This is the peer reviewd version of the followng article:

ALADIN: Distilling Fine-grained Alignment Scores for Efficient Image-Text Matching and Retrieval / Messina, N., Stefanini, M., Cornia, M., Baraldi, L., Falchi, F., Amato, G., Cucchiara, R.. - (2022), pp. 64-70. (19th International Conference on Content-based Multimedia Indexing, CBMI 2022 Graz, Austria Sept. 14-16, 2022) [10.1145/3549555.3549576].

Association for Computing Machinery
*Terms of use:*

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

09/11/2026 06:13

(Article begins on next page)

# ALADIN: Distilling Fine-grained Alignment Scores for Efficient Image-Text Matching and Retrieval

Nicola Messina
ISTI-CNR
Pisa, Italy
nicola.messina@isti.cnr.it

Matteo Stefanini
Univ. of Modena and Reggio Emilia
Modena, Italy
matteo.stefanini@unimore.it

Marcella Cornia
Univ. of Modena and Reggio Emilia
Modena, Italy
marcella.cornia@unimore.it

Lorenzo Baraldi
Univ. of Modena and Reggio Emilia
Modena, Italy
lorenzo.baraldi@unimore.it

Fabrizio Falchi
ISTI-CNR
Pisa, Italy
fabrizio.falchi@isti.cnr.it

Giuseppe Amato
ISTI-CNR
Pisa, Italy
giuseppe.amato@isti.cnr.it

Rita Cucchiara
Univ. of Modena and Reggio Emilia
Modena, Italy
rita.cucchiara@unimore.it

## ABSTRACT

Image-text matching is gaining a leading role among tasks involving the joint understanding of vision and language. In literature, this task is often used as a pre-training objective to forge architectures able to jointly deal with images and texts. Nonetheless, it has a direct downstream application: *cross-modal retrieval*, which consists in finding images related to a given query text or vice-versa. Solving this task is of critical importance in cross-modal search engines. Many recent methods proposed effective solutions to the image-text matching problem, mostly using recent large vision-language (VL) Transformer networks. However, these models are often computationally expensive, especially at inference time. This prevents their adoption in large-scale cross-modal retrieval scenarios, where results should be provided to the user almost instantaneously. In this paper, we propose to fill in the gap between effectiveness and efficiency by proposing an ALign And DIstill Network (ALADIN). ALADIN first produces high-effective scores by aligning at fine-grained level images and texts. Then, it learns a shared embedding space – where an efficient kNN search can be performed – by distilling the relevance scores obtained from the fine-grained alignments. We obtained remarkable results on MS-COCO, showing that our method can compete with state-of-the-art VL Transformers while being almost 90 times faster. The code for reproducing our results is available at https://github.com/mesnico/ALADIN.

## CCS CONCEPTS

• **Information systems** → **Information retrieval**; **Multimedia and multimodal retrieval**; • **Computing methodologies** → **Neural networks**; **Computer vision**; **Visual content-based indexing and retrieval**; **Matching**.

## KEYWORDS

cross-modal retrieval, image-text matching, vision-and-language.

*CBMI 2022, Sept. 14-16, 2022, Graz, Austria*
© 2022 Association for Computing Machinery.
This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in *Proceedings of International Conference on Content-based Multimedia Indexing (CBMI 2022).*

**ACM Reference Format:**
Nicola Messina, Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Fabrizio Falchi, Giuseppe Amato, and Rita Cucchiara. 2022. ALADIN: Distilling Fine-grained Alignment Scores for Efficient Image-Text Matching and Retrieval. In *Proceedings of International Conference on Content-based Multimedia Indexing (CBMI 2022).* ACM, New York, NY, USA, 7 pages.

## 1 INTRODUCTION

With the growing strength of deep learning methods and the availability of large-scale data, multi-modal processing has become one of the most promising research topics. In particular, most of the focus is placed on the joint processing of images and natural language sentences. By understanding the hidden semantic connections between a text and an image, many works in literature solved challenging multi-modal problems, such as image captioning [1, 10, 33] or visual question answering [1, 3, 41]. Among these tasks, *image-text matching* has crucial importance [9, 12, 16, 25, 26]: it consists of outputting a relevance score for each given (image, text) pair, where the score is high if the image is relevant to the text and low otherwise. Although this task is usually employed as a vision-language pre-training objective, it is crucial for cross-modal *image-text retrieval*, which usually consists of two sub-tasks: *image retrieval*, where we want images relevant to a given text, and *text retrieval*, where we ask for sentences better describing an input image. Efficiently and effectively solving these retrieval tasks is strategically important in modern cross-modal search engines.

Many state-of-the-art models for image-text matching, like Oscar [21] or UNITER [8], comprise large and deep multi-modal vision-language (VL) Transformers with early fusion, which are computationally expensive, especially during the inference phase. In fact, during inference, all the (image, text) pairs from the test set should be forwarded through the multi-modal Transformer to obtain the relevance scores. This is clearly unfeasible in large datasets and unusable in large-scale retrieval scenarios, where the system latency should be as small as possible.

