Cross-Modal Retrieval in the Cooking Context: Learning Semantic Text-Image Embeddings

Micael Carvalho*
Sorbonne Université, CNRS, LIP6, F-75005 Paris, France
micael.carvalho@lip6.fr

Rémi Cadène*
Sorbonne Université, CNRS, LIP6, F-75005 Paris, France
remi.cadene@lip6.fr

David Picard
Sorbonne Université, CNRS, LIP6, F-75005 Paris, France
david.picard@lip6.fr

Laure Soulier
Sorbonne Université, CNRS, LIP6, F-75005 Paris, France
laure.soulier@lip6.fr

Nicolas Thome
CEDRIC - Conservatoire National des Arts et Métiers, F-75003 Paris, France
nicolas.thome@cnam.fr

Matthieu Cord
Sorbonne Université, CNRS, LIP6, F-75005 Paris, France
matthieu.cord@lip6.fr

ABSTRACT

Designing powerful tools that support cooking activities has rapidly gained popularity due to the massive amounts of available data, as well as recent advances in machine learning that are capable of analyzing them. In this paper, we propose a cross-modal retrieval model aligning visual and textual data (like pictures of dishes and their recipes) in a shared representation space. We describe an effective learning scheme, capable of tackling large-scale problems, and validate it on the Recipe1M dataset containing nearly 1 million picture-recipe pairs. We show the effectiveness of our approach regarding previous state-of-the-art models and present qualitative results over computational cooking use cases.

CCS CONCEPTS

• **Information systems** → **Multimedia and multimodal retrieval**; • **Computer systems organization** → **Neural networks**;

KEYWORDS

Deep Learning, Cross-modal Retrieval, Semantic Embeddings

ACM Reference Format:
Micael Carvalho, Rémi Cadène, David Picard, Laure Soulier, Nicolas Thome, and Matthieu Cord. 2018. Cross-Modal Retrieval in the Cooking Context: Learning Semantic Text-Image Embeddings. In *SIGIR '18: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, July 8–12, 2018, Ann Arbor, MI, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3209978.3210036

1 INTRODUCTION

Designing powerful tools that support cooking activities has become an attractive research field in recent years due to the growing interest of users to eat home-made food and share recipes on social platforms [34]. These massive amounts of data shared on devoted sites, such as *All Recipes*[1], allow gathering food-related data including text recipes, images, videos, and/or user preferences. Consequently, novel applications are rising, such as ingredient classification [7], recipe recognition [38] or recipe recommendation [34]. However, solving these tasks is challenging since it requires taking into consideration 1) the heterogeneity of data in terms of format (text, image, video, ...) or structure (*e.g.*, list of items for ingredients, short verbal sentence for instructions, or verbose text for users' reviews); and 2) the cultural factor behind each recipe since the vocabulary, the quantity measurement, and the flavor perception is culturally intrinsic; preventing the homogeneous semantics of recipes.

One recent approach emerging from the deep learning community aims at learning the semantics of objects in a latent space using the *distributional hypothesis* [14] that constrains object with similar meanings to be represented similarly. First used for learning image representations (also called *embeddings*), this approach has been derived to text-based applications, and recently some researchers investigate the potential of representing multi-modal evidence sources (*e.g.*, texts and images) in a shared latent space [20, 30]. This research direction is particularly interesting for grounding language with common sense information extracted from images, or vice-versa. In the context of computer-aided cooking, we believe that this multi-modal representation learning approach would contribute to solving the heterogeneity challenge, since they would promote a better understanding of each domain-specific word/image/video. In practice, a typical approach consists in aligning text and image representations in a shared latent space in which they can be compared [6, 8, 20, 21, 24, 33]. One direct application in the cooking context is to perform cross-modal retrieval where the goal is to retrieve images similar to a text recipe query or conversely text recipes similar to a image query. However, Salvador et al. [33] highlight that this solution based on aligning matching pairs can lead to poor retrieval performances in a large scale framework. Training the latent space by only matching pairs of the exact same dish is not particularly effective at mapping similar dishes close together, which induces a lack of generalization to newer items (recipes or images). To alleviate this problem, [33] proposes to use additional

*Equal contribution

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SIGIR '18, July 8–12, 2018, Ann Arbor, MI, USA
© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-5657-2/18/07…$15.00
https://doi.org/10.1145/3209978.3210036

