Multimedia Tools and Applications
https://doi.org/10.1007/s11042-023-17977-0

Cascaded transformer-based networks for wikipedia large-scale image-caption matching

Nicola Messina[1] · Davide Alessandro Coccomini[1] · Andrea Esuli[1] · Fabrizio Falchi[1]

Received: 15 July 2022 / Revised: 26 September 2023 / Accepted: 21 December 2023

Abstract

With the increasing importance of multimedia and multilingual data in online encyclopedias, novel methods are needed to fill domain gaps and automatically connect different modalities for increased accessibility. For example, Wikipedia is composed of millions of pages written in multiple languages. Images, when present, often lack textual context, thus remaining conceptually floating and harder to find and manage. In this work, we tackle the novel task of associating images from Wikipedia pages with the correct caption among a large pool of available ones written in multiple languages, as required by the image-caption matching Kaggle challenge organized by the Wikimedia Foundation. A system able to perform this task would improve the accessibility and completeness of the underlying multi-modal knowledge graph in online encyclopedias. We propose a cascade of two models powered by the recent Transformer networks able to efficiently and effectively infer a relevance score between the query image data and the captions. We verify through extensive experiments that the proposed cascaded approach effectively handles a large pool of images and captions while maintaining bounded the overall computational complexity at inference time. With respect to other approaches in the challenge leaderboard, we can achieve remarkable improvements over the previous proposals (+8% in nDCG@5 with respect to the sixth position) with constrained resources. The code is publicly available at https://tinyurl.com/wiki-imcap.

Keywords Multi-modal matching · Information retrieval · Deep learning · Transformer networks

✉ Nicola Messina
nicola.messina@isti.cnr.it

Davide Alessandro Coccomini
davidealessandro.coccomini@isti.cnr.it

Andrea Esuli
andrea.esuli@isti.cnr.it

Fabrizio Falchi
fabrizio.falchi@isti.cnr.it

[1] Institute of Information Science and Technologies, National Research Council, Via Giuseppe Moruzzi, 1, Pisa 56124, Italy

Published online: 09 January 2024

1 Introduction

With the increasing overload of digital information stored in digital databases, novel methods are needed to improve the accessibility of online content [1, 2], especially the one from big online encyclopedias. In recent years, new deep learning techniques have achieved outstanding results in many computer vision and language tasks, and numerous attempts have been made to merge the two worlds. Particularly, Transformer-based networks have recently redefined the ways in which vision and language are processed, and many architectures have been introduced to create informative common spaces, where efficient k-NN search can be performed to search images given a natural language text and vice-versa, like CLIP [3]. This ability to align text and images in a latent space has been widely employed to solve many vision-language tasks, ranging from image captioning [4] to text-guided image synthesis [5].

Among all the emerging multi-modal tasks, image-caption matching is becoming particularly important, especially for scalable and efficient cross-modal retrieval [6, 7]. Image-caption matching involves associating an image with the text that best describes it. It can be used to find the most relevant images for a given query text (*text-to-image retrieval*) or vice-versa (*image-to-text retrieval*). These are important challenges that can make multimedia content more accessible and complete. While text-to-image retrieval has important applications in multimedia search engines – where a natural language phrase is used to search for visual content [8] – no natural use-cases arise for the complementary image-to-text retrieval scenario. Recently, WikiMedia foundation issued a Kaggle competition[1] that concerns the retrieval of captions from Wikipedia pages associated with a certain image. An example of the setup is reported in Fig. 1. This task turns out to be critical in large online encyclopedias, where automatically linking images to the textual concepts referenced in the page text enables the underlying knowledge base to remain complete and up-to-date. A system able to perform this task would improve the accessibility and completeness of the underlying multi-modal knowledge graph in online encyclopedias, also assisting article writers in suggesting relevant captions for an inserted figure.

Existing solutions depend on straightforward approaches that use translations or page links, but these methods have restricted coverage. One possibility is employing free-form image captioning, where text is generated from scratch, given the image. However, most sophisticated image captioning methods nowadays struggle to handle images with intricate semantic content. Furthermore, generating text from scratch is unnecessary – other than computationally expensive – if we assume that the text describing the picture is already present somewhere on the page where the image resides. For these reasons, the best option is to frame the problem as a caption retrieval task, given the image content as context.

In principle, recent cross-modal models that produce an informative common latent space like CLIP [3] could solve this problem. However, there are some problems in adopting CLIP-like models to solve this task: i) the task is framed as a multi-modal retrieval task, where the query information is represented not only by the image itself but also its URL, which carries important priors useful for an effective matching; ii) CLIP latent space may struggle if not adapted to Wikipedia images, whose distribution may diverge consistently from the CLIP training set; iii) the dataset employed in the competition – the WIT dataset [9] – is also multi-lingual (captions are written in 108 different languages), requiring the development of an ad-hoc *multi-modal multi-language* retrieval pipeline.

Another important requirement for the developed solution is to be *efficient*, given that the model should ideally process billions of Wikipedia pages. The WIT dataset [9] on which the

[1] https://www.kaggle.com/c/wikipedia-image-caption/overview

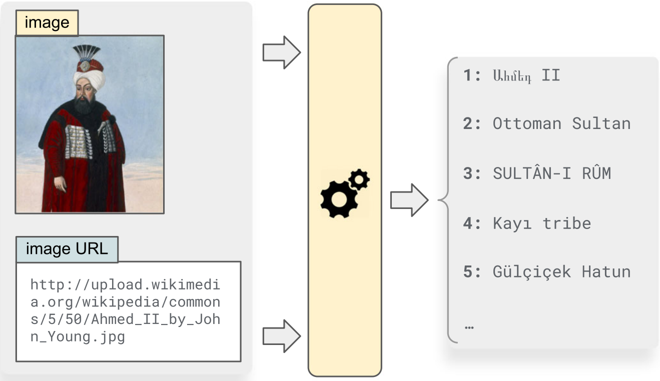


Fig. 1 Given an image and its URL, the objective is to find the most relevant caption

competition is based – alongside other proposed Wikipedia-based collections [10, 11] – is composed of millions of image-text pairs that require scalable architectures to be processed. Therefore, in our work, we devise an efficient and effective two-stage pipeline for retrieving the most relevant captions given both textual and visual information (image URL and image itself) for solving the challenge proposed by the Wikimedia foundation.

Specifically, we propose a cascade of two image-text matching models based on large pre-trained Transformer models. The first model, called Multi-modal Caption Proposal (MCProp), is based on the common space matching approach and uses XLM-RoBERTa and CLIP as text and image feature extractors, respectively. Being very efficient at inference time, this model is used to quickly propose potentially relevant candidates. The second model, Caption Re-Rank (CRank), is a fine-tuned XLM-RoBERTa pairwise classifier. This model is less efficient but more accurate and is used to re-score and reorder the candidates from the first stage. After training each stage separately, we run extensive studies on the two modules to understand the validity of the whole pipeline on a carefully chosen validation set. Finally, we perform inference on the test set provided by the challenge, keeping only the top five captions as required by the challenge rules.

We compare our performance with the other competitors. Our approach achieves the fifth position on the final private leaderboard, with a final nDCG of 0.53, a +8% improvement with respect to the team in the sixth position. The detailed experimental analysis performed on the two modules shows the effectiveness and the efficiency of the proposed pipeline against the other competitors.

To summarize, in this paper, we propose the following contributions:

- We introduce a multi-lingual multi-modal architecture composed of a cascade of Transformer-based networks for solving the novel, challenging task of image-based caption retrieval on Wikipedia pages.
- We show the effectiveness and the efficiency of the proposed solution, achieving the fifth position in the Wikimedia challenge proposed on Kaggle.
- We perform extensive experimentation on the two models to validate the proposed pipeline.