For achieving such a performance objective, many approaches in the literature project image and text embeddings in a common space where similarity is measured through simple dot products. This allows the introduction of an *offline* phase, in which all the dataset items are encoded and stored, and an *online* phase in which

only the query is forwarded through the network and compared with all the offline-stored elements. Although these approaches are very efficient, they are usually not sufficiently effective as the ones relying on early modality fusion using large VL Transformers.

In the light of these observations, in this paper we propose an ALign And DIstill Network model (*ALADIN*), which exploits the knowledge acquired by large VL Transformers to craft an efficient yet effective model for image-text retrieval. In particular, we employ late fusion approaches so that the two visual and textual pipelines are kept separated until the final matching phase. The first objective consists of *aligning* image regions with sentence words, using a simple yet effective alignment head. Then, a common visual-textual embedding space is learned by distilling the scores from the alignment head using a learning-to-rank objective. In this case, we use the learned alignment scores as ground-truth (teacher) scores.

We show that, on the widely used MS-COCO dataset, the alignment scores can reach results comparable with large joint vision-language models such as UNITER and OSCAR, while being far more efficient, especially during inference. On the other hand, the distilled scores used to learn the common space can defeat previous common space methods on the same dataset, opening the way toward metric-based indexing for large-scale retrieval.

To sum up, in this paper, we propose the following contributions:

- We employ two instances of a pre-trained VL Transformer as a backbone for extracting separate visual and textual features.
- We adopt a simple yet effective alignment method for producing high-quality scores instead of the poorly-scalable output of large joint VL Transformers.
- We create an informative embedding space by framing the problem as a learning-to-rank task and distilling the final scores using the scores in output from the alignment head.

## 2 RELATED WORK

In the last years, many works tackled the image-text matching task. The work in [12] paved the way for the common space approach for cross-modal matching. They showed the effectiveness of the hinge-based triplet ranking loss with hard-negative mining. Many works followed their footsteps [20, 25, 27, 30, 34, 36], trying out BERT [11] as a text extractor other than a simple GRU and showing the effectiveness of region-based features [1] as visual representation. After the success of BERT-like models in Natural Language Processing [11, 18, 22], many works tried to employ the Transformer Encoder to jointly process images and text, like VilBERT [23], OSCAR [21], VL-BERT [35], or VinVL [40]. These methods tackle image-text matching as a binary classification problem, where an (image, sentence) pair is input to the complex Transformer architecture which is trained to predict the probability that the sentence relates to the image. Although these architectures are very effective, they are computationally expensive at inference time, as they need to process every (image, sentence) pair to obtain the scores on the whole test set. For this reason, many methods keep the visual and textual pipelines separated, without cross-talking between them [13, 25, 27, 32, 36]. Doing so, they can be forwarded independently at inference time, at the cost of losing effectiveness. Our work is inspired by the recent success of knowledge distillation [2, 4, 7, 38, 42], used to transfer knowledge from a large model to a smaller and more efficient one. We propose to use scores distillation to learn a visual-textual common space, employing the knowledge acquired by a pre-trained VL Transformer. In this case, the knowledge distillation is framed as a learning-to-rank problem [5, 6, 28], widely used in literature but, as far as we know, never used for distilling cross-modal scores.

## 3 PROPOSED METHOD

The proposed architecture is composed of two different stages. The first stage, which we refer to as *backbone*, is composed of the layers of a pre-trained large vision-language transformer – VinVL [40], an extension to the powerful OSCAR model [21]. In the backbone, the language and the visual paths do not interact through cross-attention mechanisms so that the features from the two modalities can be extracted independently at inference time.

The second stage, instead, is composed of two separate *heads*: the *alignment* head and *matching* head. The alignment head is used to pre-train the network to efficiently align the visual and the textual concepts in a fine-grained manner, as done in TERAN [25]. Differently, the matching head is used to construct an informative cross-modal common space, that can be used to efficiently represent images and text as fixed-length vectors for use in large-scale retrieval. The scores from the matching head are distilled using the scores from the alignment head as guidance. The overall architecture is shown in Figure 1.

In the following, we dive into the building blocks of the architecture – *i.e.*, the backbone, the alignment head, and the matching head.

### 3.1 Vision-Language Backbone

As the backbone for feature extraction, we use the pre-trained layers from VinVL [40], an extension to the large-scale vision-language OSCAR model [21]. Our goal is to obtain suitable vectorial representations for the image $\mathcal{V}$ and the text $C$ in input. In particular, we employ the model pre-trained on the image-text retrieval task. The authors used a binary classification head on top of the CLS token of the output sequence, and the model is trained to predict if the input images and textual sentences are related or not.

In our use case, the visual and textual pipelines should be separated, so that they can be forwarded independently at inference time. For this reason, we use two instances of the VinVL architecture, in a shared-weights configuration to forward the two modalities independently, as shown in Figure 1.

As in [40], we use as visual tokens both the visual features extracted from object regions[1] and their labels, and the two subsequences are separated by a SEP token. In the end, the outputs from the last layers of the disentangled VinVL architecture are two sequences, $V = \{v_{\text{cls}}, v_1, v_2, \ldots, v_N\}$, representing the image $\mathcal{V}$, and $C = \{c_{\text{cls}}, c_1, c_2, \ldots, c_M\}$, representing the text $C$. Note that, in both sequences, the first element is the CLS token, used to collect representative information for the whole image or text.

