Pattern Recognition Letters 146 (2021) 38–45

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Deep multimodal learning for cross-modal retrieval: One model for all tasks

L. Viviana Beltrán Beltrán[a,*], Juan C. Caicedo[b], Nicholas Journet[c], Mickaël Coustaty[a], François Lecellier[d], Antoine Doucet[a]

[a] University of La Rochelle, 17000, France
[b] Fundación Universitaria Konrad Lorenz, Bogotá, Colombia
[c] University of Bordeaux, 33000, France
[d] University of Poitiers, 86000, France

ARTICLE INFO

Article history:
Received 28 June 2020
Revised 2 February 2021
Accepted 23 February 2021
Available online 13 March 2021

Keywords:
Embeddings
Multimodal learning
Deep learning
Cross-modal
Visual question answering
Evaluation study

ABSTRACT

We investigate the effectiveness of a successful model in Visual-Question-Answering (VQA) problems as the core component in a cross-modal retrieval system that can accept images or text as queries, in order to retrieve relevant data from a multimodal document collection. To this end, we adapt the VQA model for deep multimodal learning to combine visual and textual representations for information search, and we call this model "Deep Multimodal Embeddings (DME)". Instead of training the model to answer questions, we supervise DME to classify semantic topics/concepts previously identified in the document collection of interest. In contrast to previous approaches, we found that this model can handle any multimodal query with a single architecture while producing improved or competitive results in all retrieval tasks. We evaluate the model performance with 3 widely known databases for cross-modal retrieval tasks: Wikipedia Retrieval Database, Pascal Sentences, and MIR-Flickr-25k. The results show that the DME model learns effective multimodal representations, resulting in strongly improved retrieval performance, specifically in the top results of the ranked list, which are the most important to users in the most-common scenarios of information retrieval. Our work represents a new baseline for a wide set of different methodologies for cross-modal retrieval.

1. Introduction

Searching data in documents generally relies on one data modality (image or text) while users generally expect to take advantage of all the available data. By leveraging images, text contents, and their semantic relationships in an integrated fashion, an ideal system should propose a diversity of ways in which a multimedia collection could be used. In this paper, we propose to adapt a VQA (Visual-Question Answering) architecture for multimodal fusion. VQA systems are generally trained to answer natural language questions related to the content of images. One of the most popular and pioneer works in this area is presented by Antol et al. [1]. They introduced a large database to perform the task of free-form and open-ended Visual Question Answering that has helped advance the research in this area. They also introduced guidelines and baselines with fundamental requirements in computer vision, natural language processing, and commonsense reasoning for systems aiming to solve the VQA task. Since then, there has been great progress in the design and understanding of VQA systems, and the vision community has introduced databases and improvements to make the systems accurate [2]. Given their natural design to process images and text, we aim to solve the question of **how useful is the multimodal representation encoded by such models to solve queries in cross-modal retrieval problems?**

More precisely, we propose to evaluate how VQA systems could be used to create a common latent representation for multimodal data. Our goal is to evaluate the representation learning capability of VQA network architectures to encode the input modalities in a meaningful way for cross-modal retrieval tasks. Results indicate that these architectures can be retrained for cross-modal search without the need for special layers or additional model developments, making them ideal as a baseline for future multimodal indexing research. To assess our hypothesis, we used three very popular datasets in two different configurations: the topic classification performance with the Wikipedia Retrieval database [3];

* Corresponding author
E-mail addresses: vbeltran@univ-lr.fr (L.V.B. Beltrán), journet@labri.fr (J.C. Caicedo), mickael.coustaty@univ-lr.fr (M. Coustaty), francois.lecellier@univ-poitiers.fr (F. Lecellier), antoine.doucet@univ-lr.fr (A. Doucet).

https://doi.org/10.1016/j.patrec.2021.02.021
0167-8655/

the retrieval tasks with the Wikipedia database, the Pascal Sentences database [4], and the MIR-Flickr-25k database [5]. Our extensive experiments simultaneously evaluate all cross-modal retrieval tasks under the same computational framework as well as the uni-modal tasks.

Our results show that this strategy provides competitive results for all multimodal retrieval tasks as compared to other works that might be over-optimized for a single retrieval task and that results in more complex models. Our approach performs particularly well on the top-K relevant results, which are especially important in the most frequent use cases of information retrieval. We compare our approach against recent and specialized models that represent a variety of methodologies, allowing researchers to compare performance in different cross-modal retrieval tasks under different assumptions.

Surprisingly, even though the VQA architecture is a well established model that can be used for other image-text problems, previous work had not investigated its usefulness for cross-modal retrieval. Even the most recent approaches for cross-modal retrieval do not take advantage of the innovations introduced in VQA architectures. The main contributions of our work include:

1. The re-interpretation of the VQA architecture for cross-modal retrieval, followed by an extensive experimental evaluation of its capabilities in this problem.
2. Our work is the first to evaluate all cross-modal and uni-modal tasks with a single model trained only once without fine-tuning in specific tasks, reaching highly competitive results as well as state-of-the-art results in some cases, even though the baselines may be exclusively over-optimized for single tasks. This shows that a unifying model for multimodal retrieval is possible.
3. The simplicity of our approach serves as a baseline for future research and can inspire extensions to push performance even further. We expect future researchers taking advantage of this well established architecture to reach higher performance in their models.

This paper is organized as follows. While Section 2 describes related work in multimodal retrieval, Section 3 presents the architecture of our system DME. Section 4 details the experimental framework, and provides subsequent results and discussion. Finally, Section 5 concludes the paper and suggests potential ideas for future work.

