Dividing and Conquering Cross-Modal Recipe Retrieval: from Nearest Neighbours Baselines to SoTA

Mikhail Fain, Niall Twomey, Andrey Ponikar, Ryan Fox, and Danushka Bollegala

Abstract—We propose a novel non-parametric method for cross-modal recipe retrieval which is applied on top of precomputed image and text embeddings. By combining our method with standard approaches for building image and text encoders, trained independently with a self-supervised classification objective, we create a baseline model which outperforms most existing methods on a challenging image-to-recipe task. We also use our method for comparing image and text encoders trained using different modern approaches, thus addressing the issues hindering the development of novel methods for cross-modal recipe retrieval. We demonstrate how to use the insights from model comparison and extend our baseline model with standard triplet loss that improves state-of-the-art on the Recipe1M dataset by a large margin, while using only precomputed features and with much less complexity than existing methods. Further, our approach readily generalizes beyond recipe retrieval to other challenging domains, achieving state-of-the-art performance on Politics and GoodNews cross-modal retrieval tasks.

Index Terms—Cross-modal retrieval, baselines, nearest neighbours.

1 Introduction

In this work we are exploring the problem of cross-modal recipe retrieval between food images and textual cooking recipes. A solution to this problem has a number of applications, such as searching for the correct recipe using a photo [1], automatically determining the number of calories in a dish [2] and improving the performance of various recipe recommendation and ranking systems [3]. This task involves searching for an exact matching text of the recipe given its image among candidate textual recipes from a held-out test set.

The problem of recipe retrieval is challenging due to the diverse nature of food images and the subtle differences between recipes. For example, the photos of dishes made by following the same recipe could look completely different from each other, and very similar dish photos could be associated with very different ingredients and procedures[1].

After the release of the Recipe1M dataset [4] containing diverse food images and recipes, cross-modal recipe retrieval became one of the standard benchmarks for image-to-document retrieval tasks, with numerous improvements made on the original model [4]. Despite this progress, we observe a few issues that, if not addressed, may hinder our understanding of added value delivered by new methods. Namely, these issues are: the lack of strong baselines and the complexity of identifying strengths and weaknesses of individual model components across methods.

Problem 1: *lack of strong baselines*. In the seminal paper on cross-modal recipe retrieval, Salvador *et al.* [4] presented a baseline model using a classic statistical models for learning joint embeddings, Canonical Correlation Analysis (CCA) [5]. This model achieved the top-1 accuracy of 14.0 on a test set of 1,000 recipes. A deep learning model presented in the same paper almost doubled the top-1 accuracy reaching 25.6. By 2019, after two years of steady improvements [6], [7], [8], [9], this metric was doubled again to reach 51.8 [10]. While this could reflect the improvements in retrieval methods, it could also be due to mis-specified baselines in this domain. If the latter is the case, a strong and simple baseline for cross-modal retrieval could facilitate faster scientific progress and put the previously reported results into perspective.

Problem 2: *challenges in identifying model's components strengths*. Cross-modal recipe retrieval models are very complex and have many interacting parts [10]. Thus the performance gains are not easy to attribute to improvements in a particular component (e.g. the image or text encoder). This makes it extremely hard to understand and rank the various elements in the model, as ablation studies only address performance gains within the same method. The prior work on cross-modal recipe retrieval [7], [8], [9], [10] mitigates this issue by using an image and textual processing pipeline that is similar to the one introduced by Salvador *et al.* [4]. However, the setups still sufficiently differ to prevent a fair comparison [7], [8], and, in addition, being tied to specific architectures may limit the scope of future research.

We introduce Cross-Modal k-Nearest-Neighbours (CkNN) in Section 3 to address the aforementioned problems. We use k-Nearest-Neighbours (kNN) to search over modalities using the correspondences available in the training set. Our results are presented in Section 4: specifically, we apply CkNN to challenging non-literal image-to-text retrieval tasks by leveraging standard approaches for independently representing images and text

- *M. Fain, N. Twomey and A. Ponikar are with Cookpad Ltd. R. Fox is with Facebook (his contributions here were made prior to joining Facebook), and D. Bollegala is with the University of Liverpool.*
E-mail: andrey-ponikar@cookpad.com

[1] As an example, consider visually distinguishing different types of creamy soups from each other using only photos

using a self-supervised classification objective. This is a solution for PROBLEM 1 since we define a straightforward competitive baseline on cross-modal recipe retrieval task. Since CkNN depends directly on distance measures across different modalities, we use it for comparing the efficacy of image and text encoders, addressing PROBLEM 2. We demonstrate how to use the insights from encoder comparison and go beyond our baseline results to improve state-of-the-art (SoTA) by a large margin on RECIPE1M while still using only precomputed features and standard approaches. We further show that our approach generalizes to other challenging cross-modal retrieval tasks, reaching SoTA performance on POLITICS [11] and GOODNEWS [12] datasets.

The combination of our contributions raises the bar for baseline models and model component analysis in cross-modal retrieval and sets a new SoTA reference performance on RECIPE1M. We hope that our approach would encourage further development of advanced end-to-end methods.

2 RELATED WORK

The problem of cross-modal retrieval has been researched extensively in the Computer Vision community with a primary focus on datasets where there exists a clear mapping between objects in the image scene and a concise textual description of the image [13]. These tasks are exemplified by the standard FLICKR30K [14] and MS-COCO [15] benchmarks. The majority of the solutions create separate representations for the two different modalities, projecting them to the same shared space and performing a similarity search within that space [16], [17]. Such models are typically based on neural networks and are trained end-to-end [18], [19], [20]. The recently proposed methods exploited semantic category labels to learn discriminative features for cross-modal retrieval [21], [22], [23]. Adversarial learning [24] has also been employed to aid cross-modal retrieval [25]. Further work shows the benefits of applying attention on top of the object detection pipeline to capture fine-grained relationships between vision and language, creating a better aligned joint embedding space [26]. Wang *et al.* [27] extend these approaches to move away from the shared embedding space completely.

Some of the above ideas could be adapted successfully to cross-modal *recipe* retrieval task, which is a subproblem of the general cross-modal retrieval problem featuring a subtle image-text relationship. In this case, one modality is a recipe image, and the second one is a structured text, consisting of a recipe title, a list of ingredients in free form and a list of instructions, also written in free form. It was introduced by Salvador *et al.* [4], who used margin loss for learning the shared embedding space. The image processing pipeline was based on ResNet-50 [28]. Recipe ingredients were normalized using a separate model involving bi-directional LSTM [29] and further encoded with word2vec [30]. The list of instructions was encoded using skip-thoughts [31]. The encoded ingredients and instructions were then passed through to separate LSTMs and concatenated, thus generating the encoding of the recipe text. The resulting model was trained end-to-end (except for word2vec and skip-thoughts vectors which were pretrained separately), improving the top-1 accuracy of **CCA** [5] baseline from 14.0 to 25.6 on a test set of size 1,000 [4]. This model is further referred to as Pic2Recipe.