The rest of the paper is organized as follows: Section 2 summarizes some of the most influential works related to the presented task; Section 3 explains the two-stage pipeline for efficiency and effectively retrieves the most relevant captions; Section 4 present in-depth experiments on the two proposed models, showing the validity of the entire pipeline; Section 5 draws final conclusions and present possible extensions to this work; Finally, Appendix A, B and C report implementation details, together with some more in-depth experimentation done at inference time to try to boost the performance of the proposed model.

The code for reproducing our results is publicly available on GitHub[2].

2 Related works

In this section, we review some relevant literature for the explored task, with particular attention to *transformer-based networks*, *image captioning*, *cross-modal retrieval*, together with some works on multimedia understanding in Wikipedia.

Transformer-based networks

A big step forward for multi-modal understanding has been achieved with the introduction of Transformers [12], which proved to be extremely powerful both in the field of natural language processing with models like BERT [13], ELMo [14], GPT-3 [15], RoBERTa [16] and ALBERT [17], and in the field of computer vision with the introduction of Vision Transformers [18] and its variants like Swin Transformer [16], CrossViT [19] and Twins-SVT [20]. Since Transformers are so effective in both fields, they were used to extract common representations between multi-modal data so that they could later be compared or processed altogether. For example, some works showed the effectiveness of Transformers, from image captioning [21] to text-driven object detection [22], while TERAN [6] and, more recently, CLIP [3] demonstrated the power of transformers for cross-modal retrieval. Recently, researchers devised large vision language transformers like VinVL [23], ViLT [24], or Flamingo [25], pretrained on massive amounts of data, to solve many downstream vision-language tasks. Driven by these recent advances, our model employs some state-of-the-art transformer networks as image and text backbones for processing images and captions, respectively.

Image captioning

When trying to obtain a textual description from an image, one of the possibilities is *image captioning*, which consists of generating natural language text conditioned on the given input image. Early neural models for image captioning [26–28] encoded visual information using a single feature vector representing the whole image. Therefore, they were not able to exploit information about objects and their spatial relationships. In the last years, the concept of attention, which underpins the operation of Transformers, has proven to be crucial for image captioning tasks. Indeed, when deciding which combination of natural language words best describes an image, it is required to identify its most important and discriminating parts. The first application of Transformers on this task can be found in [29], in which the authors proposed the novel Conceptual Caption dataset and proved the effectiveness of Transformers in the captioning task. In [30], the authors exploit an image Transformer to obtain captions by attending the different image regions. Differently, in [21], the authors propose a meshed-memory Transformer, which uses mesh-like connectivity at decoding stage to exploit the activations at different depths of the network. Some works also use GANs as a framework to learn to caption images. Specifically, in [31], the authors present a novel method relying on conditional- GAN, which introduces an extension to traditional encoder-decoder architectures based on reinforcement learning (RL).

[2] https://github.com/mesnico/Wiki-Image-Caption-Matching

Sentence and image-sentence retrieval

Our research is more focused on sentence or image-sentence matching. This setup differs slightly from image captioning: in the matching task the captions must not be generated, but only carefully chosen among a set of candidates. Sentence matching obtained a huge boost in the last few years, thanks to large pre-training of Transformer models, such as BERT [13], RoBERTa [16], and multi-lingual versions of them like XLM-RoBERTa [32]. These models have been recently extended to work with images. Some works use BERT-like processing on both visual and textual modalities, such as ViLBERT [33], ImageBERT [34], Pixel-BERT [35], VL-BERT [36]. Nevertheless, all these methods require several network evaluations that scale quadratically with the number of items in the inference set. In fact, all the possible image-caption pairs should be input into the network to obtain the matching score. For this reason, many methods rely instead on the projection of visual and textual information into the same common space, where only a simple dot-product is needed to obtain the similarity between a given pair. In particular, in [37] the authors use VGG and ResNets visual features extractors, together with an LSTM for sentence processing, and they match images and captions exploiting hard-negatives at training time. Following, other methods focused on contrastive learning of common embedding spaces [38–40]. Differently, in [41], an adversarial learning method is proposed, and a discriminator is used to learn modality-invariant representations. The authors in [42] use a contextual attention-based LSTM-RNN which can selectively attend to salient regions of an image at each time step, and they employ a recurrent canonical correlation analysis to find hidden semantic relationships between regions and words. In [43] the authors proposed a system for combining image and text features for image retrieval. They introduced a fusion approach called Text Image Residual Gating (TIRG), in which the image feature is first gated and then added to a residual feature which works as a *modification* feature. Transformer networks have been used in image-text matching in [6, 7] for the task of multi-modal large-scale information retrieval. They introduced a novel disentangled transformer architecture that separately reasons on the two different modalities and enforces a final common abstract concept space.

Differently from all these works, in our efficient candidate proposal model we dealt with multi-modal queries composed of a text (the image URL) and the image itself. Therefore, we have a slightly different setup, in which an (image, text) pair is used to retrieve the captions (another text).

Multi-modal understanding in Wikipedia

Some relevant works have been recently proposed to tackle important challenges in Wikipedia. In order to train Wikipedia-scale multi-modal models, many datasets have been crawled by scraping Wikipedia pages, like WIT [9], WikiWeb2M [10], and AToMiC [11], which collect millions of image-text pairs. These datasets have been used in recent large vision-language pre-training multilingual models [44]. Especially, MURAL [45] extends ALIGN [46] by performing both image-text matching and translation pair matching, while REVEAL [47] proposes handling image captioning and image-text retrieval using an external memory network that is pretrained on massive data. The work that mostly resembles our task and motivation is the context-driven Wikipedia captioning method by [48], which employs images, descriptions, and sections in the article to generate a very precise caption. However, they do not frame the task as a retrieval problem, and they train a complex encoder-decoder transformer model, which is hard to deploy in real-world large-scale scenarios.

3 Method

The data provided in the Kaggle challenge consists of three main fields: *image URLs*, *images*, and *captions*. The challenge consists of finding the most relevant caption given the image URLs, the images, or both as a query. Given that the test set is composed of around $n_t = 92K$ elements, using a large Transformer to compute relevance score for every (query, target) pair is infeasible, as we would need to compute n_t^2 relevance scores to get the ranking for the whole test set. Driven by this concern, we decided to adopt a cascade of two different models to produce the final rankings. The first one, which we call Multi-modal Caption Proposal (MCProp) model, employs both the textual information in the image URLs and the visual information in the images as a compound query to infer the caption. This model projects queries and captions in the same common space, where cosine similarity is used to measure the similarity between a query and a caption. With this model, efficient k-nearest neighbor search can be performed to create a rank for every query of all the n_t captions. The top-ranked elements are then used as candidates by the second model, called Caption Re-Rank (CRank) model. This is a large Transformer fine-tuned for pair classification, i.e., a binary classifier that classifies a (query, caption) pair as either a match or a non-match. This second model employs only the textual information in the image URL to infer the caption without relying again on the visual information. The highest match probabilities returned by MCProp determine the top-5 captions selected for every image, as requested by the challenge. Following, we present in detail the two models.

3.1 The multi-modal caption proposal model

The core idea of the Multi-modal Caption Proposal method is to transform the query and target data to transfer them to a common feature space. In this space, we can calculate the similarity in an efficient and scalable manner using cosine similarities. The model comprises two pipelines, the *query* pipeline and the *caption* pipeline. In turn, a query is composed of an *image URL* and an *image*. Therefore, in total, we need to process two textual fields and one visual field.

Visual features are extracted from images via the image encoder part of a CLIP network [3]. CLIP is a powerful multi-modal model composed of an image encoder and a text encoder that is trained to predict the correct visual-textual pairings. Being pre-trained on a multi-modal task, the image encoder module is a very good fit for our task.

