MATHÉMATIQUES
VISION
APPRENTISSAGE

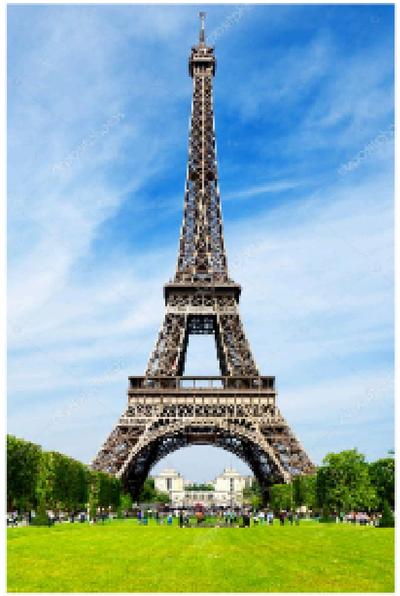
école
normale
supérieure
paris—saclay

TOPIC A - OBJECT RECOGNITION AND COMPTER VISION

Aymane El Firdoussi

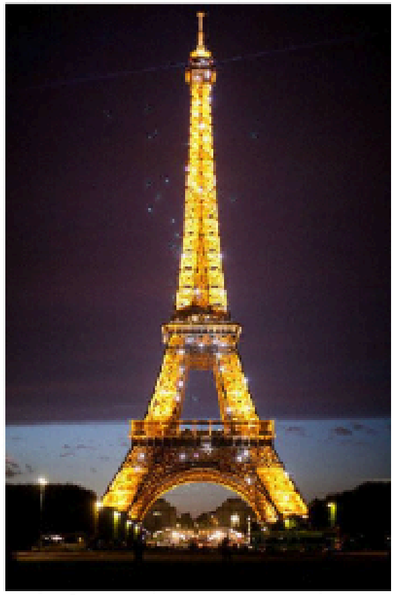
INTRODUCTION

Composed Image Retrieval (CoIR)



reference image

composed



target image

At night

modification text

Goal of the project

- Reproduce a result of the CoVR (or CoVR-2) paper about the performance of BLIP (or BLIP-2) on the CIRR dataset.
- Try a small extension of the original paper.

CoVR: Learning Composed Video Retrieval from Web Video Captions

Lucas Ventura[1,2], **Antoine Yang**[2], **Cordelia Schmid**[2], **Gül Varol**[1]

