Multimodal Learned Sparse Retrieval with Probabilistic Expansion Control

Thong Nguyen[1][0000-0003-0607-0723]*, Mariya Hendriksen[2][0000-0003-0314-2955]*,
Andrew Yates[1][0000-0002-5970-880X], and Maarten de Rijke[1][0000-0002-1086-0202]

[1] University of Amsterdam, The Netherlands
[2] AIRLab, University of Amsterdam, The Netherlands
`t.nguyen2@uva.nl`, `m.hendriksen@uva.nl`,
`a.c.yates@uva.nl`, `m.derijke@uva.nl`

Abstract. Learned sparse retrieval (LSR) is a family of neural methods that encode queries and documents into sparse lexical vectors that can be indexed and retrieved efficiently with an inverted index. We explore the application of LSR to the multi-modal domain, with a focus on text-image retrieval. While LSR has seen success in text retrieval, its application in multimodal retrieval remains underexplored. Current approaches like LexLIP and STAIR require complex multi-step training on massive datasets. Our proposed approach efficiently transforms dense vectors from a frozen dense model into sparse lexical vectors. We address issues of high dimension co-activation and semantic deviation through a new training algorithm, using Bernoulli random variables to control query expansion. Experiments with two dense models (BLIP, ALBEF) and two datasets (MSCOCO, Flickr30k) show that our proposed algorithm effectively reduces co-activation and semantic deviation. Our best-performing sparsified model outperforms state-of-the-art text-image LSR models with a shorter training time and lower GPU memory requirements. Our approach offers an effective solution for training LSR retrieval models in multimodal settings. Our code and model checkpoints are available at `github.com/thongnt99/lsr-multimodal`

1 Introduction

Learned sparse retrieval (LSR) [6, 7, 40] typically employs transformer-based encoders to encode queries and documents into sparse lexical vectors (i.e., bags of weighted terms) that are compatible with traditional inverted index. LSR has several nice properties. It provides an approach for effective and efficient neural retrieval, like dense retrieval, but with different advantages and trade-offs. For example, sparse representations have the potential to be interpretable because they are aligned with a vocabulary, and they leverage inverted index software rather than approximate nearest neighbor search [40]. Empirically, LSR also shows advantages over single-vector dense models on retrieval generalization benchmarks [6, 16].

While LSR and dense retrieval are common in text retrieval, dense retrieval has taken the lead in multi-modal search. This is evident in state-of-the-art text-image pretraining methods like BLIP [22] and ALBEF [23], which rely on dense architectures.

[*] These authors contributed equally.

The preference for dense models arises because images, unlike text, consist of continuous pixel values, presenting a challenge when they are mapped to discrete lexical terms. For multi-modal LSR, LexLIP [51] and STAIR [2] are the only two recent methods that exhibit competitive results on standard benchmarks. However, both require complex multi-step training on extensive text-image pairs: LexLIP with up to 14.3 million pairs and STAIR with a massive 1 billion pairs, encompassing public and private data.

We approach the multi-modal LSR (MMLSR) problem by using a pre-trained dense model and training a small sparse projection head on top of dense vectors, using image-text dense scores as a supervision signal. Naively learning the projection layer leads to issues of (i) high dimension co-activation and (ii) semantic deviation. Issue (i) happens when text and image sparse vectors excessively activate the same output dimensions, forming a sub-dense space inside the vocabulary space. Issue (ii) means that produced output terms do not reflect the content of captions/images, making them not human-interpretable. To counter (i) and (ii), we propose a single-step training method with probabilistic term expansion control. By disabling term expansions, we force the projection to produce meaningful terms first, then gradually allow more term expansions to improve the effectiveness while also randomly reminding the model not to fully rely on expansion terms. This process is handled using Bernoulli random variables with a parameter scheduler to model the expansion likelihood at both caption and word levels.

Opting for dense to sparse projection, instead of training an MMLSR model from scratch, provides several advantages. First, it is aligned with the broader community effort to reduce the carbon footprint of training deep learning models [30]. By keeping the dense encoders frozen and learning a lightweight projection layer, we can avoid the double GPU training/inference cost of two models (dense & sparse) while having more flexibility. Our approach enables the pre-computation of dense text and image vectors, allowing easy integration or removal of the projection layer based on available (dense or sparse) software and infrastructure. Moreover, this transformation may shed light on the interpretability of dense vectors, possibly contributing to a deeper understanding of the fundamental distinctions between these two multi-modal retrieval paradigms.

To understand the effectiveness and efficiency of the proposed training method, we conduct extensive experiments on two dense multi-modal models (BLIP, ALBEF) and two scene-centric [14] datasets (MSCOCO [27], Flickr30k [48]). We analyze the problems of dimension co-activation and semantic deviation under different settings.