We do not use the textual pipeline of CLIP as our textual backbone. The main reason for this is that this challenge is inherently multilingual, and CLIP is not trained on multiple languages. For this reason, we use instead a pre-trained large language Transformer model, XLM-RoBERTa [32], as a textual pipeline.

The overall architecture is shown in Fig. 2. Specifically, the image encoder of CLIP outputs an aggregated visual feature $\bar{\mathbf{v}}$; differently, XLM-RoBERTa outputs a set $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M\}$ of textual features, that are used by the heads of the original model for solving the many downstream tasks. Instead, we post-process these features by means of other Transformer Encoder layers in a way similar to TERN [7] and TERAN [7], obtaining $\{\mathbf{w}'_1, \mathbf{w}'_2, \ldots, \mathbf{w}'_M\}$. We use the token embedding from the first element of the output sequence, the CLS token, as a final representation for the input text: $\mathbf{c} = \mathbf{w}'_1$. The CLS token has been introduced in the BERT architecture [13], and it is a special token – usually the first element in the sequence – which is aimed at collecting global information from the sentence. All the other tokens are needed inside both XLM-RoBERTa and the subsequent Transformer Encoder

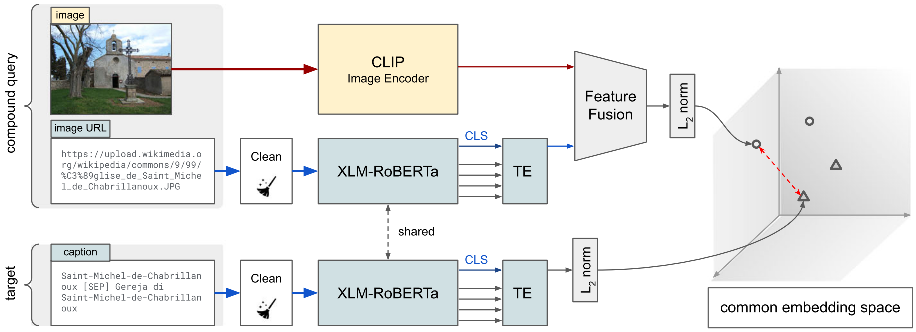

compound query
image
image URL
https://upload.wikimedia.org/wikipedia/commons/9/99/%C3%89glise_de_Saint_Michel_de_Chabrillanoux.JPG
Clean
CLIP
Image Encoder
XLM-RoBERTa
CLS
TE
Feature Fusion
L_2 norm
shared
target
caption
Saint-Michel-de-Chabrillanoux [SEP] Gereja di Saint-Michel-de-Chabrillanoux
Clean
XLM-RoBERTa
CLS
TE
L_2 norm
common embedding space

Fig. 2 The Multi-modal Caption Proposal model. The XLM-RoBERTa backbone is shared among the query and caption pipelines, and the representations are further specialized using downstream Transformer Encoders (TE), whose architecture is reported on the left side of Fig. 1 in [12]. The final CLS tokens in output from the TEs are used as final embeddings for the image URL and caption, respectively

layer for computing self-attention but are discarded afterward since they are not needed for the downstream task. Notice that the XLM-RoBERTa backbone is shared among the image URL and the caption pipelines. In fact, once the image URL has been properly cleaned, it resembles a valid natural language text that can be processed with standard pre-trained textual models. In order to specialize the representations to the downstream task, the two downstream Transformer Encoder modules from the two pipelines do not share the weights. Concerning input preprocessing, the image URL is cleaned by removing the extension, the part preceding the actual filename, and cleaning special characters like underscores or dashes, which are replaced by a space.

The image URL and the image are fused using an attentive fusion module described in the next paragraph.

3.1.1 Attentive feature fusion

The proposed problem is challenging since two different modalities (the image URL and the image) can be used as a compound query to infer the image caption. It would be interesting to automatically infer the relative importance of the two components of the query to solve the matching task. The *attentive feature fusion* module, inspired by other works in this direction [49, 50], serves precisely this purpose. This module is composed of a sub-network that computes two attention values, one for each query component. Specifically, the network is a simple MLP with a final sigmoid layer that takes as input the concatenation among the two input vectors **u** (image URL) and **v** (the image) and outputs two scalars, α_u and α_v:

$$\alpha_u, \alpha_v = \text{sigmoid}(\text{MLP}([\mathbf{u}, \mathbf{v}])) \tag{1}$$

where $[\cdot, \cdot]$ denotes the concatenation operation. Thanks to the final sigmoid layer, these values lay in the range [0, 1]. Those values are then used for computing the final query representation **q** as a weighted average between the normalized input vectors:

$$\mathbf{q} = \alpha_u \frac{\mathbf{u}}{\|\mathbf{u}\|} + \alpha_v \frac{\mathbf{v}}{\|\mathbf{v}\|} \tag{2}$$

The vectors are normalized so their intrinsic magnitude is 1. Doing so, the α_u and α_v values are forced to be directly comparable and more easily interpretable.

3.1.2 Training

In order to match images and captions in the same common space, we use a hinge-based triplet ranking loss, focusing the attention on hard negatives, as in [6, 7, 37, 38]. Specifically, given the final query representation **q** and the target caption feature **c** we use the following loss function:

$$\begin{aligned} L_{match}(\mathbf{q}, \mathbf{c}) = \max_{\mathbf{c}'}[\gamma + S(\mathbf{q}, \mathbf{c}') - S(\mathbf{q}, \mathbf{c})]_{+} + \\ \max_{\mathbf{q}'}[\gamma + S(\mathbf{q}', \mathbf{c}) - S(\mathbf{q}, \mathbf{c})]_{+}, \end{aligned} \tag{3}$$

where $[x]_{+} \equiv \max(0, x)$. $S(\mathbf{q}, \mathbf{c})$ is the similarity function between the query vector and the target caption features. We used the standard cosine similarity as $S(\cdot, \cdot)$. As in [37], the hard negatives are sampled from the mini-batch and not globally, for performance reasons.

3.2 The caption re-rank model

The Caption Re-Rank (CRank) model is a binary classifier based on the XLM-RoBERTa model. More specifically, the network consists of the XLM-RoBERTa model, i.e., the encoder part of a Transformer model, with the pooled output of the CLS token connected to a linear layer with an output size equal to the number of labels. The overall architecture is shown in Fig. 3.

The classification task aims to determine if an image URL and a caption match or not. We use a processed version of the image URL to represent the image. We do not use visual information from the image in this phase. As for the Multi-modal Caption Proposal model, the URL is processed by removing any URL of the path component preceding the actual filename and any file type extension. Any underscore or dash characters in the remaining string are replaced by a space. The input of the matching process is the concatenation of the processed URL and the caption text, with a SEP token separating them.

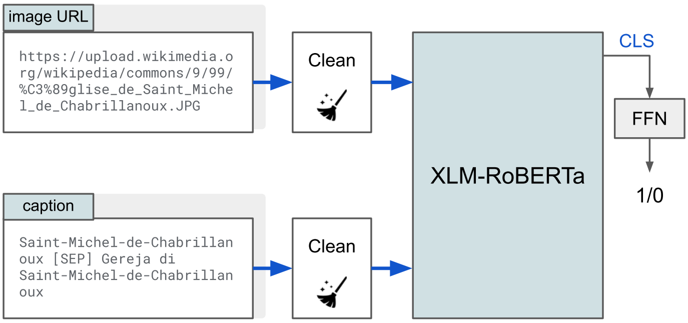


image URL
https://upload.wikimedia.org/wikipedia/commons/9/99/%C3%89glise_de_Saint_Michel_de_Chabrillanoux.JPG
caption
Saint-Michel-de-Chabrillanoux [SEP] Gereja di Saint-Michel-de-Chabrillanoux
Clean
Clean
XLM-RoBERTa
CLS
FFN
1/0

Fig. 3 The Caption Re-Rank model. It uses an XLM-RoBERTa masked language model pre-trained on the CommonCrawl dataset and fine-tuned on the image URL-caption match classification task. The CLS token in output from XLM-RoBERTa is attached to a feed-forward network (FFN) head, which outputs the final matching probability

3.2.1 Training

To fine-tune the pre-trained model for our classification task, we trained CRank using all the (image URL, caption) pairs available in the training dataset. The dataset explicitly defines only examples of matches. To get examples of non-matching pairs, we used a simple negative sampling strategy that randomly pairs image URLs and captions from the dataset. We generated several non-matching pairs equal to the matching ones, obtaining a training set of about 74 million examples. The training process used a batch size of 64, with each batch containing an equal amount of matching and non-matching pairs. We trained it for 2 epochs, using the Adam algorithm with weight decay [51], requiring 65 hours per epoch on a single NVidia RTX2080 GPU.