[1] https://github.com/microsoft/scene_graph_benchmark

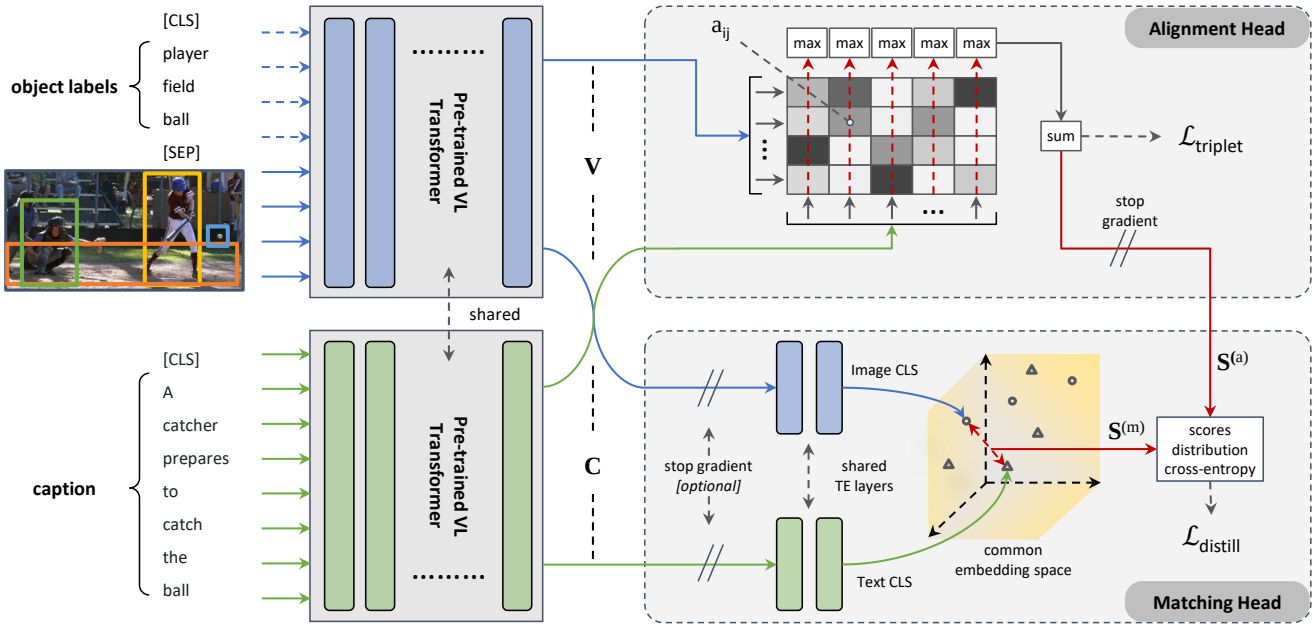


**Figure 1: Overview of our architecture. The backbone extracts visual and textual features that are used in both the matching and alignment heads. The matching head is trained by distilling the scores using the ones coming from the alignment head.**

## 3.2 Alignment Head

The alignment head comprises a similarity matrix that computes the fine-grained relevances between the visual tokens $\boldsymbol{V}$ and textual tokens $\boldsymbol{C}$. The fine-grained similarities are then pooled to obtain the final global relevance between the image and the text. In particular, we use a formulation similar to the one used in TERAN [25]. Specifically, the features in output from the backbone are used to compute a visual-textual tokens alignment matrix $\boldsymbol{A} \in \mathbb{R}^{n \times m}$, built as follows:

$$\boldsymbol{A} = a_{ij}^{kl} = \text{cosine}(\boldsymbol{v}_i, \boldsymbol{c}_j) = \frac{\boldsymbol{v}_i^T \boldsymbol{c}_j}{\|\boldsymbol{v}_i\| \|\boldsymbol{c}_j\|} \qquad i \in g_k, j \in g_l, \quad (1)$$

where $g_k$ is the set of indexes of the region features from the $k$-th image and $g_l$ is the set of indexes of the words from the $l$-th sentence. At this point, the similarities $s_{kl}$ between the image $k$ and the caption $l$ are computed by pooling the similarity matrix $\boldsymbol{A}$ along dimensions $(i, j)$ through an appropriate pooling function. Guided by [25], we use the *max-over-regions sum-over-words* policy, which computes the following final similarity score:

$$S^{(a)} = s_{kl}^{(a)} = \sum_{j \in g_l} \max_{i \in g_k} A_{ij}. \quad (2)$$

The dot-product similarity used to compute $\boldsymbol{A}$ in Eq. 1 resembles the computation of the cross-attention between visual and textual tokens. The difference boils down to the interaction between the visual and textual pipelines, which happens only at the very end of the whole architecture. This *late cross-attention* makes the sequences $\boldsymbol{V}$ and $\boldsymbol{C}$ cacheable, eliminating the need to forward the whole architecture whenever a new query – either visual or textual – is issued to the system. The computation of $S^{(a)}$, involving only simple non-parametric operations, is very efficient and can be easily implemented on GPU to obtain high inference speeds.

