Transformer Decoders with MultiModal Regularization for Cross-Modal Food Retrieval

Mustafa Shukor[1*] Guillaume Couairon[1,2] Asya Grechka[1,3] Matthieu Cord[1,4]

[1]Sorbonne University [2]Meta AI [3] Meero [4] Valeo.ai

A. Appendix

A.1. AdaMargin Triplet Loss

Here we detail another variant that we propose for adaptive triplet with dynamic margin. The number of active triplets (the triplets corresponding to non-zero triplet loss l) in the batch reveals the difficulty of the task; if it is small, that means most of the triplets already satisfy the triplet condition (*i.e.*, the difference between the distance of positive pairs is smaller than the distance of the negative pairs by a margin). Thus, we propose to inversely weigh the margin by the number of active triplets δ in each batch. Thus the loss can be written as:

$$\mathcal{L}_m(\mathcal{B}_a, \mathcal{B}_p, \mathcal{B}_n) = \sum_{x_a \in \mathcal{B}_a} \sum_{x_p \in \mathcal{B}_p} \sum_{x_n \in \mathcal{B}_n} l(x_a, x_p, x_n, \alpha/\delta) \tag{1}$$

Where $\mathcal{B}_a$, $\mathcal{B}_p$ and $\mathcal{B}_n$ are the set of anchors, positive and negative examples in the batch. We keep the α/δ between 0.05 and 0.3, and δ is computed based on $\alpha = 0.3$

A.2. Ablation Study

In this section, we investigate the importance of different design choices. Our baseline (B) consists of the dual encoders, ViT for image encoder, hiererchical transformer (HT) for recipe encoder (*i.e.*, with T and HT), and trained with Adamine loss.

HTD: In Table 1, we investigate the importance of the title for the instructions and ingredients, thus we compare our approach with another variant that takes only the ingredients (resp. instructions) as K and V with the instructions (resp. ingredients) as queries. We can notice that the title does not bring additional improvement to the instructions and ingredients.

ITEM: We did an ablation study for ITEM in Table 2. In particular, we test the module with all recipe elements as K and V (ITEM (a)), with the title alone (ITEM (t)) and with ingredients alone (ITEM (n)). The best results are obtained with all recipe elements, which also validate that all recipe elements are important for multimodal fusion.

*Corresponding author: mustafa.shukor@sorbonne-universite.fr

	image-to-recipe				recipe-to-image			
	medR	R1	R5	R10	medR	R1	R5	R10
Ours	1	66.9	87.0	91.0	1	67.5	87.2	91.1
v2	**1**	**67.3**	**87.3**	**91.1**	**1**	**67.6**	**87.5**	**91.2**

Table 1. Ablation study for HTD. medR (↓), Recall@k (↑) are reported on the Recipe1M test set with 1k setup. v2: in HTD, the cross attention of ingredients takes only the instructions as K and V and the one for instructions takes only the ingredients as K and V, while the title takes both of them as K and V.

	image-to-recipe				recipe-to-image			
	medR	R1	R5	R10	medR	R1	R5	R10
B*	1	66.6	87.5	91.0	1	67.6	88.1	91.1
B* + ITEM (t)	1	65.8	87.9	**91.4**	1	66.0	88.4	**91.4**
B* + ITEM (n)	1	66.7	87.9	91.1	1	67.2	88.4	91.1
B* + ITEM (a)	**1**	**67.5**	**88.1**	90.9	**1**	**68.4**	**88.6**	91.0

Table 2. Ablation study. medR (↓), Recall@k (↑) are reported on the Recipe1M test set with 1k setup. B* is with ViT, HT, HTD and MTD. We compare several ways of enhancing the image tokens, the cross attention is with; all recipe elements (ITEM (a)), only the title (ITEM(t) and only the ingredients (ITEM (n)). The cross attention between the image and all recipe elements gives the best results.