The follow-up work has been largely focused on expanding on the above setup, with the focus on improving cross-modal alignment techniques and minor changes in individual modality processing pipelines and training methods. For example, Chen *et al.* [6], [32] analyzed the importance of instructions and ingredients for cross-modal retrieval and then built a text representation designed to match Pic2Recipe performance [7], despite relying on a Convolutional Neural Network (CNN) pretrained on another food image dataset rather than learning it end-to-end. This model is denoted in this paper as AM following [10].

Carvalho *et al.* [8] made improvements to the alignment loss function using double-triplet loss in their AdaMine model, improving the top-1 accuracy to 39.8. Zhu *et al.* [9] in their R2GAN method employed Generative Adversarial Networks (GANs) [24] to help with learning the representations and reaching results similar to AdaMine. MCEN method uses stochastic latent variable model to share the information between modalities [33]. The current SoTA, ACME [10], also uses GANs in addition to a cross-modal triplet loss scheme [22] together with an effective sampling strategy [34], modality alignment using an adversarial learning strategy from [22] and a cross-modal translation consistency loss to reach an impressive 51.8 top-1 accuracy, more than doubling the original performance of Pic2Recipe. We note that we were unable to reproduce the reported performance of ACME model with the model weights released by the authors, and refer to the model achieving the slightly lower score as ACME*.

Most of the described cross-modal recipe retrieval models reported significant benefits from using a *semantic regularization* technique [4], where the image and text embeddings are constrained by an additional classification loss with the labels being the categories of the recipes.

On the related topic of building powerful food image classifiers there has been an independent body of work focused around food image datasets [35], [36], [37], [38]. The researchers have explored a variety of architectures more suitable to food images than the ResNet-50 backbone used in cross-modal retrieval [39].

For recipe text encoding, the body of literature is less organized. As there are no commonly used benchmarks for evaluating recipe text representations, the models are usually tuned as part of a bigger task, such as cross-modal retrieval [4], recipe translation [40] or ingredient pairing [41].

In addition to recipe retrieval, there exist other cross-modal retrieval domains where the relationship between text and images are more subtle than the standard benchmarks like Flickr30k and MS-COCO. Thomas and Kovashka [13] name these types of tasks as *non-literal* cross-modal retrieval and show that enforcing neighbourhood consistency between the image and text spaces during training is beneficial to performance, with their best model denoted SN outperforming the baselines on POLITICS [11] and GOODNEWS [12] datasets.

Nearest Neighbour Search has been shown to outperform many more complex deep learning methods on

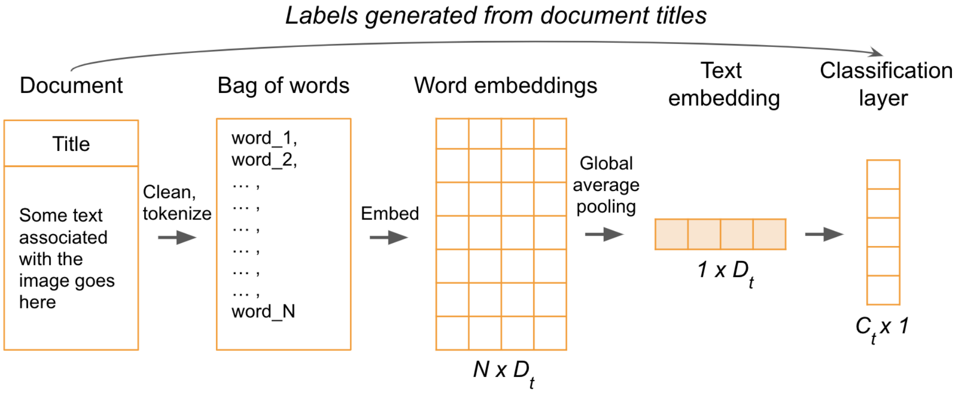

Fig. 1. Average word embeddings encoder.

neural recommendation tasks [42]. It has also been used successfully for fine-grained image retrieval as a base for many query expansion [43] and database augmentation [44] strategies, which work on top of other methods and yield significant gains on image retrieval tasks. [2]. However, to the best of our knowledge, these ideas have not been extended for cross-modal retrieval.

3 Method

Since the main purpose of this work is to create strong baselines and compare encoders rather than aim for the best possible method of cross-modal retrieval, we do not train our model end-to-end unlike existing approaches [4], [8], [10]. Instead, we adopt a widely-used approach for independently training encoders using a self-supervised classification objective [45], [46], which we setup by automatically extracting noisy labels from text. Our encoders employ basic architectures: the last layer of a CNN for images [47], [48] and the average of word embeddings (AWE) for text. [49], [50], [51]. We describe the self-supervised task for the text and image modalities respectively in Sections 3.1 and 3.2, and Section 3.3 introduces the manner in which they are integrated.

3.1 Average Word Embeddings Encoder

We use a standard average word embeddings [49], [50], [51], [52] model as a text encoder. This model assigns an embedding in D_t dimensions to each word in the input document, and computes the document representation by averaging the embeddings. Note, that this is one of the most basic approaches in the literature with the structure of the document being completely ignored by such representation.

To train the word embeddings, we use text data to define a self-supervised task. We follow the approach of [4] and extract a set of C_t noisy labels based on frequent unigrams and bigrams from document title and filter them by a threshold. The remaining text is treated as features which are used to learn a mapping to the label space. Therefore, we use the main body (ingredients and instruction) of the textual recipe as training input data, and the title to extract labels.

The details of our training setup could be described as follows. The encoder input is treated as one long, continuous document. We then add a trainable, randomly initialized word embedding layer with $D_t = 300$ dimensions. The next layer computes the mean of all the embeddings, a linear layer with sigmoid activations is added on top for multi-label classification. Binary cross-entropy is used as a loss function. See full architecture in Fig.1. We refer to this model as the *AWE-Encoder*.

While this self-supervised task is applicable for textual recipes, it does not apply to datasets where the documents do not have identifiable title (such as in GoodNews dataset we explore in Section 4.5, which consists of short captions paired with images). In this case, we train word embeddings using unsupervised FastText [53] method instead of the classification objective.

3.2 Image Encoder

Similar to AWE-Encoder, our image encoder is trained using the set of C_i noisy labels extracted from text. Image representations are produced using ResNet-50 [28], and we apply binary cross-entropy loss for multi-label classification. To create the image embeddings, we extract features from the last convolutional layer of the network [47], [48] after global average pooling. We refer to this model as *ResNet-Encoder*.