The loss function used to force this network to produce suitable similarities $s$ for each (image, text) pair is the *hinge-based triplet ranking loss*, used in previous works [12, 20, 25]. Formally,

$$\mathcal{L}_{\text{triplet}} = \sum_{k,l} \max_{l'} [\alpha + s_{kl'} - s_{kl}]_+ + \max_{k'} [\alpha + s_{k'l} - s_{kl}]_+, \quad (3)$$

where $s_{kl}$ is the similarity estimated between image $k$ and caption $l$, and $[x]_+ \equiv \max(0, x)$; the values $k', l'$ are the indexes of the image and caption hard negatives found in the mini-batch as done in [12], and $\alpha$ is a margin that defines the minimum separation that should hold between positive and negative pairs.

Given that the alignment head is directly connected to the backbone, we fine-tuned the backbone on this new alignment objective. More details on the training procedure are reported in Section 3.4.

## 3.3 Matching Head

The matching head uses the same sequences $\boldsymbol{V}$ and $\boldsymbol{C}$ given from the backbone and employs them to produce the features $\tilde{\boldsymbol{v}} \in \mathbb{R}^d$ for the image $\mathcal{V}$ and $\tilde{\boldsymbol{c}} \in \mathbb{R}^d$ for the caption $C$. These representations are forced to lay in the same $d$-dimensional embedding space. In this space, $k$-neirest-neighbor search can be efficiently computed — using metric space approaches or inverted files — to quickly retrieve images given a textual query or vice-versa. Specifically, we forward $\boldsymbol{V}$ and $\boldsymbol{C}$ through a 2-layer Transformer Encoder (TE):

$$\bar{\boldsymbol{V}} = \text{TE}(\boldsymbol{V}); \qquad \bar{\boldsymbol{C}} = \text{TE}(\boldsymbol{C}). \quad (4)$$

As in [27], the TE shares the weights among the two modalities, and the final vectors encoding the whole image and caption are the CLS tokens in output from the TE layers: $\tilde{\boldsymbol{v}} = \bar{\boldsymbol{V}}[0] = \bar{\boldsymbol{v}}_{\text{cls}}$ and

$\tilde{\boldsymbol{c}} = \tilde{\boldsymbol{C}}[0] = \tilde{\boldsymbol{c}}_{\text{cls}}$. The final relevances are simply computed as the cosine similarities between the the vector $\tilde{\boldsymbol{v}}_k$ from the $k$-th image and $\tilde{\boldsymbol{s}}_l$ from the $l$-th sentence: $\boldsymbol{S}^{(\text{m})} = s_{kl}^{(\text{m})} = \text{cosine}(\tilde{\boldsymbol{v}}_k, \tilde{\boldsymbol{s}}_l)$.

In principle, we could optimize the common space using the same hinge-based triplet ranking loss in Eq. 3 already used to train the alignment head. Instead, in the light of the good effectiveness-efficiency trade-off of the alignment head, we propose to learn a distribution for $\boldsymbol{S}^{(\text{m})}$ using the previously-learned $\boldsymbol{S}^{(\text{a})}$ as teachers.

Specifically, we frame the problem of distilling the distribution of $\boldsymbol{S}^{(\text{m})}$ from $\boldsymbol{S}^{(\text{a})}$ as a *learning-to-rank* problem. We employ the mathematical framework developed in the ListNet approach [6], which models the probability of an object being ranked at the top, given the scores of all the objects. Differently from this framework, here we need to optimize for two different entangled distributions: the distribution of text-image similarities when sentences are used as queries, and the distribution of image-text similarities when instead images are used as queries. In particular, given a textual query $k$ and an image query $l$, the probabilities of the image $i$ and text $j$ to be the top-one elements respectively with respect to $\boldsymbol{S}^{(\text{a})}$ are:

$$P_{\boldsymbol{S}^{(\text{a})}}(i) = \frac{\exp(s_{ik}^{(\text{a})})}{\sum_{t=1}^{B} \exp(s_{tk}^{(\text{a})})}; P_{\boldsymbol{S}^{(\text{a})}}(j) = \frac{\exp(s_{lj}^{(\text{a})})}{\sum_{t=1}^{B} \exp(s_{tj}^{(\text{a})})} \tag{5}$$

where $B$ is the batch size, as the learning procedure is confined to the images and sentences in the current batch. Therefore, during training, only $B$ images are retrieved using the query $k$, and $B$ textual elements are retrieved using the query $l$. Similarly, an analogous probability can be defined over $\boldsymbol{S}^{(\text{m})}$:

$$P_{\boldsymbol{S}^{(\text{m})}}(i) = \frac{\exp(\tau s_{ik}^{(\text{m})})}{\sum_{t=1}^{B} \exp(\tau s_{tk}^{(\text{m})})}; P_{\boldsymbol{S}^{(\text{m})}}(j) = \frac{\exp(\tau s_{lj}^{(\text{m})})}{\sum_{t=1}^{B} \exp(\tau s_{tj}^{(\text{m})})} \tag{6}$$

where $\tau$ is a temperature hyper-parameter which compensates for the fact that $\boldsymbol{S}^{(\text{m})}$ ranges in $[0, 1]$. We empirically found that $\tau = 6.0$ works well in practice. The final matching loss can be formulated as the cross-entropy between the $P_{\boldsymbol{S}^{(\text{a})}}$ and $P_{\boldsymbol{S}^{(\text{m})}}$ probabilities, for both the image-to-text and text-to-image cases.