2. Related work

There is a large number of past and ongoing work in cross-modal retrieval research. We will focus on cross-modal methods that cover a variety of existing strategies. First, there exist strategies based on probabilistic models that find correlations among modalities. These are now classic techniques and are based on finding projection functions that map modalities to a subspace in which the correlation among them is maximized. Most of them propose or adapt their techniques based on the commonly known canonical correlation analysis (CCA) algorithm [6,7].

The second group of methods is based on tensor factorization and graph embedding theory. These are based on tensor computations (usually matrix factorization models, in the case of two modalities), manifold learning, or in establishing graph relations between the modalities and finding projection matrices (PM) to map the data to an embedding space restricted by some regularization and constraint terms [8,9]. Some problems of these models are the complexity created when non-linear mappings are included (kernel functions), which leads to higher computational requirements without significant improvements in their performance.

Table 1
Classification accuracy evaluated by using different sets of data: First, testing different image architectures when only visual data is passed, second, evaluating different sequence lengths for the representation of the samples when only text is passed, and finally, by testing a different number of hidden layers in the multimodal component when visual and textual data is passed. Results are reported for the use case, the Wikipedia Retrieval database along with baselines results for the uni-modal evaluations.

Modality	Model	Acc
Images	VGGNet19 + 1 hidd + Class	0.4733
	VGGNet19 + 2 hidd + Class	0.4574
	VGGNet19 + 3 hidd + Class	0.4531
	ResNet + 1 hidd + Class	**0.5165**
	ResNet + 2 hidd + Class	0.4920
	ResNet + 3 hidd + Class	0.5122
	Baseline ResNet + SVM	0.4271
Texts	26-LSTM + 1 hidd + Class	0.7272
	46-LSTM + 1 hidd + Class	0.7359
	76-LSTM + 1 hidd + Class	0.7301
	100-LSTM + 1 hidd + Class	**0.7474**
	Baseline BERT + SVM	0.7445
Multimodal	Visual Net + Text Net + Class	0.7243
	Visual Net + Text Net + 1 hidd + Class	0.7142
	Visual Net + Text Net + 2 hidd + Class	**0.7561**

Methods based on deep learning are also popular and have been used successful in the context of representation learning mainly for visual feature learning, where the textual modality guides the process of learning semantic features [10]. These can be based on Restricted Boltzmann Machines [11,12], or use deep multi-layer and combined architectures [13–15]. These models for data fusion are usually composed of two (or more) sub-networks, one per modality [16,17] as in the case of deep adversarial networks [18–21], which are optimized together or that share a common last layer. Another approach is to connect the different modalities all over the network, either as multiple sub-networks [22], or by reinforcing the fusion in the form of several fusion layers [23]. This last approach can be seen as a multitask model with three outputs, one for each modality and one multimodal.

Combined approaches are also found in the literature, where the visual part is passed through a neural network, and the textual part is passed through a model such as a bag of words, and a late fusion component is used to combine both of them [24]. Recent works are trying to take advantage of raw data in databases such as descriptions and captions in the form of topics [25]. The majority of recent methods follow another tendency when addressing the problem of cross-modal retrieval as the learning of hash codes representing embeddings for different modalities optimized in a joint framework and by evaluating different lengths of hash codes. These models are mainly based in deep learning [26–29], but can also use matrix factorization [30,31], and both of them have been applied with successful results. These methods are based on binary supervision, considering multi-modal relationships as either completely similar or completely dissimilar. The most important aspect of these methods is the importance of the code length that has a significant influence on the final result. Usually, larger codes obtain better results [32]. Other applications include robotics [33], audio-video fitting [34], and cooking activities [35] (See Supplementary Material, Table 1 for a summary of works in the state of the art).

In the present paper, we analyze a successful end-to-end approach that allows us to perform uni-modal and cross-modal retrieval, by exploiting specialized models. Unlike most previous works, the ability to resolve queries of various types without additional training or finetuning is one of its strongest features. The experiments provide a general comparison with the state of the art by including several different approaches that report results in our target databases.

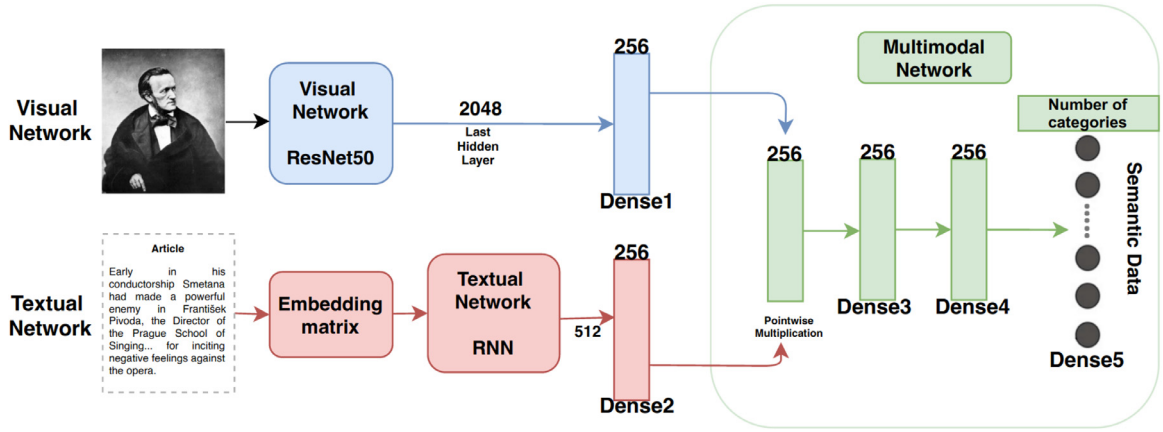

Fig. 1. Overview of the architecture for DME. There are three components: 1 - The visual network (in blue), which takes the last hidden layer of a pretrained ResNet50 model with 2048-dim as input for the visual neural network, followed by a fully-connected layer with output 256-dim. 2 - The textual network (in red) composed of an embedding matrix of GloVe vectors of 300-dim, followed by two LSTM cells, and a fully-connected layer with output 256-dim. 3 - Finally, a multimodal network (in green), where the different modalities are fused. This component contains a set of fully connected layers with the last layer (output layer) used as a classifier during the training phase. At the retrieval phase, we can use different layers to compute the embeddings. Either by using layers of each modality (Dense1 for visual data, Dense2 for text data) or by using layers after the pointwise multiplication (Dense3, Dense4, or Dense5) that contain multimodal information.