3.3 Cross-Modal k-Nearest-Neighbours (CkNN)

CkNN belongs to the category of alignment modules, which attempt to match the representations of different modalities. It is applied on top of an image encoder e_i with a distance measure $d_i(\cdot,\cdot)$ in the image embedding space, and a text encoder e_t with a distance measure $d_t(\cdot,\cdot)$ in the text embedding space. In this work we use cosine similarity as a distance measure for both the text and image embedding spaces.

CkNN, depicted in Fig.2, uses the training data to represent a candidate text T in the image embedding space using the following algorithm denoted as $\mathrm{CkNN}_i(T)$:

1) Encode a candidate text document T using $e_t(T)$.
2) Find the k_t nearest neighbours based on the text embeddings using d_t, denoted $\mathcal{R}_T$.
3) Extract the set of images $\mathcal{I}_T$ associated with $\mathcal{R}_T$.

[2] https://landmarksworkshop.github.io/CVPRW2019/

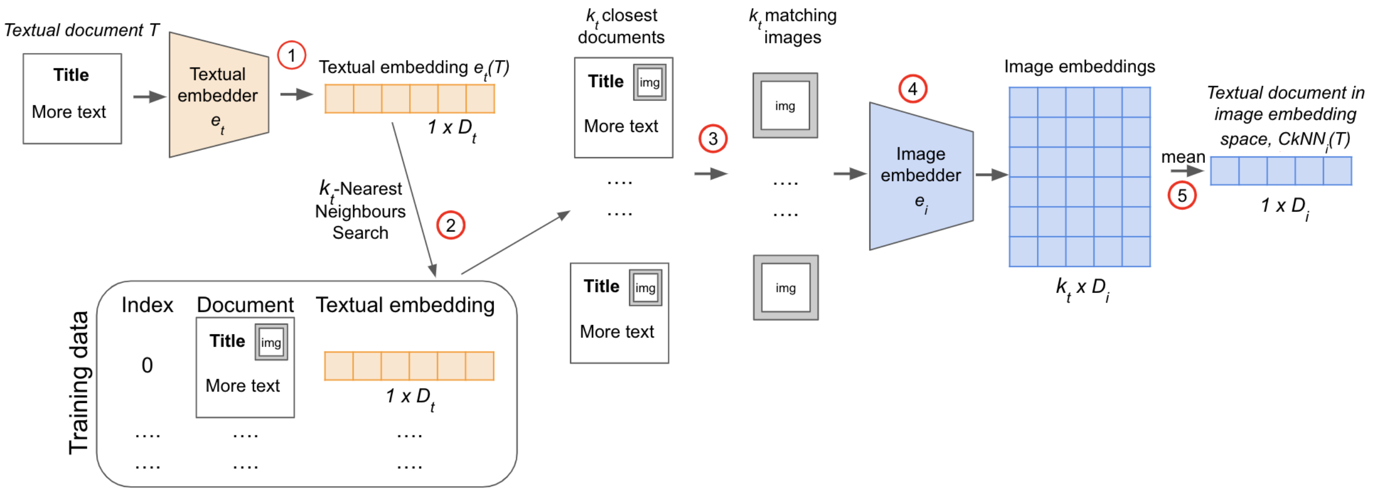

Fig. 2. A schematic representation of the core idea of our CkNN method: representing a text T in the image embedding space using nearest neighbour search in the training data. Orange and blue shapes respectively denote text and image modalities. Numbers in red circles correspond to $\mathrm{CkNN}_i(T)$ algorithm steps described in Section 3.3

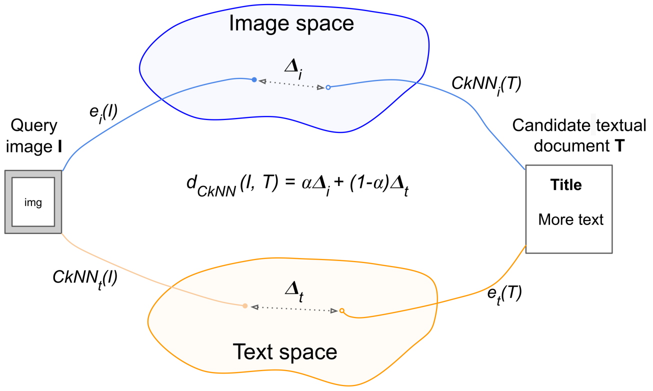

Fig. 3. CkNN distance $d_{\mathrm{CkNN}}(I, T)$ between a query image I and a candidate textual recipe T to be used in ranking for the cross-modal retrieval task

4) Encode each $I \in \mathcal{I}_T$ in the image embedding space using $e_i(I)$.
5) Return the mean vector $\frac{1}{|\mathcal{I}_T|} \sum_{I \in \mathcal{I}_T} e_i(I)$.

The complementary algorithm ($\mathrm{CkNN}_t(I)$) produces text embeddings for an image query, I, allowing a separate neighbourhood size, k_i. Thus, we now have two ways to calculate the distance between a query image I and a candidate text T: (a) in the image embedding space, and (b) in the text embedding space. We define the total distance between an image and a document as the linear combination of the two shown in Eq. (1) and illustrated in Fig.3. This distance can be used to rank the set of candidates based on a particular query.

$$\begin{aligned} d_{\mathrm{CkNN}}(I, T) = & d_i(e_i(I), \mathrm{CkNN}_i(T))\, \alpha + \\ & d_t(\mathrm{CkNN}_t(I), e_t(T))\, (1 - \alpha) \end{aligned} \tag{1}$$

We emphasize that we developed the non-parametric CkNN method for the purpose of creating cross-modal retrieval baselines rather than achieving state-of-the-art results. Further, the encoders from Sections 3.2 and 3.1 are trained using basic methods, and the resulting embeddings are not designed to be used for retrieval using cosine similarity. We refrain from applying more advanced methods for representation learning [45] as well as metric learning [54] for our modelling components to keep the focus on baselines and standard approaches.

4 Experiments and Results

4.1 Recipe1M Dataset and Metrics

The Recipe1M dataset [4] consists of over 1M textual cooking recipes (including recipe title, list of ingredients in free form, and list of cooking instructions). In addition, 402,760 of those recipes are linked to one or more corresponding images (887,706 images in total). The dataset is split into dedicated training, validation and test sets.

In our experiments we use the standard metrics on image-to-textual-recipe retrieval task for the Recipe1M dataset as proposed by [4]:

- Recall at Top K (R@K) for $K = 1, 5, 10$ describes the percentage of images for which the correct textual recipe is among the top K of the ranked results list (higher is better).
- Median Rank (medR) is the median of the rank of the correct textual recipes across all images (lower is better).