$$\mathcal{L}_{\text{distill}} = -\sum_{i=1}^{B} P_{\boldsymbol{s}^{(\text{a})}(i)} \log(P_{\boldsymbol{s}^{(\text{m})}}(i)) - \sum_{j=1}^{B} P_{\boldsymbol{s}^{(\text{a})}(j)} \log(P_{\boldsymbol{s}^{(\text{m})}}(j)) \tag{7}$$

Notice that accurate and dense teacher scores are needed to obtain a good estimate of the teacher distributions $P_{\boldsymbol{s}^{(\text{a})}}(i)$ and $P_{\boldsymbol{s}^{(\text{a})}}(j)$. This partly motivates our choice of first researching an effective and efficient alignment head that could output the scores to be used as ground-truth for the matching head.

### 3.4 Training

During the training phase, we initially respect the following constraints: (a) the backbone is finetuned only when training the alignment head, and (b) the gradients do not flow backward through $\boldsymbol{S}^{(\text{a})}$ when training the matching head (as depicted in Figure 1 through the *stop-gradient* indication). The constraint (b) comes from the fact that the scores $\boldsymbol{S}^{(\text{a})}$ are used as teacher scores. Therefore, they should not modify the weights of the backbone, because it is assumed that the backbone is already trained with the alignment head. Given these constraints, we train the network in two steps. First, we train the alignment head by updating the backbone weights using $\mathcal{L}_{\text{triplet}}$ (**ALADIN A/ft.** in the experiments). Then, we freeze the backbone and we learn the matching head by updating the weights of the 2-layer Transformer Encoder using $\mathcal{L}_{\text{distill}}$ (**ALADIN D** in the experiments). Note that the formalism *X/ft.* signifies that the gradients coming from that head loss X are used to finetune the backbone. Possible head losses are X={T, D, A} for T=triplet, D=distillation, and A=alignment, where T and D come from the matching head, while A from the alignment head. When */ft.* is omitted, it means that the backbone remains frozen.

We explore also the joint training of the two heads. Specifically, we relax constraint (a), so that gradients coming from the two heads can update the backbone. Sticking to the previous formalism, we refer to this experiment as **ALADIN A/ft. + D/ft.**. Nevertheless, when directly applying this training schema, we experienced some instabilities. If the alignment head — working as a teacher for the matching head — is not warmed-up, it can not initially provide good teacher scores. The consequence is that noisy gradients backpropagate through the matching head and interfere with the finetuning of the backbone. For this reason, we warmup the backbone by pre-training it with the alignment loss $\mathcal{L}_{\text{triplet}}$ (as in the *ALADIN A/ft.* setup).

## 4 EXPERIMENTS

In this section, we report detailed results for validating our approach. In addition to the training setups described in 3.4, we consider two more schemes as baselines: **ALADIN T** trains the matching head using the standard hinge-based triplet ranking loss without distillation, starting from a pre-trained backbone (*i.e.* ALADIN A/ft.) and leaving it fixed; similarly, **ALADIN T/ft.** lacks the alignment head and the backbone is finetuned only with the gradients from the matching head.

### 4.1 Dataset and Metrics

We perform our experiments on the widely-used MS-COCO dataset, which contains a large corpus of images scraped from the web. Each image is annotated with 5 textual descriptions. We follow the splits introduced by [15], which reserves 113,287 images for training, 5,000 for validating, and 5,000 for testing. In literature, a smaller test set comprising only 1,000 images is often used. For a fair comparison, we report the results on both 5K and 1K test sets. In the case of 1K images, the results are computed by performing a 5-fold cross-validation and averaging the results.

As commonly done to evaluate cross-modal retrieval models [12, 17, 20, 23, 29], we use the recall@$k$ metric for evaluating the ability of our model to correctly retrieve relevant texts or images. Specifically, the recall@$k$ measures the percentage of queries able to retrieve the correct item among the first $k$ results.

### 4.2 Alignment Head Results

We first compare the results obtained with our alignment head against some recent methods comprising large-scale pre-trained Transformer models (Table 1). We consider only the *Base* versions and not the *Large* ones, for hardware limitations. For a fair comparison, we initialize our backbone with the weights of VinVL Base [40].