3. Proposed method

In order to design search systems that allow both text to image, image to text queries, and uni-modal queries using a single model, we make use of an architecture inspired by Visual Question Answering (VQA) models [1], and we extend it for Deep Multimodal Learning. This model matches the inputs and outputs of a multimodal system (see Fig. 1 for details), and it can be used to compute embeddings for different modalities. In what follows, we describe the main three components of this architecture.

3.1. Visual network

We evaluated two CNN architectures to compute visual representations: VGGNet [36] and ResNet50 [37]. We use pre-trained weights from object classification using the ImageNet dataset and keep the network frozen as a feature extractor. The feature vector produced by VGGNet has 4096 dimensions, while the ResNet produces 2048. These activations from the last hidden layer are used as inputs to a fully connected layer with output 256-dim to reduce the dimensionality and match the connection with the multimodal network.

3.2. Textual network

This network is a recurrent neural network (RNN) that models the sequence of words in sentences. This provides the advantage of having an ordered representation of word sequences in a fixed-length feature vector, which is in contrast to orderless bag-of-words models commonly used for information retrieval. LSTMs are powerful neural networks to learn sequence models [38] and are used in VQA systems to represent the words used to formulate questions [39]. In the case of other sequences of text, we can evaluate the performance by identifying the most 'relevant' words present in the input text. While question words such as 'Where' and 'What' are important to understand what is the requested information in the questions of VQA systems, in this work, we are more interested in nouns. Nouns are words with semantic and discriminatory content that are possibly more related to the general content/topic/concept of an image (such as 'King', 'Cat', 'Church', etc.) than question words. To process sentences with the RNN, each word in the text is first encoded into 300-dimensional embeddings using the GloVe model [40]. The vocabulary used consists of all the words seen in the training database, and those with an embedding representation in the pre-trained GloVe embeddings. The matrix of word representations is used as an input to the RNN, which has two LSTM cells followed by a fully-connected layer with intermediate dropout layers and activation functions ReLU to get a 256-dim vector. We also explored different sequence lengths for the input texts and evaluated the varying impact on performance.

3.3. Multimodal network

For the architecture, we implement a deep multi-layer neural network with four layers as follows. We use an element-wise fusion layer (pointwise multiplication or Hadamard product), where both modalities, visual and textual, are combined. We found this operation to be critical for obtaining improved performance, which is defined as:

$$(V \circ T)_i = (V)_i (T)_i \; for \; all \; 0 \leq i \leq n, \tag{1}$$

where $n = dimV = dimT$. This fusion layer is followed by two fully connected layers with dropout, and the network has a classification layer in the output (categories or concepts that semantically discriminate the samples in the database). The model is trained in an end-to-end configuration with cross-entropy as the loss function. The last activation function is a softmax function applied to the final output vector. Once the system is trained, we can compute embeddings using different layers of the model, and evaluate their quality in retrieval tasks. We can compute separate embeddings by using layers before the modalities are combined, but this would imply to use only data of the target network, which either visual or textual, therefore not multimodal. We are interested in computing features in a latent space that can match any data modality, and therefore we use features from the outputs of layers Dense3, Dense4, or Dense5 as embeddings for cross-modal retrieval (green layers in Fig. 1). To compute multimodal features with a single modality we simply remove the other network and skip the pointwise product, which is equivalent to using a constant vector of ones as a replacement.

4. Experiments

In this section, we include the results of our experiments. First, we introduce the databases used in our analysis. Then, we present

the implementation details and finally, the results for our ablation studies, including classification results and, multimodal retrieval results that include the comparison against the state-of-the-art (See Supplementary material, Appendices 2,3 and 4 for additional details in the setting of our experiments).

4.1. Databases

For our experiments, we make use of three databases: Wikipedia Retrieval, Pascal Sentences, and MIR-Flickr-25k. We choose them for two main reasons: First, because they meet our data requirements as they contain images, text descriptions, and category information. And second, they are widely used in cross-modal retrieval works (see Section 4.3.2) which is our target task. We use the Wikipedia retrieval database, widely used database in cross-modal retrieval research, to make an extended analysis of the classification performance of the model. Pascal Sentences and MIR-Flickr-25k are also widely used in retrieval tasks because they have good ground truth annotations, and well organized textual representation for each image.

Wikipedia retrieval database [3] The database consists of 2866 image-document pairs with a train/test partition of 2173 and 693 examples respectively. Each image-text pair is labeled with one of 10 semantic classes. The median text length is 200 words. After analyzing the distribution of words, we kept only those with frequencies between 10 and 150. We explored different sequence lengths for training our model, including 26 words, as well as 46, 76, and 100. We separated the train set, in train and validation, in a percentage of 80%-20%. This gives us a partition of 1738/432 training and validation samples, respectively.