Our contributions. The main contributions of our paper are: (i) We propose a line of research for efficiently converting a multi-modal dense retrieval model to a multi-modal LSR model. (ii) We train a lightweight projection head to convert dense to sparse vectors and show that our sparsified models are faithful to dense models while outperforming previous multi-modal LSR models. The training is efficient and does not require ground-truth labels. (iii) We identify the issues of high dimension co-activation and semantic deviation and propose a training method to address them.

2 Related Work

Learned sparse retrieval (LSR). Learned sparse retrieval is a family of neural retrieval methods that encode queries and documents into sparse lexical vectors that can be indexed and searched efficiently with an inverted index. There are many LSR approaches in the literature on text retrieval [7, 39, 49]; they are mainly built up from two types of encoder: MLP and MLM [40]. The MLP encoder uses a linear feedforward layer placed on top of the transformers's last contextualized embeddings to predict the importance of input terms (similar to term-frequency in traditional lexical retrieval). The MLP encoder has no term expansion capability. On the other hand, the MLM encoder utilizes the logits of the masked language model (MLM) for weighting terms and selecting expansion terms. Splade [6, 7] is a recent state-of-the-art text-oriented LSR approach that employs the MLM encoder in both query and document side, while other methods [3, 24, 33] use MLP encoders on both sides or only on the query side. Although it seems to be more beneficial to have expansion on both queries and documents, a recent study [40] found that query and document expansion have a cancellation effect on text retrieval (i.e., having expansion on the document side reduces the usefulness of query expansion) and one could obtain near state-of-the-art results without query expansion.

Unlike prior work focused on converting sparse to dense representations for hybrid ad-hoc text retrieval [25, 26], our work explores the reverse task of dense to sparse conversion in the multi-modal domain. This direction presents new challenges due to dimension co-activation and semantic deviation issues. Ram et al. [42] interpreted text dense retrieval by zero-shot projection from dense to vocabulary space using a frozen MLM head. Nguyen et al. [38] propose a simple sparse vision-language (VL) bi-encoder without query expansion and evaluate the performance on the image suggestion task. We aim for an efficient, effective, and semantically faithful drop-in sparse replacement of multi-modal dense retrieval, necessitating training of the projection layer.

Cross-modal retrieval. Cross-modal retrieval (CMR) methods construct a multimodal representation space, where the similarity of concepts from different modalities can be measured using a distance metric such as a cosine or Euclidean distance. Some of the earliest approaches in CMR utilized canonical correlation analysis [11, 18]. They were followed by a dual encoder architecture equipped with a recurrent and a convolutional component, the most prominent approaches in that area featured a hinge loss [8, 46]. Later approaches further improved the effectiveness using hard-negative mining [5].

Later, the integration of attention mechanisms improved performance. This family of attention mechanisms includes dual attention [37], stacked cross-attention [20], bidirectional focal attention [28]. Another line of work proposes to use transformer encoders [44] for this task [36], and adapts the BERT model [4] as a backbone [9, 52].

A related line of work focuses on improving the performance on CMR via modality-specific graphs [45], or image and text generation modules [12]. There is also more domain-specific work that focuses on CMR in fashion [10, 19], e-commerce [13], cultural heritage [43], and cooking [45].

3 Background

Task definition. We use the same notation as in previous work [1, 50]. We work with a cross-modal dataset $\mathcal{D}$ that includes N image-caption tuples: $\mathcal{D} = \left\{ \left(\mathbf{x}^i_{\mathcal{I}}, \{\mathbf{x}^i_{\mathcal{C}_j}\}^k_{i=1} \right) \right\}^N_{i=1}$. Each tuple comprises an image $\mathbf{x}_{\mathcal{I}}$ and k associated captions $\{\mathbf{x}_{\mathcal{C}_j}\}^k_{j=1}$.

The *cross-modal retrieval* (CMR) task is defined analogously to the standard information retrieval task: given a query and a set of candidates, we rank all candidates w.r.t.

their relevance to the query. The query can be either a caption or an image. Similarly, the set of candidate items can contain either images or captions. CMR is performed across modalities, therefore, if the query is a caption then the set of candidates are images, and vice versa. Hence, the task comprises two subtasks: (i) *caption-to-image retrieval*: retrieving images relevant to a caption query, and (ii) *image-to-caption retrieval*: retrieving relevant captions that describe an image query. We focus on *caption-to-image retrieval* only as it is more challenging, as reported by previous research [22, 23, 51].

Sparsification-induced phenomena. In this work, we investigate two phenomena arising during the sparsification process: dimension co-activation and semantic deviation.