3.2.2 Selection of candidates for classification

The number of classifications required for the image-caption match problem we are facing grows with the square of the size of the test set. Each classification requires passing the string representing the (image URL, caption) pair through the XLMRoBERTa model, which takes a non-negligible time. This is structurally different from the MCProp model, in which images and caption are projected in the common space separately, thus with a linear cost with respect to the test set size. For the MCProp model, the quadratic cost is limited to the computation of the cosine similarities between the resulting vectors, which is a faster operation that can also be computed in parallel much easily.

The overall cost for CRank can rapidly pass the limit of available computational and time resources. With a single NVidia RTX2080 GPU the time required to compute the classification scores for all the (image URL, caption) pairs in the test set is more than three months. For comparison, computing all the pairwise similarities on embedding vectors of length 768 for the same test set requires eight minutes on a desktop CPU. Multiple GPUs can reduce the cost to a more manageable time, yet the quadratic nature of the process still make it not scalable to larger dataset. For this reason, we employed this two-steps approach, in which a faster method, e.g., MCProp, selects a smaller set of promising candidates to be processed by CRank. Having a fixed number of candidate captions for image makes the cost linear with respect to the dataset size. Using 1,000 candidate captions per image, determined using MCProp or other methods, it takes only 27 hours (compared to three months) to apply CRank to the 92K image URLs from the test set. In our experiments, we show how the final effectiveness is not influenced, while the efficiency of the overall inference phase is greatly improved.

4 Experiments

4.1 Dataset

The dataset used for our experiments is the one publicly released on the Kaggle competition page. It is based on Wikipedia-based Image Text Dataset (WIT) [9] and contains 37.6 million entity-rich image-text examples with 11.5 million unique images across 108 Wikipedia languages. There are at least 12,000 examples for each language, making the dataset particularly interesting for building a model that is not necessarily relegated to a specific language. Each example contains an image URL, from which the image can be downloaded, and the target caption.

The dataset is already divided into a training and a test set. The training set contains for each (image, image URL) pair the associated caption describing the image. On the other hand, the test set separately comprises a list of (image, image URL) pairs that compose the query and a list of captions not paired with the given queries. Each of these two lists contains 92,367 entries. The ground-truth for the test set, i.e., the (image, caption) pairs, is not released. The only way to obtain the results on the test set is to submit the inferred top-5 captions for each query to the Kaggle evaluation server.

The dataset employed in this research is available on the Kaggle challenge page[3], while the full WIT dataset can be downloaded following the instructions provided by the authors on their GitHub repository[4].

4.2 Evaluation

The quality of the obtained ranking is calculated considering the top 5 most similar captions for each image and applying the normalized Discounted Cumulative Gain (nDCG), the normalized version of the well-known Discounted Cumulative Gain (DCG). The rationale behind DCG is to penalize relevant items that are preceded by non-relevant items in the ordered list of results. The DCG grows with the exponential of the graded relevance of an item while it is inversely proportional to the logarithm of the item rank. The DCG at a particular rank position p is defined as follows:

$$\mathrm{DCG}_p = \sum_{i=1}^{p} \frac{2^{\mathrm{rel}_i} - 1}{\log_2 (i + 1)},$$

where rel_i is the graded relevance of the item at position i in the results.

The nDCG normalizes DCG by its maximum theoretical value and thus returns values in the [0, 1] range. To calculate the nDCG at a specific rank position p the following formula is used:

$$\mathrm{nDCG}_p = \frac{\mathrm{DCG}_p}{\mathrm{IDCG}_p},$$

where IDCG is the ideal discounted cumulative gain. Besides the nDCG metric employed in the challenge, we also compute the recall@K metric on our validation sets. This metric is widely-used when there is only one relevant item for every given query, like in this case. The recall@K measures the percentage of queries able to retrieve the correct result among the first K retrieved items.

Given that the final nDCG is computed by averaging the nDCG values for each query, we can also estimate the confidence interval on the mean to evaluate the statistical significance of its values to better compare different experiments. Therefore, for nDCG_k values in the tables reporting results on our validation set (Tables 1 and 2), we are able to report the 95% confidence intervals.

[3] https://www.kaggle.com/competitions/wikipedia-image-caption/data

[4] https://github.com/google-research-datasets/wit

Table 1 Caption retrieval results for the multi-modal caption proposal model on the validation set

Model	Recall@K K=1	 K=5	 K=10	nDCG_5
MCProp no-imgs	48.1	55.9	58.5	0.522 ± 0.009
MCProp imgs, concat	48.0	57.9	62.2	0.533 ± 0.009
MCProp imgs, att-fusion	**48.4**	**58.6**	**62.7**	**0.538** ± 0.009

For nDCG_5, we also report the 95% confidence interval on the mean
Bold entries signify best values

4.3 Preliminary results on the MCProp model

For training the Multi-modal Caption Proposal model, we reserved 10,000 examples chosen randomly from the given training set for validating. For validating the model, we used the main metric used for the challenge, the nDCG_5. However, we also report the Recall@K, as it is one of the main metrics used in cross-modal retrieval literature.

We used CLIP, provided with a ViT as the backbone for the visual pipeline. Instead, the language backbone is a XLM-RoBERTa, a large Transformer Encoder model pre-trained on a large and multilingual textual corpus. Both the visual and textual backbones were frozen during the training.

As a query, we tried two different configurations. As a first experiment, we used a compound query built by employing both the image URL (processed with the textual pipeline) and the image (processed with the image pipeline). In the second configuration, instead, we used only the image URL. These two experiments aim to understand the role of images in solving the matching task, given that the image URL alone seems already sufficient in most cases.

When both the image URL and the image are used as a query, as mentioned in Section 3.1, we used two different fusion techniques: straightforward concatenation or the attentive fusion mechanism.

Discussion Table 1 shows the results reached on the validation set for the different experiments. In particular, we can see that when the images are concatenated to the image URL

Table 2 Caption retrieval results for the Caption Re-Rank model, and candidate selector methods, and their combination, on the small validation set (1,000 elements)

Method	Recall@K K=1	 K=5	 K=10	nDCG_5
Levenshtein similarity	32.2	38.6	40.8	0.356 ± 0.029
XLM-RoBERTa similarity	17.6	24.5	27.5	0.213 ± 0.024
MCProp imgs, att-fusion	57.7	70.9	76.2	0.647 ± 0.023
CRank (100%)	74.2	**82.7**	**86.4**	0.789 ± 0.024
CRank on Levenshtein similarity (20%)	52.6	56.5	57.3	0.548 ± 0.030
CRank on XLM-RoBERTa similarity (20%)	48.8	53.2	53.7	0.512 ± 0.030
CRank on MCProp imgs, att-fusion (20%)	**74.5**	**82.7**	**86.4**	**0.791** ± 0.024

For nDCG_5, we also report the 95% confidence interval on the mean
Bold entries signify best values