[1] http://www.allrecipes.com/

(a) im2recipe [33] (b) AdaMine (ours)

Figure 1: Comparison between (a) the classification augmented latent space learning of [33] and (b) our joint retrieval and semantic latent space learning, which combines instance-based (L_{ins}) and semantic-based (L_{sem}) losses.

data (namely, categories of meals) to train a classifier with the aim of regularizing the latent space embeddings (Figure 1a). Their approach involves adding an extra layer to a deep neural network, specialized in the classification task. However, we believe that such classification scheme is under-effective for two main reasons. First, the classifier adds many parameters that are discarded at the end of the training phase, since classification is not a goal of the system. Second, we hypothesize that given its huge number of parameters, the classifier can be trained with high accuracy without changing much of the structure of the underlying latent space, which completely defeats the original purpose of adding classification information.

To solve these issues, we propose a unified learning framework in which we simultaneously leverage retrieval and class-guided features in a shared latent space (Figure 1b). Our contributions are three-fold:

- We formulate a joint objective function with cross-modal retrieval and classification loss to structure the latent space. Our intuition is that directly injecting the class-based evidence sources in the representation learning process is more effective at enforcing a high-level structure to the latent space as shown in Figure 1b.
- We propose a double-triplet scheme to express jointly 1) the retrieval loss (e.g., corresponding picture and recipe of a pizza should be closer in the latent space than any other picture - see blue arrows in Figure 1b) and 2) the class-based one (e.g., any 2 pizzas should be closer in the latent space than a pizza and another item from any other class, like salad). This double-loss is capable of taking into consideration both the fine-grained and the high-level semantic information underlying recipe items. Contrary to the proposal of [33], our class-based loss acts directly on the feature space, instead of adding a classification layer to the model.
- We introduce a new scheme to tune the gradient update in the stochastic gradient descent and back-propagation algorithms used for training our model. More specifically, we improve over the usual gradient averaging update by performing an adaptive mining of the most significant triplets, which leads to better embeddings.

We instantiate these contributions through a dual deep neural network. We thoroughly evaluate our model on Recipe1M [33], the only English large-scale cooking dataset available, and show its superior performances compared to the state of the art models.

The remaining of this paper is organized as follows. In Section 2, we present previous work related to computer-aided cooking and cross-modal retrieval. Then, we present in Section 3 our joint retrieval and classification model as well as our adaptive triplet mining scheme to train the model. We introduce the experimental protocol in Section 4. In Section 5 we experimentally validate our hypotheses and present quantitative and qualitative results highlighting our model effectiveness. Finally, we conclude and discuss perspectives.

2 RELATED WORK

2.1 Computational cooking

Cooking is one of the most fundamental human activities connected to various aspects of human life such as food, health, dietary, culinary art, and so on. This is more particularly perceived on social platforms in which people share recipes or their opinions about meals, also known as the *eat and tweet* or *food porn* phenomenon [2]. Users' needs give rise to smart cooking-oriented tasks that contribute towards the definition of computational cooking as a research field by itself [1]. Indeed, the research community is very active in investigating issues regarding food-related tasks, such as ingredient identification [7], recipe recommendation [10], or recipe popularity prediction [34]. A first line of work consists in leveraging the semantics behind recipe texts and images using deep learning approaches [33, 34]. The objective of such propositions is to generally align different modalities in a shared latent space to perform cross-modal retrieval or recommendation. For instance, Salvador et al. [33] introduce a dual neural network that aligns textual and visual representations under both the distributional hypothesis and classification constraints [33]. A second line of work [10, 26, 37] aims at exploiting additional information (e.g., calories, biological or economical factors) to bridge the gap between computational cooking and healthy issues. For instance, Kusmierczyk et al. [26] extend the Latent Dirichlet Algorithm (LDA) for combining recipe descriptions and nutritional related meta-data to mine latent recipe topics that could be exploited to predict nutritional values of meals.