The cross-modal retrieval metrics are calculated in the same way as in the prior work. Namely, we randomly sample $N = 1,000$ (1K setup) or $N = 10,000$ (10K setup) recipes from the test set, and for each test image, ranking the textual recipes in the sample using the given model.

The metrics above are noisy due to randomization, so we follow the strategy proposed in [4] and sample the sets 10 times, reporting the average results. It should be noted that the results are still too noisy in the case of $N = 1,000$ due to sampling errors [8], thus in this work we focus on metrics for $N = 10,000$ (although we also report the $N = 1,000$ performance for comparison against previously published results in Section 4.4).

TABLE 1
Performance of our baseline method and other reported methods on RECIPE1M image-to-textual-recipe task. The best results are shown in bold. The description of the existing methods could be found in Section 2. The metrics for the models denoted by * are computed by us using publicly available models. While we report the results for both 1K and 10K test sample size, we only rely on 10K values for our analysis since 1K results are too noisy [8]

Size of test set	Method	medR ↓	R@1 ↑	R@5 ↑	R@10 ↑
1K	CCA [5]	15.7	14.0	32.0	43.0
	Pic2Recipe [4]	5.2	25.6	51.0	65.0
	AM [7]	4.6	25.6	53.7	66.9
	AdaMine [8]	2.0	39.8	69.0	77.4
	R2GAN [9]	2.0	39.1	71.0	81.7
	ACME [10]	**1.0**	**51.8**	**80.2**	**87.5**
	ACME* [10]	1.8	49.0	77.1	85.2
	MCEN [33]	2.0	48.2	75.8	83.6
	(Ours) CkNN+AWE+ResNet	2.0	45.7	75.9	84.2
10K	Pic2Recipe* [4]	39	7.3	20.3	29.0
	AM [7]	39.8	7.2	19.2	27.6
	AdaMine [8]	13.2	14.9	35.3	45.2
	R2GAN [9]	13.9	13.5	33.5	44.9
	ACME [10]	**6.7**	**22.9**	**46.8**	**57.9**
	ACME* [10]	7.5	20.6	44.3	55.7
	MCEN [33]	7.2	20.3	43.3	54.4
	(Ours) CkNN+AWE+ResNet	9.1	19.1	41.3	52.5

To focus on one metric for clarity in our evaluation section, we select the recall at rank-1 with the sample size of 10,000 (**10K-R@1**), but we note that the results are consistent for all of the metrics. Throughout the paper, we only use the Recipe1M validation set for validation purposes, and report the results on the test set.

4.2 Nearest Neighbour Baselines

For RECIPE1M, we used the recipe title to extract $C_t = 3453$ labels for training AWE-Encoder and treated recipe text as flat documents. We used the recipe title and ingredient list to extract $C_i = 5036$ labels for training ResNet-Encoder.

We then used CkNN to align the modalities. Using the validation split, we found that $\alpha = 0.1,\ k_t = 15,\ k_i = 3$ was suitable in all cases. ResNet-Encoder was trained for 40 epochs with Adam optimizer [55], a batch size of 512 and an initial learning rate of 0.0001. AWE-Encoder was trained for 15 epochs with Adam optimizer, a batch size of 128 and an initial learning rate of 0.002. The training hyperparameters were manually tuned on validation data.

We report the performance of our proposed baseline method (Section 3) on RECIPE1M in Table 1. We can see that on RECIPE1M, we outperform the majority of the competing methods, even though these models employ sophisticated modelling components.

4.3 Analysis: Comparison of RECIPE1M Image and Text Models

Since CkNN modality alignment fully decouples the image and text encoders and depends directly on image and text distance measures d_i and d_t (Eq. 1), the performance on cross-modal retrieval could indicate how good the individual modality embeddings are for retrieval purposes. As observed in Section 4.2, the performance of CkNN is relatively robust against a wide range of hyperparameter values. This, along with the fact that CkNN does not require training, also contributes to CkNN being a suitable choice for comparing encoders.

We thus compare encoders pretrained with different methods by applying CkNN to all combinations of image and text encoders and report $10K - R@1$ metric on the RECIPE1M test set [3] since it is the most studied of the large datasets we consider. For example, we take the image encoder trained jointly as part of SoTA ACME [10] model and use it in combination with the text encoder trained jointly as part of Pic2Recipe [4]. To study this, we first require pre-trained image and text models:

4.3.1 Image Models

For comparison purposes, we used the existing image encoders trained for cross-modal recipe retrieval tasks for which we could run the inference code, which are Pic2Recipe [4], AdaMine [8] and ACME* [10].

We further used the following baseline image encoders pretrained on different public domain datasets to extract the embeddings from their last convolutional layer: ImageNet-Pretrained pretrained on ImageNet [56] and Food-Pretrained pretrained on the concatenation of Food-101 [35], ChineseFoodNet [36] and iFood-2018[4] datasets. We also used ResNet-Encoder model (Section 4.2), as well as a random embeddings model denoted as Random.

4.3.2 Text Models

The only two publicly available textual recipe encoders that we could run at the moment of writing are **ACME*** and **Pic2Recipe**, described in detail in Section 2. Despite our best efforts, we did not manage to run the code for the AdaMine text encoder.

[3] Here we use only such training recipes for CkNN for which the embeddings are available for all the encoders being compared. This accounts for the small discrepancy with numbers reported in Section 4.4.

[4] https://github.com/karansikka1/Foodx

TABLE 2
10K-R@1 metric on the RECIPE1M image-to-textual-recipe task, computed for various image and text encoders combined using our CkNN approach (higher is better). All the image models are built on ResNet-50 [28] backbone. Although the results are not optimal, they are competitive with published methods [10], and allow for direct comparisons between different image and textual recipe embeddings.

Text Model Image Model	ACME*	Pic2Recipe	TF-IDF	AWE-Encoder	Random
ACME*	17.9	13.3	10.6	15.6	0.01
Pic2Recipe	8.5	7.1	5.3	7.5	0.01
AdaMine	12.6	9.9	7.5	11.2	0.01
ImageNet-Pretrained	4.4	3.4	3.8	5.0	0.01
Food-Pretrained	7.8	6.2	6.5	8.6	0.01
ResNet-Encoder	16.6	13.1	12.0	17.4	0.01
Random	0.01	0.01	0.01	0.01	0.01

As a baseline unsupervised encoder we represented the textual recipes as a bag-of-subword-units with term frequency-inverse document frequency (TF-IDF) weights calculated on the RECIPE1M training set. We applied singular value decomposition on top of this representation, reducing the dimensionality of the embedding to $D_t = 2000$. The model is denoted as TF-IDF. We also used AWE-Encoder model described in Section 4.2, as well as a random embeddings model denoted as Random.