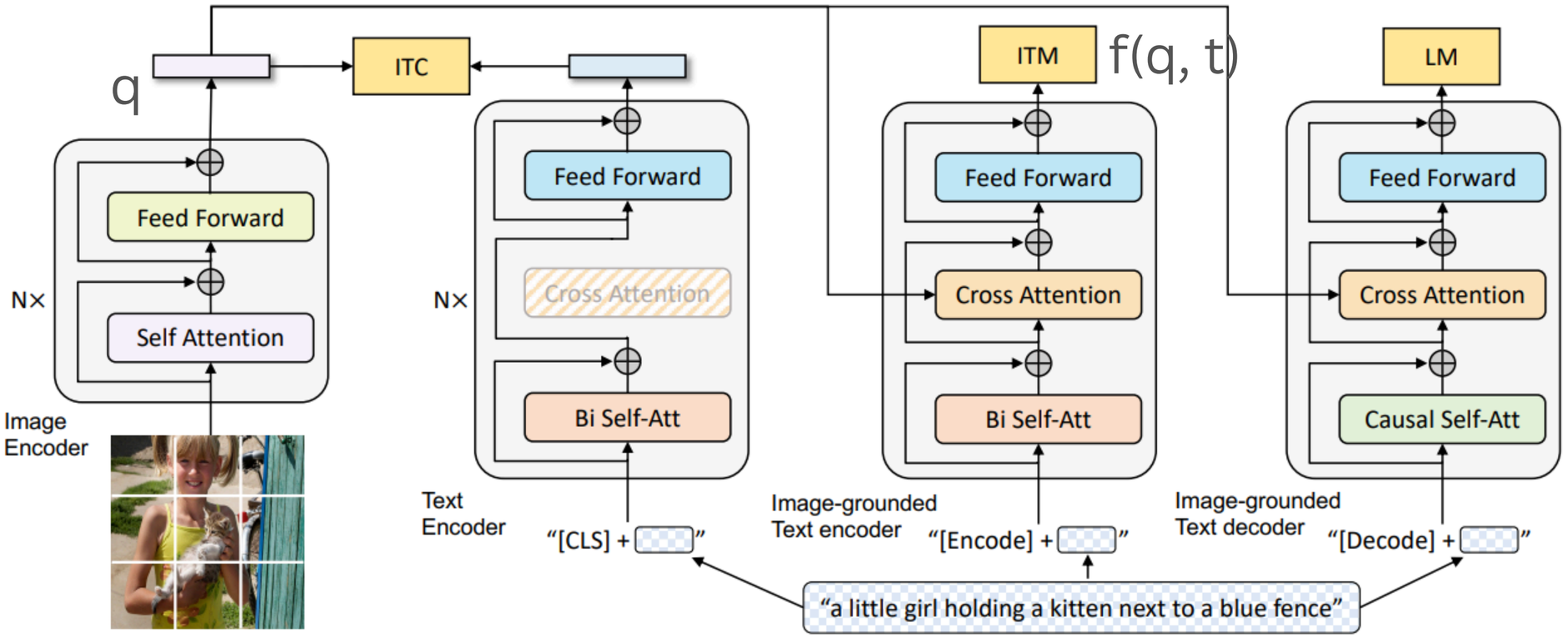
[1]LIGM, École des Ponts, Univ Gustave Eiffel, CNRS, France
[2] Inria, ENS, CNRS, PSL Research University, France
lucas.ventura@enpc.fr

PLAN

1- Reproducting CoVR paper results
2- On the impact of the embeddings
3- Conclusion