These researches are boosted by the release of food-related datasets [6, 9, 11, 22]. As a first example, [9] proposes the Pittsburgh fast-food image dataset, containing 4,556 pictures of fast-food plates. In order to solve more complex tasks, other initiatives provide richer sets of images [6, 7, 13, 33, 38]. For example, [6] proposes the Food-101 dataset, containing around 101,000 images of 101 different categories. Additional meta-data information is also provided in the dataset of [5] which involves GPS data or nutritional values. More recently, two very large-scale food-related datasets, respectively in English and Japanese, are released by [33] and [13]. The Cookpad dataset [13] gathers more than 1 million of recipes described using structured information, such as recipe description, ingredients, and process steps as well as images. Salvador et al. [33] have collected a very large dataset with nearly 1 million recipes, with about associated 800,000 images. They also add extra information corresponding to recipe classes and show how this new semantic information may be helpful to improve deep cross-modal retrieval systems. To the best of our knowledge, this dataset [33] is

the only large-scale English one including a huge pre-processing step for cleaning and formatting information. The strength of this dataset is that it is composed of structured ingredients and instructions, images, and a large number of classes as well. Accordingly, we focus all our experimental evaluation on this dataset.

2.2 Cross-modal Retrieval

Cross-modal retrieval aims at retrieving relevant items that are of di erent nature with respect to the query format; for example when querying an image dataset with keywords (image vs. text) [28]. The main challenge is to measure the similarity between di erent modalities of data. In the Information Retrieval (IR) community, early work have circumvented this issue by annotating images to perceive their underlying semantics [19, 36]. However, these approaches generally require a supervision from users to annotate at least a small sample of images. An unsupervised solution has emerged from the deep learning community which consists in mapping images and texts into a shared latent space F in which they can be compared [40]. In order to align the text and image manifolds in F, the most popular strategies are based either on 1) global alignment methods aiming at mapping each modal manifold in F such that semantically similar regions share the same directions in F; 2) local metric learning approaches aiming at mapping each modal manifold such that semantically similar items have a short distances in F.

In the first category of works dealing with global alignment methods, a well-known state-of-the-art model is provided by the Canonical Correlation Analysis (CCA) [17] which aims at maximizing the correlation in F between relevant pairs from data of di erent modalities. CCA and its variations like Kernel-CCA [4, 28] and Deep-CCA [3] have been successfully applied to align text and images [42]. However, global alignment strategies such as CCA do not take into account dissimilar pairs, and thus tend to produce false positives in cross-modal retrieval tasks.

In the second category of work, local metric learning approaches consider cross-modal retrieval as a ranking problem where items are ranked according to their distance to the query in the latent space. A perfect retrieval corresponds to a set of inequalities in which the distances between the query and relevant items are smaller than the distances between the query and irrelevant items [29, 39, 41]. Each modal projection is then learned so as to minimize a loss function that measures the cost of violating each of these inequalities [20, 24]. In particular, [12, 33] consider a loss function that minimizes the distance between pairs of matching cross-modal items while maximizing the distance between non-matching cross-modal pairs. However, ranking inequality constraints are more naturally expressed by considering triplets composed of a query, a relevant item, and an irrelevant item. This strategy is similar to the Large Margin Nearest Neighbor loss [39] in which a penalty is computed only if the distance between the query and the relevant item is larger than the distance between the query and the irrelevant item.

Our contributions di er from previous work according to three main aspects. First, we propose to model the manifold-alignment issue, which is generally based only on the semantic information [4, 17, 28, 42], as a joint learning framework leveraging retrieval and class-based features. In contrast to [33] which adds an additional classi cation layer to a manifold-alignment neural model, we directly integrate semantic information in the loss to re ne the structure of the latent space while also limiting the number of parameters to be learned. Second, our model relies on a double-triplet (instead of pairwise learning in [12, 33] or a single triplet as in [39]) to t with the joint learning objectives. Third, we propose a new stochastic gradient descent weighting scheme adapted to such a dual deep embedding architecture, which is computed on minibatches and automatically performs an adaptive mining of informative triplets.

Figure 2: AdaMine overview.

3 ADAMINE DEEP LEARNING MODEL

3.1 Model Overview

The objective of our model AdaMine (ADAptive MINing Embeding) is to learn the representation of recipe items (texts and images) through a joint retrieval and classi cation learning framework based on a double-triplet learning scheme. More particularly, our model relies on the following hypotheses:

H1: Aligning items according to a retrieval task allows capturing the ne-grained semantics of items, since the obtained embeddings must rank individual items with respect to each other.

H2: Aligning items according to class meta-data allows capturing the high-level semantic information underlying items since it ensures the identi cation of item clusters that correspond to class-based meta-data.

H3: Learning simultaneously retrieval and class-based features allows enforcing a multi-scale structure within the latent space, which covers all aspects of item semantics. In addition, we conjecture that adding a classi cation layer sequentially to manifold-alignment as in [33] might be under-e ective.