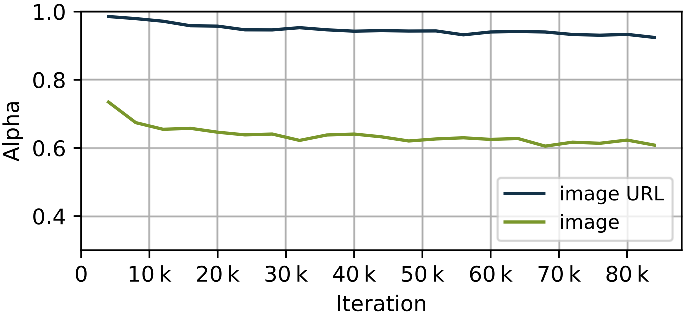


Fig. 4 Weights estimated by the model for the image URL and the image (α_u and α_v) on the validation set as training progresses

information, the overall metrics slightly raise. Nevertheless, the performance increase is not substantial, especially for recall@1, which seems to downgrade (2nd row of Table 1). Better results are obtained when using the attentive fusion approach to merge image URLs with visual information (3rd row of Table 1). The use of the attentive fusion allowed us to inspect the relevance given by the model to the two different modalities that compose the query. Figure 4 shows the evolution of the two attention values on the validation set. As we can notice, the weight provided by the model for the visual pipeline remains at 65% with respect to the weight assigned to the image URL pipeline. This confirms the evidence that the visual information contributes less to the matching task in this scenario.

In Appendix B, we explored another inference methodology that exploited the fact that captions assigned to an image are no more eligible for another image.

4.4 Preliminary results on the CRank model

The CRank model takes as input a (image URL, caption) pair and returns a classification score. The higher the score, the higher the confidence of a match. Given a pool of candidate captions for an image URL, all the (image URL, caption) pairs must be classified by CRank, sorting then the pairs by classification score, higher to lower.

As detailed in Section 3.2, the computational complexity and cost of CRank made this method not directly applicable to the test set without having access to substantial computational resources, which was not our case. In order to test the performance of CRank independently of the candidate selection method, we used a smaller validation set of 1,000 elements (held out from the training data). The small size of this validation set made it possible to apply CRank to the whole set without performing a candidate selection first.

We then applied several methods that act as candidate selectors. Each candidate selector is used to rank all the captions, and then only the top-ranked 20% of the captions are re-ranked using the classification scores produced by CRank. In this way, we can measure which method for selecting candidates works better in combination with CRank.

Discussion Table 2 shows that CRank obtains very high nDCG when applied to the whole validation set, placing the right caption in the right position 74% of the time and 86% of the time among the first 10 results. As comparative baselines, we tested three methods. We selected the top 5 most similar captions using the Levenshtein similarity with the cleaned image URL. Also, we used the embedding vector for the CLS token produced by the XLM-

RoBERTa model before the classification layer for each caption and each image URL to compute pairwise cosine similarity as the measure to select the top 5 captions. Finally, we applied MCProp to the small validation set. Comparing the baselines, it is interesting to see that the XLM-RoBERTa similarity method performs worse than the others. This indicates that the embeddings extracted from XLM-RoBERTa are so specialized for the classification task that they are no longer suitable for language representation. The Levenshtein-based model obtains an average performance. This is a remarkable result considering that this method is fundamentally not able to handle pairs of texts in different languages. On this smaller validation set, the MCProp method performs very well, placing the proper caption in the first position 57% of the time.

We then used the baseline methods as a way to select a reduced set of candidates (20% of the 1,000 elements in the validation set), which are then re-ranked by CRank. This two-steps procedure makes CRank applicable to larger test sets. In all cases, the CRank re-rank produces a sensible improvement over the baselines.

When Levenshtein and XLM-RoBERTa are used the final scores are lower than the maximum achieved by CRank on the full validation set. This indicates that both Levenshtein and XLM-RoBERTa fail to put the correct result in the top 20% of their ranks for a sensible number of cases. On the contrary, the combination of MCProp and CRank produces a slight increase in recall@1 and nDCG, with respect to the use of CRank only. This is due to a positive interaction between the two methods: (i) all the elements placed by CRank in its top 5 positions are also placed by MCProp in its top 20% positions; (ii) a few cases of tie in CRank that caused the top result to be not the correct one are solved by MCProp that puts the tied element in the correct order[5].

4.5 Final results

This section reports the final results obtained on the private leaderboard of the Wikipedia challenge on Kaggle. Until the end of the competition, only the results on the public test set, composed of only 25% of the full test set, were shown to the teams. Five submissions per day were accepted on the public leaderboard. Therefore, teams had the opportunity to optimize their methods on the small test subset. The results in the private leaderboard were instead computed at the end of the challenge, on 85% of the test set. For all these reasons, despite not being publicly accessible, the results in the private leaderboard are more representative of the final model ranking.

In Table 3, we report our models' performance on the private test set. Unfortunately, it was not possible to evaluate confidence intervals as we do not have access to the challenge test set. We can notice how the choices made using the validation sets are well reflected on the wider test corpus. In particular, the two-cascaded model pipeline obtains higher results among all the baseline methods. This confirms the ability of the MCProp model to propose relevant candidates and the proficiency of the CRank to move correct items toward the top of the ranked list.

In the Appendix C, we further tried to eliminate the need to explicitly deal with multiple languages by translating the whole test set into English. This trial did not improve the results reported in Table 3, confirming the strength of multilingual models in solving this complex matching task.

[5] Whenever there is a tie in the CRank ranking, the order of the tied elements in the list of candidate elements is preserved.

Table 3 nDCG$_5$ of our models on the private leaderboard (85% test data)

Similarity model	Base	CRank
Levenshtein	0.215	0.374
XLM-RoBERTa similarity	0.175	0.269
MCProp imgs, att-fusion	**0.421**	**0.533**

The *Base* column reports results obtained using the given similarity model only; the *CRank* column shows the results using these methods to propose candidates for the CRank model (1,000 candidates)
Bold entries signify best values

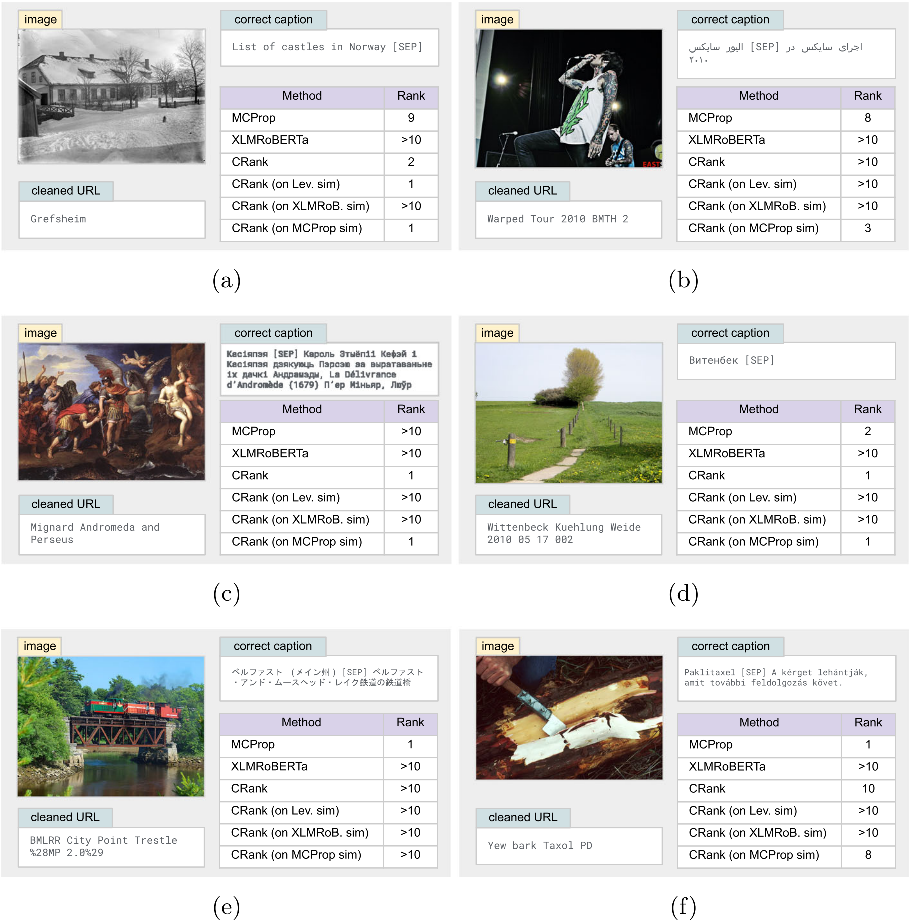
In Table 4, we report the results on the private leaderboard for the top-10 methods. More than 100 teams took part in this challenge, and we positioned fifth. The first participant significantly overperformed the rest of the participants, with the second also performing distinctly well. We obtained a remarkable overall result, performing very similar to the two methods ahead of us and distancing the sixth team by a consistent margin (+8% in nDCG@5). These results confirm the validity of the approach, which already demonstrated promising outcomes on the validation set. The larger performance gap between our method and the top-performing teams can be justified by our choice of keeping the right balance between effectiveness and efficiency and our need for limited training computational resources. However, we argue that these limitations can easily be solved when deploying the method in large production environments offering large computing facilities.