RESULTS REPRODUCTION

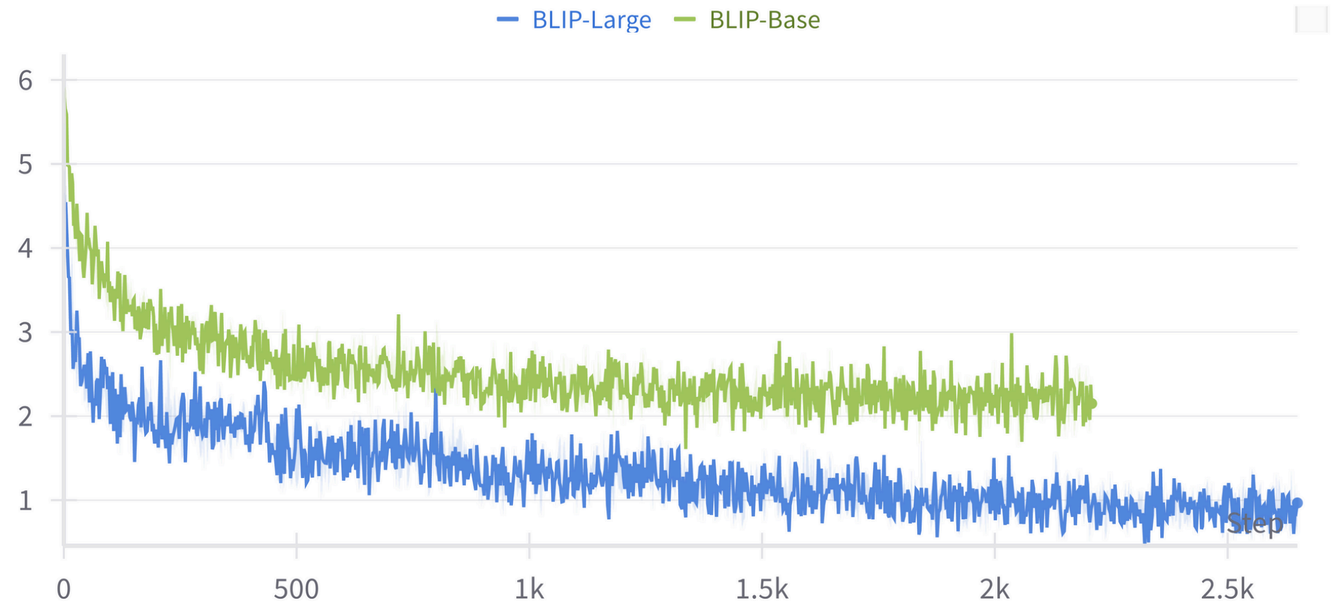
BLIP model in a nutshell



- We want to maximize the cosine-similarity between f and the target image embedding (minimizing the HNCE loss).

Training

- Trained BLIP-Base (Capfilt ckpt) and BLIP-Large (finetuned on COCO) models on the CIRR dataset using 4 GPUs (NVIDIA P4) with a batch-size of 16 (paper 1024) and with 16-bit Mixed precision (to accelerate training).
- We train / evaluate our models on the Compose Image Retrieval on Real-life images (CIRR) dataset.



Results

- We get lower results compared to the ones in the paper due to the drastic decrease in the batch size (from 1024 to 16).

Model	R@1	R@5	R@10	R@50
BLIP-Large	49.16	79.76	88.65	97.49
BLIP-Large*	27.03	67.42	80.08	95.07
BLIP-Base*	21.84	59.02	73.60	93.23

ON THE IMPACT OF THE EMBEDDINGS

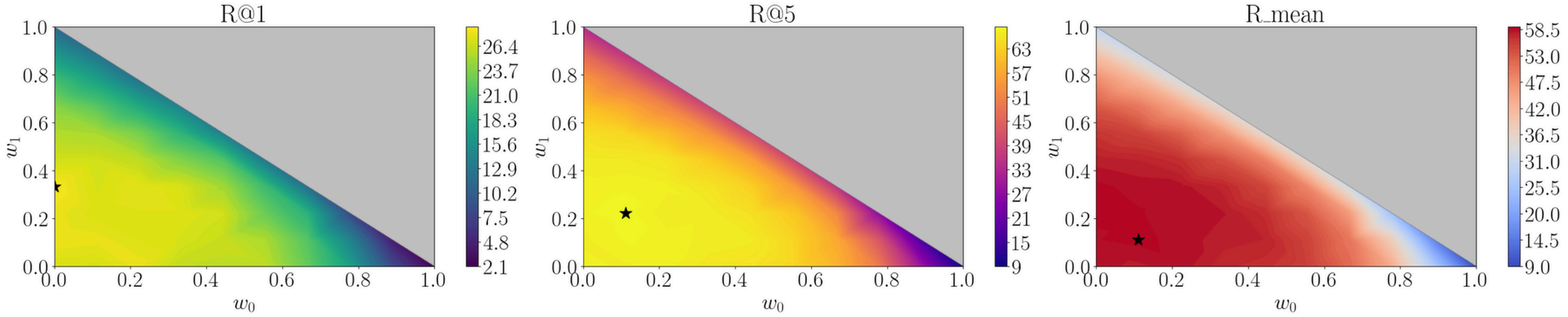
Fixing by Mixing

- We want to investigate on the efficiency of the multimodal-embedding f(q,t).
- We introduce the “Mixed” embedding:

$$m(q,t) = w_0 q + w_1 t + w_2 f(q,t), \quad \sum_{i=0}^{2} w_i = 1 \qquad (1)$$

Analyzing the results

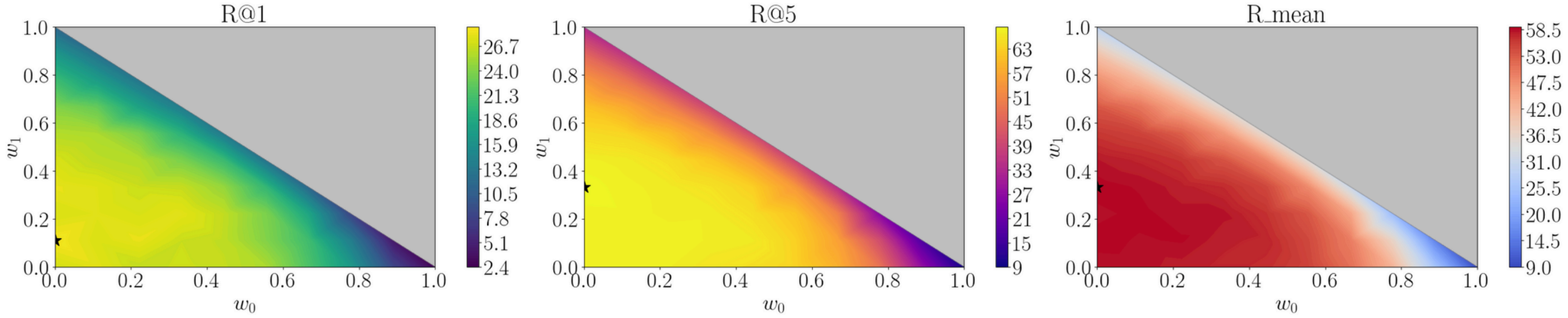
- Counter-intuitive result: that we don’t have the same w_0 and w_1 that maximizes all recalls.



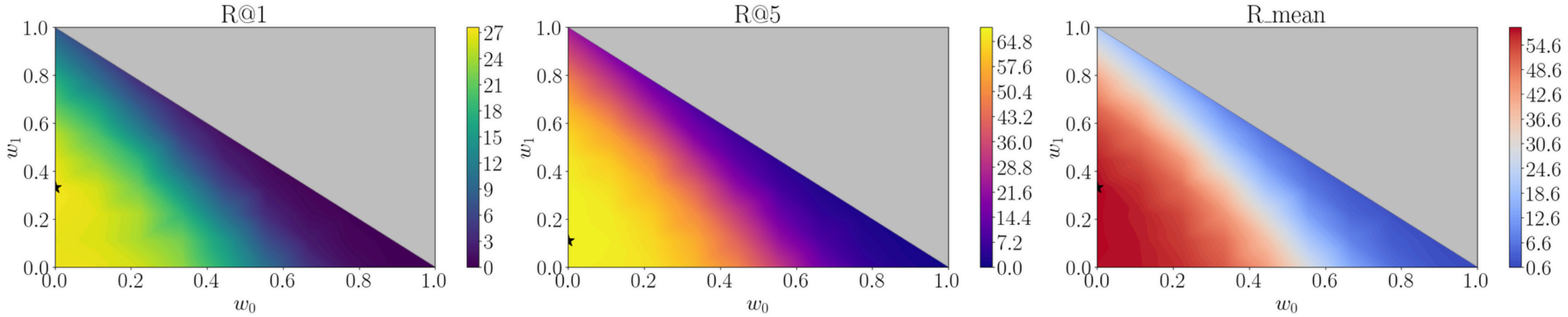
Model	R@1	R@5	Avg
baseline	27.03	67.42	58.18
Mixed	**28.32**	**68.14**	**58.69**

About the aggregation rule

- I also tried Median and Max-pooling instead of simple averaging, and got the following results:



Median



Max-pooling

CONCLUSION

Conclusion & Potential future directions

- Overall, the multimodal embedding is sufficient (not much difference between m and f).
- Also tried to learn the weights (w_0, w_1 and w_2), but it took too much credits and time, and didn’t give satisfying results (needed tuning).
- Explore the mixed embedding on other datasets: FashionIQ, WebVid-CoVR dataset, etc.
- Examine other approach to compare vectors instead of Cosine similarity ?

ACKNOWLEDGMENT

I would like to thank my project supervisors: Mr. Lucas Ventura and Dr. Gül Varol for their constant help and their fast e-mail answering (even during holidays)

Thank you all for your attention !