Based on these hypotheses, we propose to learn the latent space structure (and item embeddings) by integrating both retrieval objective and semantic information in a single cross-modal metric learning problem (see the Latent Space in Figure 2). We take inspiration from the learning-to-rank retrieval framework by building a learning schema based on query/relevant item/irrelevant item triplets noted $(x_q; x_p; x_n)$. Following hypothesis H3, we propose a double-triplet learning scheme that relies on both instance-based and semantic-based triplets, noted respectively $(x_q; x_p; x_n)$ and $(x_q^0; x_p^0; x_n^0)$, in order to satisfy the multi-level structure (ne-grained and high-level) underlying semantics. More particularly,

we learn item embeddings by minimizing the following objective function:

$$L_{total}(\theta) = L_{ins}(\theta) + \lambda L_{sem}(\theta) \quad (1)$$

where θ is the network parameter set, L_{ins} is the loss associated with the retrieval task over instance-based triplets (x_q, x_p, x_n), and L_{sem} is the loss coming with the semantic information over semantic-based triplets (x'_q, x'_p, x'_n). Unlike [33] that expresses this second term L_{sem} acting as a regularization over the L_{ins} optimization, in our framework, it is expressed as a joint classi cation task.

This double-triplet learning framework is a di cult learning problem since the trade-o between L_{ins} and L_{sem} is not only in uenced by λ but also by the sampling of instance-based and semantic-based triplets and depends on their natural distribution. Furthermore, the sampling of violating triplets can be di cult as the training progresses which usually leads to vanishing gradient problems that are common in triplet-based losses, and are ampli- ed by our double-triplet framework. To alleviate these problems, we propose an adaptive sampling strategy that normalizes each loss allowing to fully control the trade-o with λ alone while also ensuring non-vanishing gradients throughout the learning process.

In the following, we present the network architecture, each component of our learning framework, and then discuss the learning scheme of our model.

3.2 Multi-modal Learning Framework

3.2.1 Network Architecture. Our network architecture is based on the proposal of [33], which consists of two branches based on deep neural networks that map each modality (image or text recipe) into a common representation space, where they can be compared. Our global architecture is depicted in Figure 2.

The image branch (top-right part of Figure 2) is composed of a ResNet-50 model [15]. It contains 50 convolutional layers, totaling more than 25 million parameters. This architecture is further detailed in [15], and was chosen in order to obtain comparable results to [33] by sharing a similar setup. The ResNet-50 is pretrained on the large-scale dataset of the ImageNet Large Scale Visual Recognition Challenge [32], containing 1.2 million images, and is ne-tuned with the whole architecture. This neural network is followed by a fully connected layer, which maps the outputs of the ResNet-50 into the latent space, and is trained from scratch.

In the recipe branch (top-left part of Figure 2), ingredients and instructions are rst embedded separately, and their obtained representations are then concatenated as input of a fully connected layer that maps the recipe features into the latent space. For ingredients, we use a bidirectional LSTM [16] on pretrained embeddings obtained with the word2vec algorithm [31]. With the objective to consider the di erent granularity levels of the instruction text, we use a hierarchical LSTM in which the word-level is pretrained using the skip-thought technique [25] and is not ne-tuned while the sentence-level is learned from scratch.

3.2.2 Retrieval loss. The objective of the retrieval loss L_{ins} is to learn item embeddings by constraining the latent space according to the following assumptions (Hypothesis H1): 1) ranking items according to a similarity metric so as to gather matching items together and 2) discriminating irrelevant ones. We propose to use a loss function ℓ_{ins} based on a particular triplet (x_q, x_p, x_n) consisting of a query x_q, its matching counterpart in the other modality x_p and a dissimilar item x_n. The retrieval loss function L_{ins} is the aggregation of the individual losses ℓ_{ins} over all triplets. The aim of ℓ_{ins} is to provide a ne-grained structure to the latent space where the nearest item from the other modality with respect to the query is optimized to be its matching pair. More formally, the individual retrieval loss $\ell_{ins}(\theta; x_q, x_p, x_n)$ is formalized as follows:

$$\ell_{ins}(\theta; x_q, x_p, x_n) = \left[d(x_q, x_p) + \alpha - d(x_q, x_n) \right]_+ \quad (2)$$

where $d(x, y)$ expresses the cosine distance between vectors x and y in the latent space F.