**Table 1: Experiment results using scores from the alignment head. The comparison is performed with entangled visual-textual Transformer models.**

| | | 1K Test Set | | | | | | 5K Test Set | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Text Retrieval | | | Image Retrieval | | | Text Retrieval | | | Image Retrieval | | |
| **Model** | Training Data | $k=1$ | $k=5$ | $k=10$ | $k=1$ | $k=5$ | $k=10$ | $k=1$ | $k=5$ | $k=10$ | $k=1$ | $k=5$ | $k=10$ |
| 12-in-1 [24] | 4.4M | - | - | - | 65.2 | 91.0 | 96.2 | - | - | - | - | - | - |
| VilBERT [23] | 3.1M | - | - | - | 58.2 | 84.9 | 91.5 | - | - | - | - | - | - |
| Unicoder-VL [19] | 3.8M | 84.3 | 97.3 | 99.3 | 69.7 | 93.5 | 97.2 | 62.3 | 87.1 | 92.8 | 46.7 | 76.0 | 85.3 |
| UNITER (Base) [8] | 5.6M | - | - | - | - | - | - | 63.3 | 87.0 | 93.1 | 48.4 | 76.7 | 85.9 |
| OSCAR (Base) [21] | 6.5M | - | - | - | - | - | - | 70.0 | 91.1 | 95.5 | 54.0 | 80.8 | 88.5 |
| VinVL (Base) [40] | 8.9M | - | - | - | - | - | - | **74.6** | **92.6** | **96.3** | **58.1** | **83.2** | **90.1** |
| **ALADIN A/ft.** | 8.9M | **88.1** | **99.1** | **99.7** | **75.4** | **95.2** | 97.9 | 70.0 | 90.7 | 95.6 | 54.4 | 81.0 | 88.6 |
| **ALADIN A/ft. + D/ft.** | 8.9M | 87.6 | 98.5 | **99.7** | 75.0 | **95.2** | **98.0** | 69.9 | 91.3 | 95.7 | 54.7 | 81.0 | 88.7 |

**Table 2: Experimental results using scores from the matching head. The comparison is performed with methods using disentangled visual-textual pipelines.**

| | | 1K Test Set | | | | | | 5K Test Set | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Text Retrieval | | | Image Retrieval | | | Text Retrieval | | | Image Retrieval | | |
| **Model** | Training Data | $k=1$ | $k=5$ | $k=10$ | $k=1$ | $k=5$ | $k=10$ | $k=1$ | $k=5$ | $k=10$ | $k=1$ | $k=5$ | $k=10$ |
| TERN [27] | 0.6M | 65.5 | 91.0 | 96.5 | 54.5 | 86.9 | 94.2 | 40.2 | 71.1 | 81.9 | 31.4 | 62.5 | 75.3 |
| SAEM (ens.) [37] | 0.6M | 71.2 | 94.1 | 97.7 | 57.8 | 88.6 | 94.9 | - | - | - | - | - | - |
| CAMERA (ens.) [30] | 0.6M | 77.5 | 96.3 | 98.8 | 63.4 | 90.9 | 95.8 | 55.1 | 82.9 | 91.2 | 40.5 | 71.7 | 82.5 |
| TERAN (ens.) [25] | 0.6M | 80.2 | 96.6 | 99.0 | 67.0 | 92.2 | 96.9 | 59.3 | 85.8 | 92.4 | 45.1 | 74.6 | 84.4 |
| DSRAN (w. BERT) [36] | 0.6M | 80.6 | 96.7 | 98.7 | 64.5 | 90.8 | 95.8 | 57.9 | 85.3 | 92.0 | 41.7 | 72.7 | 82.8 |
| **ALADIN T** | 8.9M | 79.2 | 96.7 | 99.1 | 68.9 | 92.8 | 96.6 | 57.9 | 84.8 | 91.8 | 46.0 | 74.8 | 84.1 |
| **ALADIN D** | 8.9M | 83.1 | 97.4 | 99.3 | 70.5 | 93.6 | 97.3 | 62.7 | 87.5 | 93.5 | 47.4 | 76.2 | 85.4 |
| **ALADIN T/ft.** | 8.9M | **84.9** | **98.5** | 99.6 | 71.9 | 93.8 | 97.0 | 63.6 | 87.4 | 93.5 | 49.7 | 77.7 | 86.3 |
| **ALADIN A/ft. + D/ft.** | 8.9M | 84.7 | 98.0 | **99.8** | **72.7** | **94.5** | **97.5** | **64.9** | **88.6** | **94.5** | **51.3** | **79.2** | **87.5** |
| CLIP (0-shot) [31] | *0.4B* | - | - | - | - | - | - | *58.4* | *81.5* | *88.1* | *37.8* | *62.4* | *72.2* |
| ALIGN [14] | *1.8B* | - | - | - | - | - | - | *77.0* | *93.5* | *96.9* | *59.9* | *83.3* | *89.8* |

Notice that, at test time, all the reported models except ours need to compute a number of network forward steps in the order of $O(n^2 r)$, where $n$ is the number of images and $r$ is the number of sentences associated to each image ($r = 5$ in case of MS-COCO). In fact, due to cross-attention links between visual and textual pipelines, intermediate representations cannot be cached for being reused with a different query. Instead, given the disentangled pipelines, our model enables caching of the image and text features in output from the backbone for speeding up the retrieval with never seen queries, with a number of network forward steps in the order of $O(n + nr)$. As we can notice from Table 1, this disentanglement comes at the cost of a slight reduction of the overall effectiveness, as we can notice by comparing our approach to the VinVL model. Nevertheless, our model ALADIN A/ft. can perfectly compete, and partially overtake, all the previous entangled visual-textual Transformer models on both image and sentence retrieval tasks. From the results on the ALADIN A/ft. + D/ft. model, we can notice that when the distillation loss is also active the alignment scores are pretty comparable to ALADIN A/ft. In particular, on the 5K test set, we observe slight improvements in both image and sentence retrieval. This evidence suggests that the distillation loss has the collateral effect of regularizing its own teacher scores, as done in recent works on self-distillation [7, 39].