Pascal sentences [4] This contains 1000 samples of images associated with several sentences and descriptions of their content (approximately 5 sentences per image) from 20 categories, 50 images per category. We follow the standard partition of 600/400 (i.e., 30/20 samples per category) for training and test samples. For texts, we concatenate all sentences for each sample, and extract all words. We set the sequence length as the average number of words in all samples, resulting in a sequence length of 25 words. We did not prune the distribution of words in this dataset.

MIR-Flickr-25k [5] This database consists of 25,000 images downloaded from the social photography site Flickr and annotated using a set of annotations with a total of 24 semantic concepts. We set the sequence length in this dataset to 15 tags. Note that these are no sentences; instead the text annotations are just a collection of tags. The median number of tags per image is 5, and we use all tags independently of their frequency. In order to compare with the state of the art, we make two partitions of the data, (1) as in Xie et al. [29], Luo et al. [41]), taking approximately 2000 samples as the test partition, and (2) taking 95% of the data as training, 5% for test and reporting the average of 3 experiments, following the setup of [26]).

4.2. Implementation details

To implement our approach,[1] we used the Keras framework, and trained the model described in Fig. 1 with a learning rate of 0.001. The input size of the convolutional network is 224×224 pixels. We explored data augmentation strategies for images and text and obtained better results when using minimal augmentation for images. For the text, we augment the word sequences by creating different combinations of up to N random tokens from all the tokens available for each image (where N is the maximum sequence length in the experiment). The sequence fixed-length for each database is as follows: For Wikipedia Retrieval database, we explore the most effective sequence length with a final result of 100 tokens (see Section 4.3.1); for Pascal sentences we fixed it in 25 tokens; and finally, for MIR-Flickr-25k we fixed the length in 15 tokens. To create the train-validation folds, we follow similar configurations for each database in previous works. This resulted as follows: one fold for Wikipedia retrieval database; one fold for Pascal Sentences; and finally, five folds for MIR-Flickr-25k. The indexes of each partition are shared in the repository as *Indexes*. We systematically explored the batch size (16-256) and dropout rates (0.0–0.5) parameters and conducted experiments with different optimization algorithms (best Adam and RMSPROP [42]) (See Supplementary material, Appendix 4: Detailed Parameter tuning for more information).

[1] Code available at https://github.com/lvbeltranb/DME

4.3. Results

This last section gathers the results of our evaluations. First, we worked on the classification analysis. The two next parts are related to the retrieval results for all the selected databases in a unimodal and a cross-modal way.

4.3.1. Classification accuracy

To the best of our knowledge, there are no results for classification accuracy reported for Wikipedia Database. To get a better understanding of the system performance, we evaluated the classification accuracy in each separate component: the visual and the textual networks. We report detailed results of classification accuracy in the supplementary material, and discuss the highlights next.

For image classification, we explored embeddings as well as the number of suitable layers before merging the data with the multimodal network. The best results of visual classification are obtained when using the last hidden layer from a pre-trained ResNet with 50 convolutional layers to compute visual embeddings, followed by 1 fully connected layer and the classification layer. For text classification, we explored sequence lengths leaving fixed the rest of the network. The best configuration for the architecture includes two LSTM cells followed by 1 fully connected layer plus the classification layer. We found the length of the sequence input to work well with 100 tokens. Finally, we explored the number of layers in the multimodal component for pair classification, where the best configuration uses 2 hidden layers in the multimodal network.

We also compared the performance of our neural network classification model against a representative baseline method based on SVM classifiers for each modality. The SVM baseline used BERT embeddings for text [43], and ResNet embeddings for images [37]. The baseline classification performance was 0.7445 for text, and 0.4271 for images, while our approach reached 0.7474 for text and 0.5165 for images (see Table 1). The final accuracy for the multimodal model is 0.7561.

In our experiments, we observed that the performance does not degrade in the multimodal system. Although the visual modality has significantly lower performance than the text modality, due to the fact that while samples from the same category at the textual level share key words between them, the visual content can vary drastically, it also opens interesting questions for future research to balance the contribution of performance from each modality.

4.3.2. Unimodal and cross-modal retrieval

This section focuses on the evaluation of multimodal retrieval tasks. Most previous works report results in only one or two retrieval tasks; in our work, we evaluated all cross-modal retrieval tasks using the same trained model, including 'Img2Img' (also known as Query-by-Example), 'Txt2Txt', Txt2Img' (also known as Query-by-String), and 'Img2Txt' (also known as image captioning).

Table 2
Mean Average Precision (MAP) on the Wikipedia retrieval database under different retrieval tasks using the first K results in the ranking (8, 50, and 500). For each task, we present the results of prior work under the same configuration of our evaluated approach. Results with '' were not reported by the authors.

Task	Method	K		
		8	50	500
Txt2Txt	BERT [43]	**0.735**	0.60	0.406
	DME (ours)	0.681	**0.641**	**0.576**
Img2Img	CMSTH [9]	–	–	**0.449**
	ResNet [37]	0.433	0.323	0.193
	DME (ours)	**0.455**	**0.399**	0.299
Img2Txt	CMSTH [9]	–	–	0.337
	RE-DNN [44]	0.351	0.281	–
	DME (ours)	**0.451**	**0.435**	**0.426**
Txt2Img	CMSTH [9]	–	–	**0.387**
	RE-DNN [44]	0.23	0.241	–
	DME (ours)	**0.489**	**0.49**	0.382

Table 3
Mean Average Precision (MAP) on the Wikipedia retrieval database for different retrieval tasks when evaluating 100% of samples in the ranked list. The best two results are highlighted in green (first) and orange (second). For each one of the tasks, we put the results of methods in the state-of-the-art, under the same configuration of our approach.