Finally, in Fig. 5, we can qualitatively appreciate the outcomes from the proposed pipeline. We report results on the 1000-item validation set for which we have the ground truth available, and we leave out the Levenshtein distance since, as a baseline, it always produces ranks > 10 in all our reported examples. Figure 5a and b show success cases, where CRank successfully attracts relevant items already found by MCProp towards the top of the list. Figure 5c and d show scenarios in which the CRank alone is already able to retrieve the correct caption due to the clear syntactic correspondence between the URL and the caption – although in different languages – which makes the figure unnecessary. For example, the caption Витенбек in Figure 5d is the Russian translation for *Wittenbeck*, a municipality in northern Germany, also present in the figure URL. Figure 5e and f show failure cases, where MCProp can find

Table 4 nDCG$_5$ of the top-10 performing methods on the private leaderboard

#	Team name	Score
1	新东方人工智能研究院	0.734
2	sigs21group	0.614
3	Basic Go	0.559
4	Sadidul Islam	0.535
5	**Fan Dani (ours)**	**0.533**
6	GMago123	0.492
7	OddAsparagus11	0.444
8	Tetsuro Asano	0.436
9	Lab_EKB	0.432
10	b0w1d	0.414

Bold entries signify best values



image

correct caption

List of castles in Norway [SEP]

Method	Rank
MCProp	9
XLMRoBERTa	>10
CRank	2
CRank (on Lev. sim)	1
CRank (on XLMRoB. sim)	>10
CRank (on MCProp sim)	1

cleaned URL

Grefsheim

(a)

image

correct caption

اجرای سایکس در [SEP] الیور سایکس
۲۰۱۰

Method	Rank
MCProp	8
XLMRoBERTa	>10
CRank	>10
CRank (on Lev. sim)	>10
CRank (on XLMRoB. sim)	>10
CRank (on MCProp sim)	3

cleaned URL

Warped Tour 2010 BMTH 2

(b)

image

correct caption

Kaciяпэя [SEP] Кароль Этыёпii Кефэй i Kaciяпэя дзякуюць Пэрсэю за выратаваньне ix дачкі Андрамэды, La Délivrance d'Andromède (1679) П'ер Міньяр, Люўр

Method	Rank
MCProp	>10
XLMRoBERTa	>10
CRank	1
CRank (on Lev. sim)	>10
CRank (on XLMRoB. sim)	>10
CRank (on MCProp sim)	1

cleaned URL

Mignard Andromeda and Perseus

(c)

image

correct caption

Витенбек [SEP]

Method	Rank
MCProp	2
XLMRoBERTa	>10
CRank	1
CRank (on Lev. sim)	>10
CRank (on XLMRoB. sim)	>10
CRank (on MCProp sim)	1

cleaned URL

Wittenbeck Kuehlung Weide 2010 05 17 002

(d)

image

correct caption

ベルファスト （メイン州）［SEP］ベルファスト・アンド・ムースヘッド・レイク鉄道の鉄道橋

Method	Rank
MCProp	1
XLMRoBERTa	>10
CRank	>10
CRank (on Lev. sim)	>10
CRank (on XLMRoB. sim)	>10
CRank (on MCProp sim)	>10

cleaned URL

BMLRR City Point Trestle %28MP 2.0%29

(e)

image

correct caption

Paklitaxel [SEP] A kérget lehántják, amit további feldolgozás követ.

Method	Rank
MCProp	1
XLMRoBERTa	>10
CRank	10
CRank (on Lev. sim)	>10
CRank (on XLMRoB. sim)	>10
CRank (on MCProp sim)	8

cleaned URL

Yew bark Taxol PD

(f)

Fig. 5 Qualitative examples, showing our approach's success and failure cases. Specifically, examples (a) and (b) show the cases in which CRank successfully attracts relevant items already found by MCProp towards the top of the list; examples (c) and (d) show scenarios in which the CRank alone is already able to retrieve the correct caption; examples (e) and (f) show failure cases, where MCProp can find the correct result in a good position and CRank worsens their rank

the correct result in a good position and CRank worsens their rank. This is probably due to the lack of visual information in CRank which is important for discriminating among many possible existing matching candidates if only the URL is considered.

5 Conclusions

In this paper, we proposed a system able to match images with corresponding multilingual captions. This is an important tool for managing large encyclopedia websites like Wikipedia, where most article images do not have any written context connected to the image. Driven by the power of recent Transformer-based models, we addressed the matching problem using a cascade of two models, Multi-modal Caption Proposal (MCProp), to efficiently propose relevant caption candidates and Caption Re-Rank (CRank), which re-ranks the proposals using a fine-tuned XLM-RoBERTa model. The results obtained on the validation sets show that the MCProp model is an effective model for proposing candidates, and CRank is able to bring the correct results (chosen among the candidates) towards the top of the ranked list.

We participated in the Wikipedia image-caption matching challenge proposed on Kaggle, reaching the fifth position on the private leaderboard among more than 100 participating teams with a system running on limited computational resources.

Although promising, this approach suffers from some known limitations. In particular, it seems that MCProp is not completely exploiting the visual information, and, as shown in qualitative results, the cascaded pipeline cannot always enhance results because CRank is only exploiting textual information. Future research directions include the use of additional contextual data available in the provided dataset during training to regularize the model and improve generalization. Also, it would be interesting to experiment with other fusion techniques for MCProp and exploit visual information also in CRank, or try approaching the problem by distilling efficient vision-language scores from large pre-trained vision-language transformers as done in ALADIN [52].

Appendix

Appendix A: Implementation details

For MCProp, the CLIP image encoder is taken from the official CLIP implementation[6]. Specifically, we used a `ViT-B/32` backbone, employing the pre-trained weights. We took the CLS output from the last encoder layer as the final image embedding. Differently, for the XLM-RoBERTa textual backbone, we employed the Huggingface implementation[7]. Specifically, it uses byte-level Binary Pair Encoding [53] with a vocabulary of 50k tokens, and we used the weights of the model pre-trained on the Masked Language Modeling (MLM) objective [32] on documents in 100 languages from the CommonCrawl dataset [54]. We employed the same XLM-RoBERTa backbone also for CRank, by fine-tuning it on a sentence-matching objective on the available training data. We report a summary of modules configuration in Table 5.