3.2.3 Semantic loss. L_{sem} is acting as a regularizer capable of taking advantage of semantic information in the multi-modal alignment, without adding extra parameters to the architecture nor graph dependencies. To leverage class information (Hypotheses H2), we propose to construct triplets that optimize a surrogate of the k-nearest neighbor classi cation task. Ideally, for a given query x_q, and its corresponding class $c(x_q)$, we want its associated closest sample $x_{?,q}$ in the feature space to respect $c(x_q) = c(x_{?,q})$. This enforces a semantic structure on the latent space by making sure that related dishes are closer to each other than to non-related ones. To achieve this, we propose the individual triplet loss ℓ_{sem}:

$$\ell_{sem}(\theta; x'_q, x'_p, x'_n) = \left[d(x'_q, x'_p) + \alpha - d(x'_q, x'_n) \right]_+ \quad (3)$$

where x'_p belongs to the set of items with the same semantic class $c(x'_q)$ as the query, and x'_n belongs to the set of items with di erent semantic classes than the one of the query.

Contrary to the classi cation machinery adopted by [33], ℓ_{sem} optimizes semantic relations directly in the latent space without changing the architecture of the neural network, as shown on Figure 1. This promotes a smoothing e ect on the space by encouraging instances of the same class to stay closer to each other.

3.3 Adaptive Learning Schema

As commonly used in Deep Learning, we use the stochastic gradient descent (SGD) algorithm which approximates the true gradient over mini-batches. The update term is generally computed by aggregation of the gradient using the *average* over all triplets in the mini-batch. However, this *average* strategy tends to produce a vanishing update with triplet losses. This is especially true towards the end of the learning phase, as the few active constraints are averaged with many zeros coming from the many inactive constraints. We believe this problems is ampli ed as the size of the training set grows. To tackle this issue, our proposed *adaptive* strategy considers an update term Δ_{adm} that takes into account informative triplets only (i.e., non-zero loss). More formally, given a mini-batch B, P^r_q the set of matching items with respect to a query x_q and P^s_q the set of items with the same class as x_q, the update term Δ_{adm} is de ned by:

$$\Delta_{adm} = \sum_{x_q \in B} \sum_{x_p \in B \setminus P^r_q} \sum_{x_n \in B \cap P^r_q} \frac{\nabla \ell_{ins}(\theta; x_q, x_p, x_n)}{\Delta^r_0} + \sum_{x_p \in B \setminus P^s_q} \sum_{x_n \in B \cap P^s_q} \frac{\nabla \ell_{sem}(\theta; x_q, x_p, x_n)}{\Delta^s_0} \quad (4)$$

with $\mathcal{L}_r^0$ and $\mathcal{L}_s^0$ being the number of triplets contributing to the cost:

$$\mathcal{L}_r^0 = \sum_{x_q \in B} \sum_{x_p \in B \setminus P_q^r} \sum_{x_n \in B \cap P_q^r} 1_{\grave{}_{ins}, 0}$$
$$\mathcal{L}_s^0 = \sum_{x_q \in B} \sum_{x_p \in B \setminus P_q^s} \sum_{x_n \in B \cap P_q^s} 1_{\grave{}_{sem}, 0} \quad (5)$$

At the very beginning of the optimization, all triplets contribute to the cost and, as constraints stop being violated, they are dropped. At the end of the training phase, most of the triplets will have no contribution, leaving the hardest negatives to be optimized without vanishing gradient issues. Remark that this corresponds to a curriculum learning starting with the average strategy and ending with the hard negative strategy like in [35], but without the burden of ending the time-step at which to switch between strategies as this is automatically controlled by the weights $\mathcal{L}_r$ and $\mathcal{L}_s$.

Remark also that an added benefit of $\mathcal{L}_{adm}$ is due to the independent normalization of each loss by its number of active triplets. Thus $\mathcal{L}_{adm}$ keeps the trade-off between $\grave{}_{ins}$ and $\grave{}_{sem}$ unaffected by difference between the number of active triplets in each loss and allows $\mathcal{L}$ to be the only effective control parameter.

4 EVALUATION PROTOCOL

The objective of our evaluation is threefold: 1) Analyzing the impact of our semantic loss that directly integrates semantic information in the latent space; 2) Testing the effectiveness of our model; 3) Exploring the potential of our model and its learned latent space for solving smart cooking tasks. All of our experiments are conducted using PyTorch[2], with our own implementation[3] of the experimental setup (i.e. preprocessing, architecture and evaluation procedures) described by [33]. We detail the experimental setup in the following.