Task	Method	MAP (100%)
Txt2Txt	BERT [43]	0.3873
	DME (ours)	0.567
Img2Img	ResNet [37]	0.195
	DME (ours)	0.283
Img2Txt	SCM_1 [3]	0.277
	RE-DNN [44]	0.340
	MDCR [8]	0.435
	Marginal SM [45]	0.332
	SCM_2 [6]	0.362
	Self_Supervised [25]	0.391
	LCALE [46]	0.367
	DME (ours)	0.422
Txt2Img	SCM_1 [3]	0.226
	RE-DNN [44]	0.352
	MDCR [8]	0.394
	Marginal SM [45]	0.241
	SCM_2 [6]	0.273
	Self_Supervised [25]	0.434
	LCALE [46]	0.357
	DME (ours)	0.360

The X2Y convention is used to represent the source (X) and target (Y) modalities. As an example, the 'Txt2Img' refers to the task where the queries are texts and the retrieved samples are images. Assuming the DME model is trained, the cross-modal retrieval process is then divided in two phases, the indexing phase and the search phase. The indexing phase generates embeddings for all the documents in the database using the target modality only. In the retrieval phase, the query modality is processed using our DME model without activating the network of the target modality (visual or textual). As a performance measure of the ranking of retrieved results, we use the mean average precision (MAP), which is a standard evaluation metric in information retrieval. We calculate the MAP on a different number of samples to retrieve. The MAP 100% corresponds to the case when we use all the samples in the test set. Therefore, for the Wikipedia database, the maximum size is 693, for Pascal it is 400, and for MIR-Flickr-25k it is 2000 in the setting (1).

Ablation study We studied the impact of two factors in the performance of the model: The distance metrics used to retrieve documents and the quality of embeddings taken from different layers in our model. The first evaluation is focused on identifying good distance metrics to retrieve points in the multimodal feature space. We compared the MAP results in the evaluated tasks on the Wikipedia database by using different functions: Kullback–Leibler divergence (KL), Euclidean distance, and normalized correlation (NC), with average performances for the fourth tasks of 0.4424, 0.4513, and **0.4917** respectively. Since normalized correlation had the best performance in almost all experiments, we use it as the similarity metric in the rest of our evaluation (see Table 2 in Supplementary material for detailed results).

The second evaluation is focused on the quality of the embeddings obtained from different layers. We compute embeddings for the visual and textual modalities using the corresponding layer in the model: Dense1 for images, Dense2 for text, and layers Dense3, Dense4, or Dense5 to obtain a fused representation after the pointwise multiplication layer (see Fig. 1). Each evaluation is made independently (only one layer is used at each time). We had the hypothesis that multimodal network can generate better embeddings in deeper layers, where it has the opportunity to learn higher level representations of the combined information. The results for the first 3 tasks are consistent with our hypothesis, the performance increases when we compute the embeddings from deeper layers (Dense1/2 < Dense3 < Dense4 < Dense5), with average performances for the 4 tasks of 0.3945, 0.3978, 0.3908 and **0.4080** (see Table 3 in Supplementary material for detailed results).

Comparison with state-of-the-art: wikipedia database In this part, we present a comparison of results for cross-modal retrieval tasks for all databases. We report the state of the art results under the same protocol of experimentation, but covering different methodologies. Previous works have not reported results for uni-modal tasks, (similar to results in Section 4.3.1), thus, we used BERT (for text), and ResNet (for images) as reproducible baselines to compare against. Up to our knowledge, we are the first reporting 'Txt2Txt' results in these databases, and we also report results for 'Img2Img' in the same way that few works have done before. By reporting results for these two tasks, we aim to establish baselines for future works too.

Tables 2 and 3 report MAP results obtained for different tasks for Wikipedia retrieval database. Table 2 presents the MAP performance when we consider the top 8, 50, and 500 results in the ranking list. Our model exhibits the best performance in the case of top 8 and top 50 results in almost all tasks, which are the most useful cases for users in real world conditions. The lowest performance in general is obtained when searching images with visual queries, which is a challenging task. However, our multimodal approach can improve performance in the top results of this task, showing the benefits of our general purpose fusion framework.

An alternative evaluation protocol uses 100% of the ranked results instead of only the top K when computing MAP. Previous work using the Semantic Correlation Matching (SCM) [3,6] reported results following this protocol. The SCM approach finds a semantic space that takes into account correlated multimodal projections, and we used it as a baseline under this protocol. To analyze the statistical significance of our approach we run a t-test comparing our model against the baseline using the average of 18 runs, with 15 different queries each in the Img2Txt and Txt2Img tasks. The null hypothesis that our approach has no difference with the baseline was rejected in in both cases with significance threshold 0.05 (p-values: Img2Txt = 0.0001, Txt2Img = 0.036). Previous works do not explicitly fuse modalities [45] but instead use the textual modality for supervising the visual representation learning. For example, a recent self-supervised model addressed the task of cross-modal retrieval, training an LDA model of text and topics as prediction target for the CNN that processes the visual modality