4.3.3 Comparison of Previously Published Model Components

The results of our experiments are shown in Table 2. The first observation is that the performance of CkNN is competitive with direct search in the embedding space for which the jointly trained encoders were optimized [4], [10]. Indeed, combining ACME* image and text encoders using CkNN drops the performance of direct search only from 20.6 to 17.9, which is still better than most previously published methods (Table 1). For Pic2Recipe, the drop is even smaller: from 7.3 to 7.1.

We further observe that the performance of individual encoders, image or text, is generally consistent across different combinations and in line with performance on the RECIPE1M dataset as reported in [10], validating our comparison framework using CkNN. ACME* (SoTA method) image encoder produces consistently higher numbers across all four text encoders than AdaMine (2nd best method) image encoder, and AdaMine is better than Pic2Recipe (3rd best method) across all metrics. ACME* text encoder outperforms Pic2Recipe text encoder across all 6 image encoders.

The two observations above indicate that modality alignment procedures used in the existing approaches add on the order of 10% improvement to 10K-R@1 metric compared to CkNN. At the same time, the performance differences due to replacing the encoders are around 50% from the 3rd to the 2nd best result, and from the 2nd best result to the SoTA result. This suggests that the large performance gap in the reported performance of existing cross-modal methods could be explained primarily by the strengths of the learned image and text embedding spaces for retrieval purposes, and not by the quality of cross-modal alignment.

4.3.4 Comparison of Independently Trained Encoders

We now compare the results obtained with the encoders trained jointly as part of the existing cross-modal methods and other text and image encoders, trained independently. We observe that the image encoders pretrained on external data combined with unsupervised TF-IDF encoder produce results on a par with some published metrics. Indeed, Food-Pretrained image encoder outperforms Pic2Recipe image encoder in combination with some of the text encoders, and its combination with TF-IDF scores close to direct search with Pic2Recipe. Even Imagenet-Pretrained with TF-IDF produces a reasonable score of 3.8 (a random model would yield 10K-R@1 score of 0.01). [5]. This shows that CkNN allows easy creation of many competitive baselines out of encoders trained in completely different ways.

Next, we analyze the performance of our proposed self-supervised encoders. Among image encoders, ResNet-Encoder performs close to ACME* and consistently outperforms other encoders. Among text encoders, AWE-Encoder reaches performance similar to ACME*, whereas Pic2Recipe and TF-IDF perform much worse. This suggests that independent training using a self-supervised classification objective can produce encoders competitive with those trained as part of the SoTA cross-modal retrieval methods. In addition, the success of a much simpler average word embeddings architecture compared to the complex ACME* and Pic2Recipe textual model architectures suggests that more research is needed to understand how to best represent textual recipes.

4.4 Beyond Baselines

In this section we consider generalisations on top of the base CkNN model. While ACME* image and text embeddings combination outperforms all others according to Table 2, we note that ResNet-Encoder and AWE-Encoder score is very close with 10K-R@1 of 17.9 and 17.4 respectively. In fact, CkNN combined with the embeddings precomputed using these encoders (*CkNN+AWE+ResNet* in Table 1) provides a strong cross-modal recipe retrieval baseline that improves upon a third-best published result [9] In this section, we show how to improve *CkNN+AWE+ResNet* baseline to achieve new SoTA, summarizing our results in Table 3.

4.4.1 Triplet Loss Alignment

Although we showed that CkNN is a suitable choice for model comparison and building cross-modal retrieval base-

[5] This combination also achieves 1K-R@1 of 15.2 on 1K test set, outperforming CCA baseline of 14.0 reported by [4]

TABLE 3
Performance of the extensions of our method on RECIPE1M image-to-textual-recipe task and the previous state-of-the-art method, ACME [10]. The best results are shown in bold, and are statistically significant. While we report the results for both 1K and 10K test sample size, we only rely on 10K values for our analysis since 1K results are too noisy [8].

Size of test set	Method	medR ↓	R@1 ↑	R@5 ↑	R@10 ↑
1K	(Previous SoTA) ACME [10]	1.0	51.8	80.2	87.5
	(Ours) CkNN+AWE+ResNet	2.0	45.7	75.9	84.2
	(Ours) CkNN+AWE+ResNext	1.3	50.5	79.5	86.7
	(Ours) Triplet+AWE+ResNet	1.0	55.9	82.4	88.7
	(Ours) Triplet+AWE+ResNext	**1.0**	**60.2**	**84.0**	**89.7**
10K	(Previous SoTA) ACME [10]	6.7	22.9	46.8	57.9
	(Ours) CkNN+AWE+ResNet	9.1	19.1	41.3	52.5
	(Ours) CkNN+AWE+ResNext	6.8	22.9	46.9	58.0
	(Ours) Triplet+AWE+ResNet	5.0	26.5	51.8	62.6
	(Ours) Triplet+AWE+ResNext	**4.0**	**30.0**	**56.5**	**67.0**

lines, there is still scope for improving alignment. As observed in Section 4.3, direct search through a jointly learned embedding space can surpass the results of our CkNN approach for ACME* and Pic2Recipe encoders. Thus, we also train a standard triplet loss alignment module to create a joint embedding space on top of the precomputed image and text embeddings.

In particular, we jointly train two feed-forward neural networks (FNN) with one hidden layer, dropout and batch normalization: one for image (g_i) and another for textual (g_t) features with triplet loss. Architectures of both neural networks are identical, and output feature size is $D = 1024$. This pipeline is depicted in Fig.4.

Each triplet consists of one feature embedding as an anchor point in image modality and a positive and negative feature embeddings from text modality. The positive instances are the different modalities of the same recipe, X_{ia} for image and X_{ta} for textual features. We use online negative instance mining [57] to choose the negative instance X_{tn} as the closest text instance to the anchor point selected from other recipes in the mini-batch. The objective $\mathcal{L}$ is given by Eq. (2).

$$\begin{aligned}\mathcal{L} = \max(0, & d(g_i(X_{ia}), g_t(X_{ta})) - \\ & d(g_i(X_{ia}), g_t(X_{tn})) + \gamma),\end{aligned} \tag{2}$$

where d is the cosine distance between two vectors and γ is the margin.

The model was trained on 238K recipes from the RECIPE1M training set, with the margin γ manually tuned to be 0.3 on 10K-R@1 metric on a validation set. We used Adam optimizer [55], a batch size of 256, initial learning rate of 0.002, and applied alternating optimization [4] to aid convergence. Training takes only 25 seconds per epoch on a Tesla M60 GPU. The hyperparameters were tuned on a validation set.