## 4.3 Matching Head Results

We compare the common space created from our matching head with other disentangled methods using similar approaches. The results are shown in Table 2. As explained above, for comparison we report also the matching head directly trained using the hinge-based triplet loss (ALADIN T and ALADIN T/ft.) without distilling the scores from the alignment head. Furthermore, for completeness, we report also the results from the recent methods CLIP (0-shot) [31] and ALIGN [14]. Although the comparison with CLIP (0-shot) may result unfair, we decided to stick with the results obtained by the authors of the original paper, to avoid all the intricacies deriving from the hyper-parameter tuning phase needed for a satisfactory fine-tuning stage. However, these models use from 100× to 1000× more training data, so we exclude them from the analysis.

All of our methods outperform the previous models, notably surpassing TERAN [25], the method that introduced the alignment matrix used in the alignment head. Concerning the experiments that non-finetune the backbones (ALADIN T and ALADIN D), we argue that scores distillation helps, especially in the recall@1, where

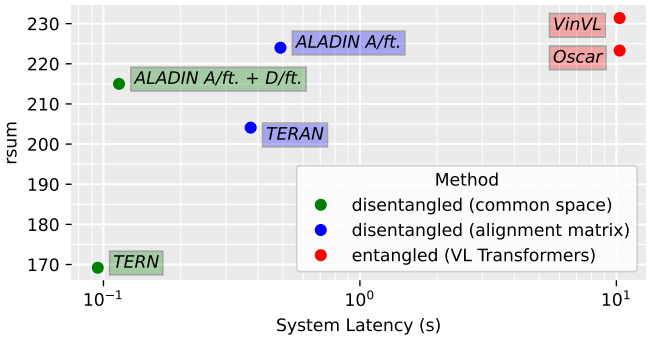


**Figure 2: Effectiveness vs efficiency. We report effectiveness as the sum of recall values on the image retrieval (rsum), and efficiency as the time needed to search the 5K test images.**

we observe an improvement of about 8% and 2% on sentence and image retrieval respectively for the 5K test set. We obtain the best results by using our model ALADIN A/ft. + D/ft., which jointly trains the alignment and distillation heads by also finetuning the backbone with the respective gradients. The alignment scores from this setup already proved to be effective in Table 1. The distilled scores in output from the matching head follow the same trend, obtaining the best results on the 5K test set.

### 4.4 Effectiveness vs Efficiency

To better show the advantage of our model in terms of computing times, in Figure 2 we plot the effectiveness vs the efficiency of our approach compared with other methods. We address image-retrieval on the 1K test set, and we report the sum of the recall values (rsum) versus the average time needed to solve a textual query. These experiments are run on a system equipped with an RTX 2080Ti and an AMD Ryzen 7 1700 Eight-Core Processor. As we can notice, the scores from the alignment head (*ALADIN A/ft.*) can directly compete with VL Transformer models, although being almost 20 times faster. Notably, the scores computed on the distilled space from *ALADIN A/ft. + D/ft.* obtain a speedup of almost 90×, with a rsum loss of only about 7% with respect to VinVL. Therefore, the proposed models help fill the gap between efficiency and effectiveness – *i.e.*, the top left zone of the diagram.

Considering the efficiency-effectiveness trade-offs of both the alignment and matching heads, the whole architecture could be deployed in real application scenarios in a two-stage configuration: first, the faster matching head proposes relevant candidates using k-NN search on the common space; then, the candidates are re-ranked using the scores from the alignment head. This pipeline would enable the alignment head, which is slower but more effective, to contribute to the final ranking while keeping the whole system highly scalable.

## 5 CONCLUSIONS

In this paper, we presented an efficient and effective architecture for visual-textual cross-modal retrieval. Specifically, we proposed to learn an alignment score by independently forwarding the visual and the textual pipelines using a state-of-the-art VL Transformer as a backbone. Then, we used the scores produced by the alignment head to learn a visual-textual common space, which can produce easily indexable fixed-length features. Specifically, we approached the problem using a learn-to-rank distillation objective, which empirically demonstrated its effectiveness over the standard hinge-based triplet ranking loss to optimize the common space. The experiments conducted on MS-COCO confirmed the validity of our approach. The results demonstrated that this method helps fill the gap between effectiveness and efficiency, enabling this system to be deployed in large-scale cross-modal retrieval scenarios.

## ACKNOWLEDGMENTS

This work has been partially supported by AI4CHSites CNR4C program (CUP B15J19001040004), by AI4Media under GA 951911, by the "Artificial Intelligence for Cultural Heritage (AI4CH)" project, co-funded by the Italian Ministry of Foreign Affairs and International Cooperation, and by the PRIN project "CREATIVE: CRossmodal understanding and gEnerATIon of Visual and tExtual content" (CUP B87G22000460001), co-funded by the Italian Ministry of University and Research.

## REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.
[2] Rohan Anil, Gabriel Pereyra, Alexandre Passos, Robert Ormandi, George E Dahl, and Geoffrey E Hinton. 2018. Large scale distributed neural network training through online distillation. In *ICLR*.
[3] Pratyay Banerjee, Tejas Gokhale, Yezhou Yang, and Chitta Baral. 2021. Weakly Supervised Relative Spatial Reasoning for Visual Question Answering. In *ICCV*.
[4] Manuele Barraco, Matteo Stefanini, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi, and Rita Cucchiara. 2022. CaMEL: Mean Teacher Learning for Image Captioning. In *ICPR*.
[5] Sebastian Bruch. 2021. An alternative cross entropy loss for learning-to-rank. In *Web Conference*.
[6] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: from pairwise approach to listwise approach. In *ICML*.
[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *ICCV*.
[8] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: UNiversal Image-TExt Representation Learning. In *ECCV*.
[9] Marcella Cornia, Lorenzo Baraldi, Hamed R Tavakoli, and Rita Cucchiara. 2020. A unified cycle-consistent neural model for text and image retrieval. *Multimedia Tools and Applications* 79, 35 (2020), 25697–25721.
[10] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *CVPR*.
[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*.
[12] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. In *BMVC*.
[13] Yan Huang, Qi Wu, Wei Wang, and Liang Wang. 2018. Image and sentence matching via semantic concepts and order learning. *IEEE Trans. PAMI* 42, 3 (2018), 636–650.
[14] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*.
[15] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.
[16] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. In *NeurIPS Workshops*.
[17] Kuang-Huei Lee, Hamid Palangi, Xi Chen, Houdong Hu, and Jianfeng Gao. 2019. Learning visual relation priors for image-text matching and image captioning

with neural scene graph generators. *arXiv preprint arXiv:1909.09953* (2019).
[18] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *ACL*.
[19] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*.
[20] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. 2019. Visual semantic reasoning for image-text matching. In *ICCV*.
[21] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*.
[22] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692* (2019).
[23] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *NeurIPS*.
[24] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning. In *CVPR*.
[25] Nicola Messina, Giuseppe Amato, Andrea Esuli, Fabrizio Falchi, Claudio Gennaro, and Stéphane Marchand-Maillet. 2021. Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders. *ACM TOMM* 17, 4 (2021).
[26] Nicola Messina, Giuseppe Amato, Fabrizio Falchi, Claudio Gennaro, and Stéphane Marchand-Maillet. 2021. Towards efficient cross-modal visual textual retrieval using transformer-encoder deep features. In *CBMI*.
[27] Nicola Messina, Fabrizio Falchi, Andrea Esuli, and Giuseppe Amato. 2021. Transformer reasoning network for image-text matching and retrieval. In *ICPR*.
[28] Przemysław Pobrotyn, Tomasz Bartczak, Mikołaj Synowiec, Radosław Białobrzeski, and Jarosław Bojar. 2020. Context-aware learning to rank with self-attention. *arXiv preprint arXiv:2005.10084* (2020).
[29] Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. 2020. ImageBERT: Cross-modal Pre-training with Large-scale Weak-supervised Image-Text Data. *arXiv preprint arXiv:2001.07966* (2020).
[30] Leigang Qu, Meng Liu, Da Cao, Liqiang Nie, and Qi Tian. 2020. Context-Aware Multi-View Summarization Network for Image-Text Matching. In *ACM Multimedia*.
[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.
[32] Nikolaos Sarafianos, Xiang Xu, and Ioannis A Kakadiaris. 2019. Adversarial representation learning for text-to-image matching. In *ICCV*.
[33] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. 2022. From Show to Tell: A Survey on Deep Learning-based Image Captioning. *IEEE Trans. PAMI* (2022).
[34] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2021. A novel attention-based aggregation function to combine vision and language. In *ICPR*.
[35] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In *ICLR*.
[36] Keyu Wen, Xiaodong Gu, and Qingrong Cheng. 2020. Learning Dual Semantic Relations with Graph Attention for Image-Text Matching. *IEEE Transactions on Circuits and Systems for Video Technology* (2020).
[37] Yiling Wu, Shuhui Wang, Guoli Song, and Qingming Huang. 2019. Learning fragment self-attention embeddings for image-text matching. In *ACM Multimedia*.
[38] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. 2020. Self-Training With Noisy Student Improves ImageNet Classification. In *CVPR*.
[39] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. 2019. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *ICCV*.
[40] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. VinVL: Revisiting Visual Representations in Vision-Language Models. In *CVPR*.
[41] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J Corso, and Jianfeng Gao. 2020. Unified Vision-Language Pre-Training for Image Captioning and VQA. In *AAAI*.
[42] Yuanen Zhou, Meng Wang, Daqing Liu, Zhenzhen Hu, and Hanwang Zhang. 2020. More grounded image captioning by distilling image-text matching model. In *CVPR*.