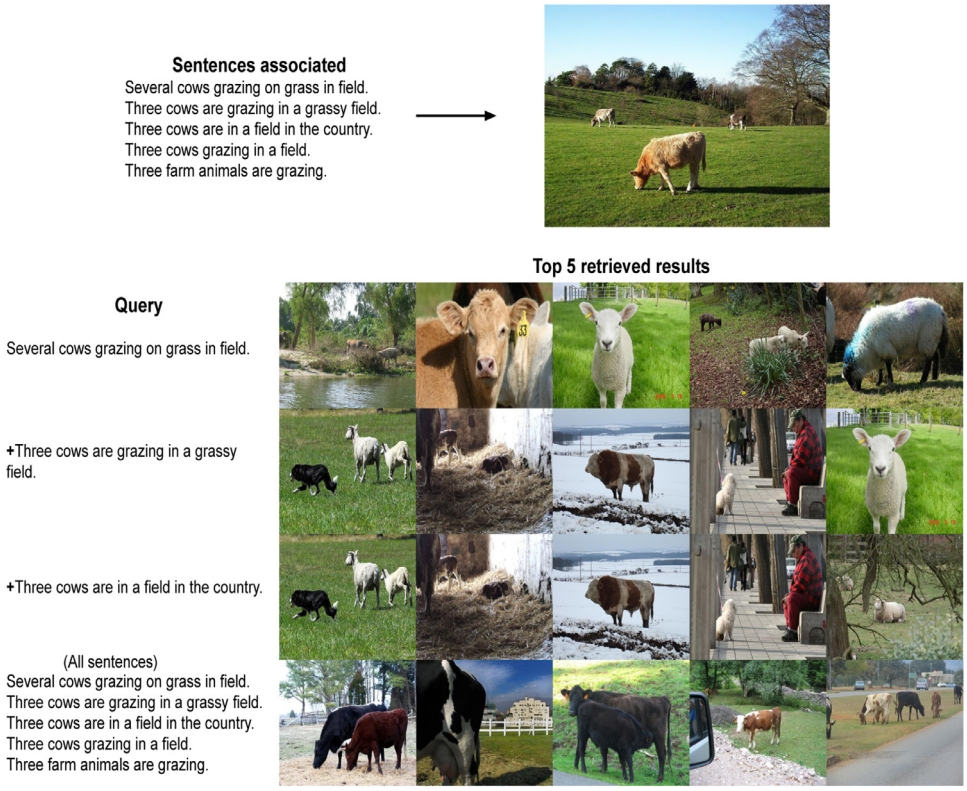

Fig. 2. Retrieval Samples: At the top, the original sample with its sentences and associated image, at the bottom, retrieved images by gradually adding more sentences to the text query.

Table 4
MAP comparison on Pascal Sentences and MIR-Flickr-25k databases under different retrieval tasks and by using the first K results (10, 100, and 200).

Task	Method	Pascal			MIR-Flickr-25k		
		10	100	200	10	100	200
Txt2Txt	BERT	0.449	0.274	0.230	0.870	0.793	0.764
	DME (ours)	**0.556**	**0.486**	**0.465**	**0.883**	**0.839**	**0.823**
Img2Img	ResNet	0.584	0.405	0.376	0.946	0.897	0.874
	DME (ours)	**0.586**	**0.495**	**0.486**	**0.962**	**0.948**	**0.943**
Img2Txt	DME (ours)	0.532	0.474	0.458	0.942	0.915	0.899
Txt2Img	DME (ours)	0.523	0.470	0.451	0.919	0.893	0.876

[25]. This model gets the best performance in the Txt2Img task, but is not competitive in the Img2Text task showing how the models can be optimized for only one task instead of being truly multimodal.

The most similar approach to our DME model is the Regularized Deep Neural Network (RE-DNN) [44] (see also Section 1). This model has three parts: the visual, textual, and multimodal subnets; each one pre-trained separately before fusion. RE-DNN is not trained end-to-end, and in contrast to our model the textual network does not use sequential word embeddings. Notably, RE-DNN does not use a pointwise multiplication layer for fusion, but instead relies on concatenation of features. The results show that our model surpasses their performance in all tasks, indicating that the choices of our model yield enhanced fusion performance. The analysis of cross-modal results in Tables 2 and 3 confirm that ours is generally the best performing approach. The one case where it is not the best is when MAP is computed on 100% of the ranked results (Table 3). However, our result is very competitive and has improved performance in all other metrics and tasks. Our approach may also be more computationally efficient given that models like MDCR learn two projection functions while we aim to learn only one generic multimodal mapping.

Comparison with state-of-the-art: pascal and MIRFlickr Tables 4–6 present the MAP performance for Pascal Sentences and MIR-Flickr-25k databases. Table 4 presents the results when MAP is evaluated by taking the top 10, 100, and 200 retrieved results (we choose these values as being suitable for both databases) showing good results, especially for the MIR-Flickr-25k database. We start comparing our performance against BERT (for text) and ResNet (for images) as uni-modal baselines methods.

Tables 5 and 6 present results of methods with comparable setups for both databases. We obtain competitive results in the Pascal Sentences dataset, and we reach the best performance in the MIR-Flickr-25k dataset. Our model reaches state-of-the-art performance in the MIR-Flickr-25k dataset when compared with recent cross-modal retrieval approaches, including TFNH [26], a hash-based method, and CPAH [29]. We evaluated the retrieval performance in the Txt2Img task when gradually increasing the length of the text sequence in the query by adding more sentences in the Pascal database. An illustrative example is presented in Fig. 2.

Table 5
MAP comparison on Pascal Sentences dataset. The best two results are highlighted in green (first) and orange (second).

Task	Method	MAP (100%)
Txt2Txt	BERT [43]	0.196
	DME (ours)	0.453
Img2Img	ResNet [37]	0.361
	DME (ours)	0.436
Img2Txt	MDCR [8]	0.455
	Marginal SM [45]	0.222
	Self_Supervised [25]	0.326
	LCALE [46]	0.414
	DME (ours)	0.429
Txt2Img	MDCR [8]	0.471
	Marginal SM [45]	0.173
	Self_Supervised [25]	0.360
	LCALE [46]	0.394
	DME (ours)	0.425

Table 6
MAP comparison on MIR-Flickr-25k dataset. The best two results are highlighted in green (first) and orange (second).