This model boosts 10K-R@1 metric to 26.5 on RECIPE1M, yielding a new SoTA result (*Triplet+AWE+ResNet* in Table 3). We emphasize that these SoTA results were achieved by applying a small alignment module on top of features precomputed from an independently trained ResNet-50 image encoder and an average word embeddings textual recipe encoder. This is in contrast to the complexity of the previous SoTA approach, which jointly trained image and text models using GANs; an adversarial alignment module; a novel hard negative mining strategy; translation consistency losses; classification losses; and multiple bidirectional LSTMs on top of skip-thought vectors and dedicated ingredient embeddings for textual recipe representation [10]. This suggests that the retrieval metrics on the RECIPE1M dataset still have a lot of room for improvement, and we expect large gains to be achieved with advanced end-to-end methods in future work.

4.4.2 Increasing the Capacity of the Image Encoder

Since our ResNet-Encoder is trained using a classification objective, it is an obvious extension to replace ResNet-50 with ResNext-101 [58] architecture (ResNext-Encoder), which performs better on standard classification benchmarks such as ImageNet [56].

When we use ResNext-Encoder, 10K-R@1 metric for CkNN reaches 22.9 (*CkNN+AWE+ResNext*), matching previous SoTA results from [10], and with triplet loss (*Triplet+AWE+ResNext*) 10K-R@1 is boosted to 30.0, which further improves on previous SoTA by a large margin with the relative change of 30%. It remains to be seen to what extent the existing methods would benefit from other image architectures.

4.4.3 Ablation Study: Impact of Training Data Utilization

One benefit of training text and image models separately as described in Section 3.1 and 3.2, is that one can make use of training data which was filtered out by the existing methods' preprocessing pipeline [4], [10]. Indeed, since the setup for training AWE-Encoder does not require any image-to-textual-recipe pairs and is entirely self-supervised, we are able to train it on 680K textual recipes from the RECIPE1M training set, as opposed to only 240K recipes with images filtered by ACME for joint training [10]. Similarly, we utilize 280K recipes with images for ResNet-Encoder/ResNext-Encoder compared to 240K filtered by ACME [10]. It should be noted, however, that ingredient and instruction embeddings from Pic2Recipe and ACME were also trained on the full 1M dataset [4].

To see by how much our best models' performance was improved by better training data utilization, we train our encoders only on the subset of images and recipes used by

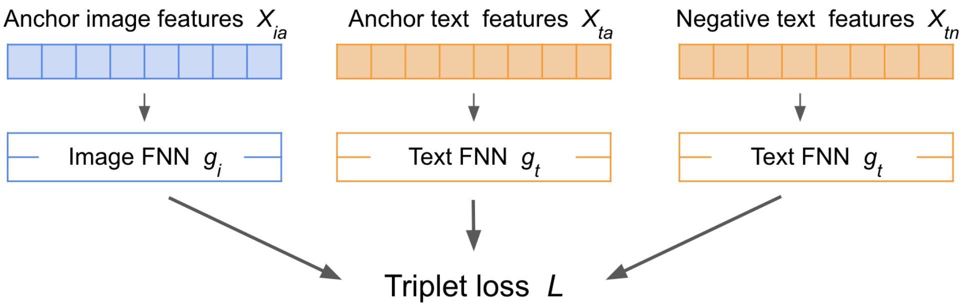
Anchor image features X_{ia}
Anchor text features X_{ta}
Negative text features X_{tn}
Image FNN g_i
Text FNN g_t
Text FNN g_t
Triplet loss L

Fig. 4. Projecting the precomputed image and textual features in the same common space using triplet loss. X_{ia} has D_i dimensions, X_{ta} and X_{tn} have D_t dimensions

TABLE 4
Ablation study of the new SoTA results. † symbol next to the model means using only training data available to ACME [10] method after preprocessing and selecting recipes paired with images. We only report on 10K test set since 1K results are too noisy [8].

Method	medR ↓	R@1 ↑	R@5 ↑	R@10 ↑
(Previous SoTA) ACME [10]	6.7	22.9	46.8	57.9
(Ours) †Triplet+AWE+ResNet	5.9	24.4	49.4	60.5
(Ours) Triplet+AWE+ResNet	5.0	26.5	51.8	62.6
(Ours) †Triplet+AWE+ResNext	4.0	28.6	54.8	65.6
(Ours) Triplet+AWE+ResNext	**4.0**	**30.0**	**56.5**	**67.0**

ACME [10] for joint training. Triplet loss model has already been using the same subset. We observe that 10K-R@1 metric dropped from 30.0 to 28.6 for ResNext, and from 26.5 to 24.4 for ResNet. While this difference is statistically significant, all the models in the ablation study still improve on previous SoTA (†*Triplet+AWE+ResNet* and †*Triplet+AWE+ResNext* in Table 4).

4.5 Performance on POLITICS and GOODNEWS

To understand how well our approach generalizes to other (non-cooking) domains, we apply our method to other large-scale publicly available datasets. We hypothesize that our method will be effective for challenging cross-modal tasks where many of the approaches developed for benchmarks such as Flickr30k will not be applicable, and thus strong baselines might be of value. We identify two such datasets. POLITICS [11] consists of 246K political articles paired with images after preprocessing. GOODNEWS [12] features 466K images from news articles paired with captions. We split GOODNEWS and POLITICS datasets into training, validation and test sets with 80-10-10 ratio following Thomas *et al* [13]. As for RECIPE1M, we focused on the Recall@1 metric on image-to-document retrieval: for each image, we randomly sample 4 candidate documents from the evaluation set, and report the proportion of query images for which the matching document was ranked higher than all the candidate documents as in prior work [13].

We then apply our CkNN baseline to these datasets, analogously to how it was done for RECIPE1M in Section 4.2. Specifically, for POLITICS dataset, we have used the first two sentences of the text document as a title following [13], and extracted $C_i = C_t = 2527$ labels for training AWE-Encoder and ResNet-Encoder. For GOODNEWS dataset we have used the captions to extract the $C_i = 2120$ labels for ResNet-Encoder, and resorted to using FastText [53] for embedding the words of the short captions as discussed in Section 3.1. A sample of extracted labels and documents were manually validated to verify relevance.

For both datasets, we used CkNN to align the modalities. We used exactly the same hyperparameters for CkNN, ResNet-Encoder and AWE-Encoder as we did for RECIPE1M in Section 4.2. We report CkNN performance on POLITICS and GOODNEWS datasets in Table 5. We further report the performance of our models on POLITICS and GOODNEWS after applying the improvements mentioned in Section 4.4 to the respective CkNN baselines described in the current section. For POLITICS, our method achieves Recall@1 of 64.8 on the image-to-text task. For GOODNEWS, it achieves Recall@1 of 88.6 on the image-to-text task. The results in each case are on par with the best published method (see Table 5 and [13]).

