
FreestyleRet	*29.6*	-	*56.1*	-	-	-
Ours	**61.9**	**81.4**	**87.2**	**60.0**	**80.2**	**86.1**

ABLATION

Performed on unseen split of FS-COCO

	R@1	R@5	R@10
ImageNet	0.0	0.2	0.5
CLIP	6.4	14.7	21.2
CLIP + triplet	22.7	42.9	54.2
CLIP + InfoNCE	52.2	74.3	81.7
CLIP + ICon	60.0	80.2	86.1
CLIP (ViT) + ICon	53.9	76.6	83.5

EXAMPLES

MORE EXPERIMENTS AND RESULTS?

CODE AND WEIGHTS!

ONLINE DEMO!?