Task	Method	MAP (100%)
Txt2Txt	BERT [43]	0.717
	DME(2)	0.759
Img2Img	ResNet [37]	0.793
	DME(2)	0.868
Img2Txt	RCCA(1) [41]	0.695
	CPAH(1) [29]	0.791
	TFNH(2) [26]	0.688
	DME(1)	0.725
	DME(2)	0.809
Txt2Img	RCCA(1) [41]	0.694
	CPAH(1) [29]	0.789
	TFNH(2) [26]	0.761
	DME(1)	0.731
	DME(2)	0.805

Note that the relevance of the top 5 results changes with different textual queries, and when more descriptions are available our system can retrieve more accurate results. We achieve these results with a single model that can handle variable sequence length as inputs for computing the multimodal representation (see other visual retrieval results in the Supplementary material, Appendices 6 and 7 along with an analysis of embeddings from Wikipedia Retrieval database).

5. Conclusions

We have analyzed and evaluated an approach that can process visual and textual modalities in a document and learns a semantic relationship between both. Inspired by visual question answering architectures, this approach can help in learning combined representations to build effective cross-modal retrieval systems. The results of our experiments underline two positive aspects of the evaluated model. First, performance does not degrade after combining the two modalities (one of the modalities can be noisier than the other). Second, our goal was to investigate the potential of a single model to retrieve relevant documents in cross-modal tasks without being optimized exclusively for any one of them. We performed an extended analysis of our approach, and without the need to fine-tune the model or results for every single retrieval task, we achieved robust results in 3 datasets for all tasks and even reached state-of-the-art performance in some of them. The work presented in this paper represents a new baseline for cross-modal learning and retrieval, given its simplicity and good performance.

As part of our future work, we consider evaluating the use of state-of-the-art language models, such as BERT, instead of the RNN architecture. Models based on Transformers are revolutionizing language and vision, and our architecture can incorporate such advances to improve performance even further.

Declaration of Competing Interest

The authors declare that they have no known competing fnancial interests or personal relationships that could have appeared to infuence the work reported in this paper.

Acknowledgements

This work has been supported by the French region of Nouvelle Aquitaine under the Animons project and by the Mires research federation.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at10.1016/j.patrec.2021.02.021

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, D. Parikh, VQA: visual question answering, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2425–2433.
[2] D. Zhang, R. Cao, S. Wu, Information fusion in visual question answering: asurvey, Inf. Fusion 52 (2019) 268–280.
[3] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G.R. Lanckriet, R. Levy, N. Vasconcelos, A new approach to cross-modal multimedia retrieval, in: Proceedings of the 18th ACM International Conference on Multimedia, ACM, 2010, pp. 251–260.
[4] A. Farhadi, M. Hejrati, M.A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, D. Forsyth, Every picture tells a story: generating sentences from images, in: European Conference on Computer Vision, Springer, 2010, pp. 15–29.
[5] M.J. Huiskes, M.S. Lew, The MIR Flickr retrieval evaluation, in: Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval, ACM, 2008, pp. 39–43.
[6] J. Costa Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G.R. Lanckriet, R. Levy, N. Vasconcelos, On the role of correlation and abstraction in cross-modal multimedia retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 36 (3) (2014) 521–535, doi:10.1109/TPAMI.2013.142.
[7] K. Wang, R. He, L. Wang, W. Wang, T. Tan, Joint feature selection and subspace learning for cross-modal retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 38 (10) (2016) 2010–2023, doi:10.1109/TPAMI.2015.2505311.
[8] Y. Wei, Y.A.O. Zhao, Z. Zhu, S. Wei, Y. Xiao, J. Feng, S. Yan, A Modality-dependent Cross-media Retrieval V(212), ACM Trans. Intell. Syst. Technol. (2015) 1–13.
[9] L. Xie, L. Zhu, P. Pan, Y. Lu, Cross-modal self-taught hashing for large-scale image retrieval, Signal Process. 124 (2016) 81–92, doi:10.1016/j.sigpro.2015.10.010.
[10] J. Wang, Y. He, C. Kang, S. Xiang, C. Pan, Image-text cross-modal retrieval via modality-specific feature learning, in: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval - ICMR '15, 2015, pp. 347–354, doi:10.1145/2671188.2749341.
[11] D. Wang, P. Cui, M. Ou, W. Zhu, Deep multimodal hashing with orthogonal regularization, in: IJCAI International Joint Conference on Artificial Intelligence 2015-Janua, 2015, pp. 2291–2297.
[12] D. Hu, X. Lu, X. Li, Multimodal learning via exploring deep semantic similarity, in: Proceedings of the 2016 ACM on Multimedia Conference - MM '16, 2016, pp. 342–346, doi:10.1145/2964284.2967239.
[13] P. Hu, D. Peng, X. Wang, Y. Xiang, Multimodal adversarial network for cross–modal retrieval, Knowl.-Based Syst. 180 (2019) 38–50.
[14] P.-Y. Huang, X. Chang, A.G. Hauptmann, et al., Improving what cross-modal retrieval models learn through object-oriented inter-and intra-modal attention networks, in: Proceedings of the 2019 on International Conference on Multimedia Retrieval, ACM, 2019, pp. 244–252.
[15] P. Hu, L. Zhen, D. Peng, P. Liu, Scalable deep multimodal learning for cross-modal retrieval, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019, pp. 635–644.
[16] M. Engilberge, L. Chevallier, P. Pérez, M. Cord, Finding beans in burgers: deep semantic-visual embedding with localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3984–3993.
[17] L. Zhen, P. Hu, X. Wang, D. Peng, Deep supervised cross-modal retrieval, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10394–10403.