Although the performance differences between our work and that of [13] is not statistically significant, our approach, which from one perspective may be viewed as a baseline, is highly competitive and matches the leading methods on these datasets and tasks. We observe that the existing methods for cross-modal recipe retrieval are specialized to and evaluated on only a single Recipe1M dataset, whereas our method is also competitive on POLITICS and GOODNEWS without any modifications, and while using exactly the same hyperparameters.

4.6 Text-To-Image Retrieval

Although in this paper we only focus on image-to-text retrieval tasks for clarity and the ease of analysis, we also report our best model's performances on the mirror text-to-image task for completeness. On RECIPE1M, the proposed method achieves medR of 4.0, 10K-R@1 of 30.5, 10K-R@5 of 56.3 and 10K-R@10 of 66.6. Similar to the image-to-text task, this a large improvement over the previous SoTA

TABLE 5
Recall@1 performance of our baselines and other approaches [13] on POLITICS and GOODNEWS for the 5-way image-to-text retrieval task. The best results within the level of statistical significance are in bold.

Method	POLITICS	GOODNEWS
Trip+NP+Sym	47.4	72.0
PVSE	62.5	87.2
Ang+NP+Sym	62.7	87.0
(Previous SoTA) SN	**64.7**	**88.5**
(Ours) CkNN+AWE+ResNet	60.5	83.8
(Ours) Triplet+AWE+ResNext	**64.8**	**88.6**

results from [10]. Text-to-image retrieval is also very good on POLITICS and GOODNEWS with Recall@1 of 64.9 and 88.5 respectively which is on par with SoTA performance on these datasets [13].

5 CONCLUSION

We conclude this work by highlighting two key takeaway messages. First, this paper provides strong evidence that nearest neighbours, when incorporated according to our methodology, offers an straightforward, yet broadly performant, baseline on a cross-modal recipe retrieval task. The test performance of our proposed model is competitive with SoTA evaluation, even though we deliberately avoided 'advanced' modelling techniques. The benefit of this is efficient, stable and reliable solutions with relatively low computational overhead in training and evaluation. Although these baseline results do not best SoTA, they come reasonably close to it. Secondly, we show that our approach greatly simplifies model exploration and model comparison. We demonstrate in a case study how to compare various models' individual components using our method and use the insights from the analysis to significantly advance SoTA on the definitive cross-modal recipe retrieval dataset, RECIPE1M, with straightforward modifications of our baseline. The resulting model has much less complexity than other methods tailored for RECIPE1M, and it generalizes well enough to match best published results on two other challenging datasets for cross-modal retrieval. We believe that our approach fills a growing need for strong baselines and a systematic comparison framework in cross-modal recipe retrieval and hope that it will facilitate further progress in this space as well as other cross-modal domains.

REFERENCES

[1] J. Marín, A. Biswas, F. Ofli, N. Hynes, A. Salvador, Y. Aytar, I. Weber, and A. Torralba, "Recipe1m: A dataset for learning cross-modal embeddings for cooking recipes and food images," *arXiv preprint arXiv:1810.06553*, vol. abs/1810.06553, 2018. [Online]. Available: http://arxiv.org/abs/1810.06553

[2] A. Myers, N. Johnston, V. Rathod, A. Korattikara, A. Gorban, N. Silberman, S. Guadarrama, G. Papandreou, J. Huang, and K. Murphy, "Im2Calories: Towards an Automated Mobile Vision Food Diary," in *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2015, pp. 1233–1241. [Online]. Available: http://ieeexplore.ieee.org/document/7410503/

[3] J. Freyne and S. Berkovsky, "Recommending food: Reasoning on recipes and ingredients," in *User Modeling, Adaptation, and Personalization*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 381–386.

[4] A. Salvador, N. Hynes, Y. Aytar, J. Marin, F. Ofli, I. Weber, and A. Torralba, "Learning cross-modal embeddings for cooking recipes and food images," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 3068–3076.

[5] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3-4, pp. 321–377, 12 1936. [Online]. Available: https://doi.org/10.1093/biomet/28.3-4.321

[6] J.-j. Chen, C.-W. Ngo, and T.-S. Chua, "Cross-modal recipe retrieval with rich food attributes," in *Proceedings of the 25th ACM International Conference on Multimedia*, ser. MM '17. New York, NY, USA: ACM, 2017, pp. 1771–1779. [Online]. Available: http://doi.acm.org/10.1145/3123266.3123428

[7] J.-J. Chen, C.-W. Ngo, F.-L. Feng, and T.-S. Chua, "Deep understanding of cooking procedure for cross-modal recipe retrieval," in *Proceedings of the 26th ACM International Conference on Multimedia*, ser. MM '18. New York, NY, USA: ACM, 2018, pp. 1020–1028. [Online]. Available: http://doi.acm.org/10.1145/3240508.3240627

[8] M. Carvalho, R. Cadène, D. Picard, L. Soulier, N. Thome, and M. Cord, "Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings," in *The 41st International ACM SIGIR Conference on Research &#38; Development in Information Retrieval*, ser. SIGIR '18. New York, NY, USA: ACM, 2018, pp. 35–44. [Online]. Available: http://doi.acm.org/10.1145/3209978.3210036

[9] B. Zhu, C.-W. Ngo, J. Chen, and Y. Hao, "R2GAN: Cross-modal recipe retrieval with generative adversarial network," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[10] H. Wang, D. Sahoo, C. Liu, E.-p. Lim, and S. C. H. Hoi, "Learning cross-modal embeddings with adversarial networks for cooking recipes and food images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 572–11 581.

[11] C. Thomas and A. Kovashka, "Predicting the politics of an image using webly supervised data," in *Advances in Neural Information Processing Systems*, 2019, pp. 3630–3642.

[12] A. F. Biten, L. Gomez, M. Rusinol, and D. Karatzas, "Good news, everyone! context driven entity-aware captioning for news images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 466–12 475.

[13] C. Thomas and A. Kovashka, "Preserving semantic neighborhoods for robust cross-modal retrieval," in *Proceedings of the European Conference on Computer Vision (ECCV)*, August 2020.

[14] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2641–2649.

[15] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[16] Q. Jiang and W. Li, "Deep cross-modal hashing," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3270–3278.

[17] Y. Wang, X. Luo, L. Nie, J. Song, W. Zhang, and X. Xu, "Batch: A scalable asymmetric discrete cross-modal hashing," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2020.