We trained MCProp with a batch size of 64 for 30 epochs, using the Adam optimizer with a learning rate of $1e-5$. Following other previous cross-modal learning works [6, 38], we set $\gamma = 0.2$. We instead trained CRank with a batch size of 64 for 2 epochs, using the AdamW optimizer with a learning rate of $4e - 5$.

Table 5 Modules details

Module	# Layers	# Parameters	# Heads	Pretrained
XLM-RoBERTa	12	125M	8	✓
CLIP (ViT-B/32 Image encoder)	12	150M	8	✓
Feature fusion	–	4M	–	✗
Transformer encoder	2	7M	4	✗

[6] https://github.com/openai/CLIP

[7] https://huggingface.co/docs/transformers/model_doc/xlm-roberta

Appendix B: Bijective matching

This challenge has an objective slightly different from the one underlying search systems. While in search systems two different queries possibly return the same relevant items, here we would like to obtain a bijective matching between the queries and the captions. In fact, every image seems to have its own description that cannot be shared with others.

We try to tackle this problem at inference time by searching for the best assignment between queries and captions. We perform this association through a *linear sum assignment* on the inferred score matrix. Formally, we are given the score matrix $\mathbf{S} \in \mathbb{R}^{n \times n}$. This matrix has n rows, one for each query, and contains in its i-th row the scores computed between the i-th query and all the n captions. The n pairs of query and caption indexes (q, c) that guarantee the best score assignment are obtained by solving the following linear optimization problem:

$$\max \sum_{i=1}^{n} \sum_{j=1}^{n} s_{i,j} x_{i,j} \tag{B1}$$

$$\text{s.t.} \sum_{j=1}^{n} x_{i,j} = 1 \quad (i = 1, 2, \dots, n), \tag{B2}$$

$$\sum_{i=1}^{n} x_{i,j} = 1 \quad (j = 1, 2, \dots, n), \tag{B3}$$

$$x_{i,j} \in \{0, 1\} \quad (i, j = 1, 2, \dots, n). \tag{B4}$$

$\mathbf{X}$ is a permutation matrix found by the optimization algorithm such that $x_{q,c} = 1$.

We need the top-k candidates for each query, and in particular, we require the top-5 for submitting to the challenge. However, a single run of the linear assignment procedure gives us only the top-1 best assignments for each query. To obtain the top-k, we run the linear assignment algorithm k times, every time eliminating from the candidate pool the elements already assigned in the previous round. In this way, the elements assigned at the i-th run are no more eligible at the $(i + 1)$-th run. We report the complete bijective matching procedure in Algorithm 1.

Algorithm 1 Bijective matching.

```
Input
    $\mathbf{S} \in \mathbb{R}^{n\times n}$   Matrix of scores
    $k$              Number of top candidates for each query
Output
    $\mathbf{I} \in \mathbb{N}^{n\times k}$   Indexes of top-k elements for each query
for $t \leftarrow 1$ to $k$ do
    $(\mathbf{q}, \mathbf{c}) \leftarrow$ LinearSumAssignment($\mathbf{S}$)
    $\mathbf{S}[\mathbf{q}, \mathbf{c}] \leftarrow 0$        ▷ Zero out the scores of assigned elements
    $\mathbf{I}[\mathbf{q}, t] = \mathbf{c}$         ▷ Caption idxs are appended to the result
end for
```

The final results for the bijective matching are shown in Table 6. We used the scores from the Multi-modal Caption Proposal model computed on the 10K validation set. The bijective matching methodology seems to obtain the best results on the validation set. Unfortunately, it did not enhance the results obtained on the test set. This could be due to the unbalanced

Table 6 nDCG$_5$ on the multi-modal caption proposal model, with and without the bijective matching on both validation and test sets

Model	Validation	Test
MCProp imgs, att-fusion	0.538	**0.421**
MCProp imgs, att-fusion, bijective	**0.556**	0.419

break Bold entries signify best values

distribution of the output scores, which makes the whole model less and less linear as the size of the inference set increases.

Appendix C: Translating to a common language

The (image URL, caption) pairs processed by the CRank model may be formed by two pieces of text in different languages. Our hypothesis in the definition of CRank was that the multilingual XLM-RoBERTa model, pre-trained on a large multilingual corpus, would be able to classify the match/non-match between image URLs and captions independently of their language.

We also explored an alternative approach in which the captions and pieces of text resulting from cleaning the image URLs (see Section 3.2) are all translated to English before being fed to CRank. The hypothesis supporting this approach is that pairwise translation models should be accurate in projecting all the text into a single language. The match/non-match classification task should be easier when run on a single language. We extended this hypothesis to MCProp, using English-translated text instead of the original multilingual one.

Table 7 reports the results from submissions that use the translation-based approach compared to our reference multilingual approach. The translation impact is negative for both models, so we rejected the hypothesis that translation could benefit the task. Results support the idea of learning to project entities from different sources, either media or language, into a common space without resorting to performing any explicit and more complex translation.

Table 7 Comparison of nDCG$_5$ of MCProp and CRank using the original multilingual text or the English-translated version

Model	Multilingual	Translated
MCProp imgs, att-fusion	0.421	0.395
CRank on MCProp imgs, att-fusion (1,000 candidates)	0.533	0.499

Evaluated on submissions on the private leaderboard

Funding Open access funding provided by Consiglio Nazionale Delle Ricerche (CNR) within the CRUI-CARE Agreement. This work was partially supported by the INAROS project (CNR4C program, CUP B53D210080600-
08), and by the AI4Media project, funded by the EC (H2020 - Contract n. 951911).

Data Availability The dataset analyzed during the current study is available in the Kaggle repository, https://www.kaggle.com/competitions/wikipedia-image-caption/data.

Code Availability The code for reproducing our results is publicly available on GitHub: https://github.com/mesnico/Wiki-Image-Caption-Matching.

Declarations

Conflicts of interest The authors have no relevant financial or non-financial interests to disclose.

References

1. Eken S, Menhour H, Köksal K (2019) Doca: a content-based automatic classification system over digital documents. IEEE Access 7:97996–98004
2. Yurtsever MME, Özcan M, Taruz Z, Eken S, Sayar A (2022) Figure search by text in large scale digital document collections. Concurr Comp-pract Exp 34(1):6529
3. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J et al (2021) Learning transferable visual models from natural language supervision. In: International conference on machine learning, pp 8748–8763. PMLR
4. Sarto S, Cornia M, Baraldi L, Cucchiara R (2022) Retrieval-augmented transformer for image captioning. In: Proceedings of the 19th international conference on content-based multimedia indexing, pp 1–7
5. Rombach R, Blattmann A, Lorenz, D, Esser, P, Ommer, B (2022) High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10684–10695
6. Messina N, Amato G, Esuli A, Falchi F, Gennaro C, Marchand-Maillet S (2021) Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders. ACM Trans Multimed Comput Commun Appl (TOMM) 17(4):1–23
7. Messina N, Falchi F, Esuli A, Amato G (2021) Transformer reasoning network for image-text matching and retrieval. In: 2020 25th International conference on pattern recognition (ICPR), pp 5222–5229. IEEE
8. Amato G, Bolettieri P, Falchi F, Gennaro C, Messina N, Vadicamo L, Vairo C (2021) Visione at video browser showdown 2021. In: International conference on multimedia modeling, pp 473–478. Springer
9. Srinivasan K, Raman K, Chen J, Bendersky M, Najork M (2021) Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In: Proceedings of the 44th International ACM SIGIR conference on research and development in information retrieval, pp 2443–2449
10. Burns A, Srinivasan K, Ainslie J, Brown G, Plummer BA, Saenko K, Ni J, Guo M (2023) Wikiweb2m: A page-level multimodal wikipedia dataset. arXiv:2305.05432
11. Yang J-H, Lassance C, Sampaio De Rezende R, Srinivasan K, Redi M, Clinchant S, Lin J (2023) Atomic: An image/text retrieval test collection to support multimedia content creation. In: Proceedings of the 46th International ACM SIGIR conference on research and development in information retrieval, pp 2975–2984
12. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems, pp 5998–6008
13. Kenton JDM-WC, Toutanova LK (2019) Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT, pp 4171–4186
14. Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018) Deep contextualized word representations. In: Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: human language technologies, vol 1 (Long Papers), pp 2227–2237
15. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A et al (2020) Language models are few-shot learners. Adv Neural Inf Process Syst 33:1877–1901

16. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) Roberta: A robustly optimized bert pretraining approach. arXiv:1907.11692
17. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R (2019) Albert: a lite bert for self-supervised learning of language representations. In: International conference on learning representations
18. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S et al (2020) An image is worth 16x16 words: transformers for image recognition at scale. In: International conference on learning representations
19. Chen, C-FR, Fan, Q, Panda, R (2021) Crossvit: cross-attention multi-scale vision transformer for image classification. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 357–366
20. Chu X, Tian Z, Wang Y, Zhang B, Ren H, Wei X, Xia H, Shen C (2021) Twins: revisiting the design of spatial attention in vision transformers. Adv Neural Inf Process Syst 34:9355–9366
21. Cornia M, Stefanini M, Baraldi L, Cucchiara R (2020) Meshed-memory transformer for image captioning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10578–10587
22. Kamath A, Singh M, LeCun Y, Synnaeve G, Misra I, Carion N (2021) Mdetr-modulated detection for end-to-end multi-modal understanding. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 1780–1790
23. Zhang P, Li X, Hu X, Yang J, Zhang L, Wang L, Choi Y, Gao J (2021) Vinvl: revisiting visual representations in vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5579–5588
24. Kim W, Son B, Kim I (2021) Vilt: Vision-and-language transformer without convolution or region supervision. In: International conference on machine learning, pp 5583–5594. PMLR
25. Alayrac J-B, Donahue J, Luc P, Miech A, Barr I, Hasson Y, Lenc K, Mensch A, Millican K, Reynolds M et al (2022) Flamingo: a visual language model for few-shot learning. Adv Neural Inf Process Syst 35:23716–23736
26. Mao J, Xu W, Yang Y, Wang J, Yuille, AL (2015) Deep captioning with multimodal recurrent neural networks (m-rnn). In: 3rd International conference on learning representations, ICLR
27. Kiros R, Salakhutdinov R, Zemel R (2014) Multimodal neural language models. In: International conference on machine learning, pp 595–603. PMLR
28. Donahue J, Anne Hendricks L, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, Darrell T (2015) Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2625–2634
29. Sharma P, Ding N, Goodman S, Soricut R (2018) Conceptual captions: a cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th annual meeting of the association for computational linguistics (vol 1: Long Papers), pp 2556–2565
30. He S, Liao W, Tavakoli HR, Yang M, Rosenhahn B, Pugeault N (2020) Image captioning through image transformer. In: Proceedings of the asian conference on computer vision
31. Chen C, Mu S, Xiao W, Ye Z, Wu L, Ju Q (2019) Improving image captioning with conditional generative adversarial nets. Proc AAAI Conf Artif Intell 33:8142–8150
32. Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, Grave É, Ott M, Zettlemoyer L, Stoyanov V (2020) Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp 8440–8451
33. Lu J, Batra D, Parikh D, Lee S (2019) Vilbert: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Advances in neural information processing systems, pp 13–23
34. Qi D, Su L, Song J, Cui E, Bharti T, Sacheti A (2020) Imagebert: cross-modal pre-training with large-scale weak-supervised image-text data. arXiv:2001.07966
35. Huang Z, Zeng Z, Liu B, Fu D, Fu J (2020) Pixel-bert: aligning image pixels with text by deep multi-modal transformers. arXiv:2004.00849
36. Su W, Zhu X, Cao Y, Li B, Lu L, Wei F, Dai J (2020) Vl-bert: pre-training of generic visual-linguistic representations. In: International conference on learning representations
37. Faghri F, Fleet DJ, Kiros JR, Fidler S (2018) VSE++: improving visual-semantic embeddings with hard negatives. In: BMVC 2018, p 12
38. Li K, Zhang Y, Li K, Li Y, Fu Y (2019) Visual semantic reasoning for image-text matching. ICCV 2019:4653–4661
39. Qu L, Liu M, Cao D, Nie L, Tian Q (2020) Context-aware multi-view summarization network for image-text matching. In: Proc. of the 28th ACM international conference on multimedia, pp 1047–1055
40. Wu Y, Wang S, Song G, Huang Q (2019) Learning fragment self-attention embeddings for image-text matching. In: Proc. of the 27th ACM international conference on multimedia, pp 2088–2096
41. Sarafianos N, Xu X, Kakadiaris IA (2019) Adversarial representation learning for text-to-image matching. In: Proc. of the IEEE international conference on computer vision, pp 5814–5824

42. Guo Y, Yuan H, Zhang K (2020) Associating images with sentences using recurrent canonical correlation analysis. Appl Sci 10(16):5516
43. Vo N, Jiang L, Sun C, Murphy K, Li L-J, Fei-Fei L, Hays J (2019) Composing text and image for image retrieval - an empirical odyssey. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)
44. Li Z, Fan Z, Chen J, Zhang Q, Huang X-J, Wei Z (2023) Unifying cross-lingual and cross-modal modeling towards weakly supervised multilingual vision-language pre-training. In: Proceedings of the 61st annual meeting of the association for computational linguistics (vol 1: Long Papers), pp 5939–5958
45. Jain A, Guo M, Srinivasan K, Chen T, Kudugunta S, Jia C, Yang Y, Baldridge J (2021) Mural: multimodal, multitask retrieval across languages. arXiv:2109.05125
46. Jia C, Yang Y, Xia Y, Chen Y-T, Parekh Z, Pham H, Le Q, Sung Y-H, Li Z, Duerig T (2021) Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning, pp 4904–4916. PMLR
47. Hu Z, Iscen A, Sun C, Wang Z, Chang K-W, Sun Y, Schmid C, Ross DA, Fathi A (2023) Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 23369–23379
48. Nguyen K, Biten AF, Mafla A, Gomez L, Karatzas D (2023) Show, interpret and tell: entity-aware contextualised image captioning in wikipedia. Proc AAAI Conf Artif Intell 37:1940–1948
49. Hazarika D, Gorantla S, Poria S, Zimmermann R (2018) Self-attentive feature-level fusion for multimodal emotion detection. In: 2018 IEEE Conference on multimedia information processing and retrieval (MIPR), pp 196–201.IEEE
50. Hori C, Hori T, Lee T-Y, Zhang Z, Harsham B, Hershey JR, Marks TK, Sumi K (2017) Attention-based multimodal fusion for video description. In: Proceedings of the IEEE international conference on computer vision, pp 4193–4202
51. Loshchilov I, Hutter F (2018) Decoupled weight decay regularization. In: International conference on learning representations
52. Messina N, Stefanini M, Cornia M, Baraldi L, Falchi F, Amato G, Cucchiara R (2022) Aladin: distilling fine-grained alignment scores for efficient image-text matching and retrieval. In: Proceedings of the 19th international conference on content-based multimedia indexing, pp 64–70
53. Wang C, Cho K, Gu J (2020) Neural machine translation with byte-level subwords. Proc AAAI Conf Artif Intell 34:9154–9160
54. Wenzek G, Lachaux M-A, Conneau A, Chaudhary V, Guzmán F, Joulin A, Grave E (2020) CCN et: extracting high quality monolingual datasets from web crawl data. In: Proceedings of the 12th language resources and evaluation conference, pp 4003–4012. European Language Resources Association, Marseille, France. https://aclanthology.org/2020.lrec-1.494

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.