[18] F. Shang, H. Zhang, L. Zhu, J. Sun, Adversarial cross-modal retrieval based on dictionary learning, Neurocomputing 355 (2019) 93–104.
[19] W. Gu, X. Gu, J. Gu, B. Li, Z. Xiong, W. Wang, Adversary guided asymmetric hashing for cross-modal retrieval, in: Proceedings of the 2019 on International Conference on Multimedia Retrieval, ACM, 2019, pp. 159–167.
[20] F. Wu, X.-Y. Jing, Z. Wu, Y. Ji, X. Dong, X. Luo, Q. Huang, R. Wang, Modality-specific and shared generative adversarial network for cross-modal retrieval, Pattern Recognit. 104 (2020) 107335.
[21] Y. Wu, S. Wang, G. Song, Q. Huang, Augmented adversarial training for cross–modal retrieval, IEEE Trans. Multimed. 23 (2020) 559–571.
[22] X. Xu, H. Lu, J. Song, Y. Yang, H.T. Shen, X. Li, Ternary adversarial networks with self-supervision for zero-shot cross-modal retrieval, IEEE Trans. Cybern. 50 (2019) 2400–2413.
[23] V. Vielzeuf, A. Lechervy, S. Pateux, F. Jurie, Centralnet: a multilayer approach for multimodal fusion, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018. 0–0
[24] I. Gallo, A. Calefati, S. Nawaz, Multimodal classification fusion in real-world scenarios, in: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 5, IEEE, 2017, pp. 36–41.
[25] Y. Patel, L. Gomez, M. Rusiñol, D. Karatzas, C. Jawahar, Self-supervised visual representations for cross-modal retrieval, in: Proceedings of the 2019 on International Conference on Multimedia Retrieval, ACM, 2019, pp. 182–186.
[26] Z. Hu, X. Liu, X. Wang, Y.-m. Cheung, N. Wang, Y. Chen, Triplet fusion network hashing for unpaired cross-modal retrieval, in: Proceedings of the 2019 on International Conference on Multimedia Retrieval, ACM, 2019, pp. 141–149.
[27] Z. Ji, W. Yao, W. Wei, H. Song, H. Pi, Deep multi-level semantic hashing for cross-modal retrieval, IEEE Access 7 (2019) 23667–23674.
[28] S. Su, Z. Zhong, C. Zhang, Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 3027–3035.
[29] D. Xie, C. Deng, C. Li, X. Liu, D. Tao, Multi-task consistency-preserving adversarial hashing for cross-modal retrieval, IEEE Trans. Image Process. 29 (2020) 3626–3637.
[30] T. Yao, X. Kong, L. Yan, W. Tang, Q. Tian, Efficient discrete supervised hashing for large-scale cross-modal retrieval, arXiv preprint arXiv:1905.01304(2019).
[31] Y. Fang, H. Zhang, Y. Ren, Unsupervised cross-modal retrieval via multi-modal graph regularized smooth matrix factorization hashing, Knowl.-Based Syst. 171 (2019) 69–80.
[32] C. Huang, X. Luo, J. Zhang, Q. Liao, X. Wang, Z.L. Jiang, S. Qi, Explore instance similarity: an instance correlation based hashing method for multi-label cross–model retrieval, Inf. Process. Manag. 57 (2) (2020) 102165.
[33] W. Zheng, H. Liu, B. Wang, F. Sun, Cross-modal surface material retrieval using discriminant adversarial learning, IEEE Trans. Ind. Inform. 15 (2019) 4978–4987.
[34] S.-W. Chung, J.S. Chung, H.G. Kang, Perfect match: Self-supervised embeddings for cross-modal retrieval, IEEE J. Sel. Top. Signal Process. 14 (2020) 568–576.
[35] H. Wang, D. Sahoo, C. Liu, E.-p. Lim, S.C. 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. 11572–11581.
[36] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556(2014).
[37] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[38] K. Greff, R.K. Srivastava, J. Koutník, B.R. Steunebrink, J. Schmidhuber, LSTM: a search space odyssey, IEEE Trans. Neural Netw. Learn. Syst. 28 (10) (2017) 2222–2232.
[39] Z. Yu, J. Yu, Y. Cui, D. Tao, Q. Tian, Deep modular co-attention networks for visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 6281–6290.
[40] J. Pennington, R. Socher, C. Manning, GloVe: global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543.
[41] J. Luo, Y. Shen, X. Ao, Z. Zhao, M. Yang, Cross-modal image-text retrieval with multitask learning, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, ACM, 2019, pp. 2309–2312.
[42] S. Ruder, An overview of gradient descent optimization algorithms, arXiv preprint arXiv:1609.04747(2016).
[43] H. Xiao, bert-as-service, 2018, (https://github.com/hanxiao/bert-as-service).
[44] C. Wang, H. Yang, C. Meinel, A deep semantic framework for multimodal representation learning, Multimed. Tools Appl. 75 (15) (2016) 9255–9276, doi:10.1007/s11042-016-3380-8.
[45] A.K. Menon, D. Surian, S. Chawla, Cross-modal retrieval: a pairwise classification approach, in: Proceedings of the 2015 SIAM International Conference on Data Mining, 2015, pp. 199–207, doi:10.1137/1.9781611974010.23.
[46] K. Lin, X. Xu, L. Gao, Z. Wang, H.T. Shen, Learning cross-aligned latent embeddings for zero-shot cross-modal retrieval., in: AAAI, 2020, pp. 11515–11522.