[18] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proceedings of the 22Nd ACM International Conference on Multimedia*, ser. MM '14. New

York, NY, USA: ACM, 2014, pp. 7–16. [Online]. Available: http://doi.acm.org/10.1145/2647868.2654902

[19] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ser. ICML'11. USA: Omnipress, 2011, pp. 689–696. [Online]. Available: http://dl.acm.org/citation.cfm?id=3104482.3104569

[20] R. Tu, X. Mao, B. Ma, Y. Hu, T. Yan, W. Wei, and H. Huang, "Deep cross-modal hashing with hashing functions and unified hash codes jointly learning," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2020.

[21] Y. Peng, J. Qi, X. Huang, and Y. Yuan, "CCL: Cross-modal correlation learning with multigrained fusion by hierarchical network," *IEEE Transactions on Multimedia*, vol. PP, 04 2017.

[22] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proceedings of the 25th ACM International Conference on Multimedia*, ser. MM '17. New York, NY, USA: ACM, 2017, pp. 154–162. [Online]. Available: http://doi.acm.org/10.1145/3123266.3123326

[23] H. T. Shen, L. Liu, Y. Yang, X. Xu, Z. Huang, F. Shen, and R. Hong, "Exploiting subspace relation in semantic labels for cross-modal hashing," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2020.

[24] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'14. Cambridge, MA, USA: MIT Press, 2014, pp. 2672–2680. [Online]. Available: http://dl.acm.org/citation.cfm?id=2969033.2969125

[25] Y. Peng, J. Qi, and Y. Yuan, "Cm-gans: Cross-modal generative adversarial networks for common representation learning," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 15, 10 2017.

[26] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[27] Z. Wang, X. Liu, H. Li, L. Sheng, J. Yan, X. Wang, and J. Shao, "Camp: Cross-modal adaptive message passing for text-image retrieval," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5764–5773.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 770–778.

[29] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997. [Online]. Available: http://dx.doi.org/10.1162/neco.1997.9.8.1735

[30] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'13. USA: Curran Associates Inc., 2013, pp. 3111–3119. [Online]. Available: http://dl.acm.org/citation.cfm?id=2999792.2999959

[31] R. Kiros, Y. Zhu, R. Salakhutdinov, R. S. Zemel, A. Torralba, R. Urtasun, and S. Fidler, "Skip-thought vectors," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'15. Cambridge, MA, USA: MIT Press, 2015, pp. 3294–3302. [Online]. Available: http://dl.acm.org/citation.cfm?id=2969442.2969607

[32] J. Chen, L. Pang, and C.-W. Ngo, "Cross-modal recipe retrieval: How to cook this dish?" in *MultiMedia Modeling*. Cham: Springer International Publishing, 2017, pp. 588–600.

[33] H. Fu, R. Wu, C. Liu, and J. Sun, "Mcen: Bridging cross-modal gap between cooking recipes and dish images with latent variable model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[34] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 815–823.

[35] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101 – mining discriminative components with random forests," in *European Conference on Computer Vision*, 2014.

[36] X. Chen, H. Zhou, and L. Diao, "ChineseFoodNet: A large-scale image dataset for Chinese food recognition," *arXiv preprint arXiv:1705.02743*, 2017. [Online]. Available: http://arxiv.org/abs/1705.02743

[37] H. Hassannejad, G. Matrella, P. Ciampolini, I. De Munari, M. Mordonini, and S. Cagnoni, "Food image recognition using very deep convolutional networks," in *Proceedings of the 2Nd International Workshop on Multimedia Assisted Dietary Management*, ser. MADiMa '16. New York, NY, USA: ACM, 2016, pp. 41–49. [Online]. Available: http://doi.acm.org/10.1145/2986035.2986042

[38] A. Singla, L. Yuan, and T. Ebrahimi, "Food/non-food image classification and food categorization using pre-trained GoogLeNet model," in *Proceedings of the 2Nd International Workshop on Multimedia Assisted Dietary Management*, ser. MADiMa '16. New York, NY, USA: ACM, 2016, pp. 3–11. [Online]. Available: http://doi.acm.org/10.1145/2986035.2986039

[39] N. Martinel, G. L. Foresti, and C. Micheloni, "Wide-slice residual networks for food recognition," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2018, pp. 567–576.

[40] T. Sato, J. Harashima, and M. Komachi, "Japanese-English machine translation of recipe texts," in *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*. Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 58–67. [Online]. Available: https://www.aclweb.org/anthology/W16-4603

[41] D. Park, K. Kim, Y. Park, J. Shin, and J. Kang, "Kitchenette: Predicting and ranking food ingredient pairings using siamese neural network," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, 7 2019, pp. 5930–5936. [Online]. Available: https://doi.org/10.24963/ijcai.2019/822

[42] M. F. Dacrema *et al.*, "Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches," Proceedings of the 13th ACM Conference on Recommender Systems (RecSys 2019), 2019.

[43] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman, "Total recall: Automatic query expansion with a generative feature model for object retrieval," in *2007 IEEE 11th International Conference on Computer Vision*, Oct 2007, pp. 1–8.

[44] P. Turcot and D. Lowe, "Better matching with fewer features: The selection of useful features in large database recognition problems," in *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, Sep. 2009, pp. 2109–2116.

[45] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arxiv:1810.04805*, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[46] C. Doersch *et al.*, "Unsupervised visual representation learning by context prediction," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ser. ICCV '15. Washington, DC, USA: IEEE Computer Society, 2015, pp. 1422–1430. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2015.167

[47] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, "Neural codes for image retrieval," in *Computer Vision – ECCV 2014*. Cham: Springer International Publishing, 2014, pp. 584–599.

[48] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, 06 2015.

[49] J. Wieting, M. Bansal, K. Gimpel, and K. Livescu, "Towards universal paraphrastic sentence embeddings," in *International Conference on Learning Representations*, 2016.

[50] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," *31st International Conference on Machine Learning, ICML 2014*, vol. 4, 05 2014.

[51] F. Hill, K. Cho, and A. Korhonen, "Learning distributed representations of sentences from unlabelled data," in *HLT-NAACL*, 2016.

[52] S. Guo and N. Yao, "Document vector extension for documents classification," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2019.

[53] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017. [Online]. Available: http://aclweb.org/anthology/Q17-1010

[54] J. Deng, J. Guo, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," *arXiv preprint arXiv:1801.07698*, 2018.

[55] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[56] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A Large-Scale Hierarchical Image Database,” in *CVPR09*, 2009.

[57] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[58] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, “Aggregated residual transformations for deep neural networks,” *arXiv preprint arXiv:1611.05431*, 2016.