4.1 Dataset

We use the Recipe1M dataset [33], the only large-scale dataset including both English cooking recipes (ingredients and instructions), images, and categories. The raw Recipe1M dataset consists of about 1 million image and recipe pairs. It is currently the largest one in English, including twice as many recipes as [27] and eight times as many images as [7]. Furthermore, the availability of semantic information makes it particularly suited to validate our model: around half of the pairs are associated with a class, among 1048 classes parsed from the recipe titles. Using the same preprocessed pairs of recipe-image provided by [33], we end up with 238,399 matching pairs of images and recipes for the training set, while the validation and test sets have 51,119 and 51,303 matching pairs, respectively.

4.2 Evaluation Methodology

We carry out a cross-modal retrieval task following the process described in [33]. Specifically, we first sample 10 unique subsets of 1,000 (1k setup) or 5 unique subsets of 10,000 (10k setup) matching text recipe-image pairs in the test set. Then, we consider each item in a modality as a query (for instance, an image), and we rank items in the other modality (resp. text recipes) according to the cosine distance between the query embedding and the candidate embeddings. The objective is to retrieve the associated item in the other modality at the first rank. The retrieved lists are evaluated using standard metrics in cross-modal retrieval tasks. For each subset (1k and 10k), we estimate the median retrieval rank (MedR), as well as the recall percentage at top K (R@K), over all queries in a modality. The R@K corresponds to the percentage of queries for which the matching item is ranked among the top K closest results.

4.3 Baselines

To test the effectiveness of our model AdaMine, we evaluate our multi-modal embeddings with respect to those obtained by state-of-the-art (SOTA) baselines:

CCA, which denotes the Canonical Correlation Analysis method [17]. This baseline allows testing the effectiveness of global alignment methods.

PWC, the pairwise loss with the classification layer from [33]. We report their state-of-the-art results for the 1k and 10k setups when available. This baseline exploits the classification task as a regularization of embedding learning.

PWC*, our implementation of the architecture and loss described by [33]. The goal of this baseline is to assess the results of its improved version PWC++, described below.

PWC++, the improved version of our implementation PWC*. More particularly, we add a positive margin to the pairwise loss adopted in [33], as proposed by [18]:

$$\grave{}_{pw++}\,{}^{1}\,;x_q;x^o = y\ d^1x_q;x^o\ \ {}_{pos}\ {}_{+} + {}^{1}1\ \ y^o\ \ {}_{neg}\ \ d^1x_q;x^o\ {}_{+} \quad (6)$$

with $y = 1$ (resp. $y = 0$) for pos. (resp. neg.) pairs. The positive margin $\mathcal{L}_{pos}$ allows matching pairs to have different representations, thus reducing the risk of overfitting. In practice, the positive margin is set to 0.3 and the negative margin to 0.9.

We evaluate the effectiveness of our model AdaMine, which includes both the triplet loss and the adaptive learning, in different setups, and having the following objectives:

Evaluating the impact of the retrieval loss: we run the AdaMine_ins scenario which refers to our model with the instance loss $\mathcal{L}_{ins}$ only and the adaptive learning strategy (the semantic loss $\mathcal{L}_{sem}$ is discarded);

Evaluating the impact of the semantic loss: we run the AdaMine_sem scenario which refers to our model with the semantic loss $\mathcal{L}_{sem}$ only and the adaptive learning strategy (the instance loss $\mathcal{L}_{ins}$ is discarded);

Evaluating the impact of the strategy used to tackle semantic information: we run the AdaMine_ins+cls scenario which refers to our AdaMine model by replacing the semantic loss by the classification head proposed by [33];

Measuring the impact of our adaptive learning strategy: we run the AdaMine_avg. The architecture and the losses are identical to our proposal, but instead of using the adaptive learning strategy, this one performs the stochastic gradient descent averaging the gradient over all triplets, as is common practice in the literature;

Evaluating the impact of the text structure: we run our whole model (retrieval and semantic losses + adaptive SGD) by considering either ingredients only (noted AdaMine_ingr) or instructions only (noted AdaMine_instr).

[2] http://pytorch.org
[3] https://github.com/Cadene/recipe1m.bootstrap.pytorch