Definition 1 (Dimension co-activation). We define *dimension co-activation* as sparse image and caption representations activating the same output dimensions, creating a sub-dense space within the vocabulary. While co-activation is essential for matching captions with images and can be measured by FLOPs, *high co-activation* results in unnecessarily long posting lists, harming the efficiency of LSR. Establishing a clear threshold for *high co-activation* is challenging, but we observe that beyond a certain point, increased FLOPs yield minimal improvements in effectiveness. To quantify this effect, we use effectiveness metrics (e.g., R@k) in combination with the FLOPs metric:

$$\mathrm{FLOPs} = \tfrac{1}{|\mathcal{C}||\mathcal{I}|} \textstyle\sum_{\mathbf{x}_{\mathcal{C}} \in \mathcal{C}} \sum_{\mathbf{x}_{\mathcal{I}} \in \mathcal{I}} \mathbf{s}_{\mathcal{C}}^{0} \cdot \mathbf{s}_{\mathcal{I}}^{0} \tag{1}$$

where $\mathcal{C}$ and $\mathcal{I}$ are caption and image collections, $\mathbf{s}_{\mathcal{C}}$, $\mathbf{s}_{\mathcal{I}}$ are sparse vectors of a caption $\mathbf{x}_{\mathcal{C}}$ and an image $\mathbf{x}_{\mathcal{I}}$.

Definition 2 (Semantic deviation). We define *semantic deviation* as the disparity between the semantic information in the visual or textual query and that in the sparse output terms. High co-activation suggests (but does not guarantee) semantic deviation.

Measuring semantic deviation directly is challenging, so we use two rough proxies, *Exact@k* and *Semantic@k*, defined as follows:

$$Exact@k = \frac{1}{k} |\{t \mid t \in \mathbf{x}_{\mathcal{C}}, t \in top_k(\mathbf{s}_{\mathcal{C}})\}| \tag{2}$$

$$Semantic@k = \frac{1}{k} \sum_{\mathbf{x}_t^i \in top_k(\mathbf{s}_{\mathcal{C}})} \max_{\mathbf{x}_t^j \in \mathbf{x}_{\mathcal{C}}} \frac{f_{enc}(\mathbf{x}_t^i) \cdot f_{enc}(\mathbf{x}_t^j)}{\|f_{enc}(\mathbf{x}_t^i)\| \|f_{enc}(\mathbf{x}_t^j)\|}. \tag{3}$$

Exact@k measures the ratio of overlapping terms between the input caption and the top-k highest weighted output terms, providing a partial picture of semantic deviation without considering synonyms. $Semantic@k$ complements $Exact@k$ by calculating the averaged cosine similarity between static embeddings obtained using model $f_{enc}(\cdot)$ of top-k output terms and input caption terms. Higher values for both metrics suggest less semantic deviation, implying better alignment of output terms with input captions.

4 Methodology

4.1 Model architecture

The architecture of our Dense2Sparse model is visualized in Figure 1. Dense2Sparse takes an image and a caption as input, projecting them into a $|V|$-dimensional space,

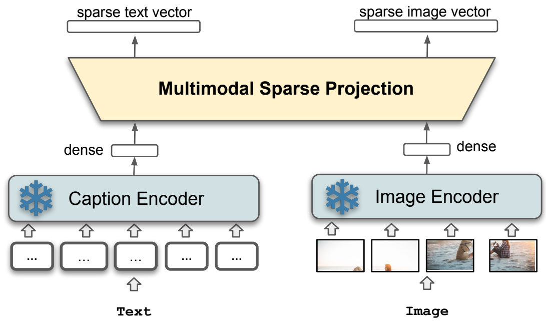


Fig. 1: The architecture of Dense2Sparse (D2S). The caption and image encoders are frozen, and the sparse projection is trained to project dense vectors to sparse vectors.

where each dimension represents the weight of a corresponding vocabulary entry. The key components include two dense encoders, an image encoder $f_\theta^{\mathcal{I}}(\cdot)$ and a caption encoder $f_\phi^{\mathcal{C}}(\cdot)$, as well as a multimodal sparse projection head $g_\psi(\cdot)$.

Dense image and text encoders. The *dense image encoder* $f_\theta^{\mathcal{I}} : \mathcal{X} \rightarrow \mathcal{Z}$ takes an input image $\mathbf{x}_{\mathcal{I}}$ and maps it into a latent space $\mathcal{Z} = \mathcal{R}^d$: $\mathbf{z}_{\mathcal{I}} = f_\theta^{\mathcal{I}}(\mathbf{x}_{\mathcal{I}})$, where $\mathbf{z}_{\mathcal{I}} \in \mathcal{R}^d$. Similarly, the *dense text encoder* $f_\phi^{\mathcal{C}} : \mathcal{X} \rightarrow \mathcal{Z}$ takes an input text (caption) $\mathbf{x}_{\mathcal{C}}$, and maps it into a latent space $\mathcal{Z} = \mathcal{R}^d$: $\mathbf{z}_{\mathcal{C}} = f_\phi^{\mathcal{C}}(\mathbf{x}_{\mathcal{C}})$, where $\mathbf{z}_{\mathcal{C}} \in \mathcal{R}^d$. We obtain dense representations using BLIP and ALBEF as a backbone. Both encoders are frozen.

Multimodal sparse projection head. The *multimodal sparse projection head* $g_\psi : \mathcal{Z} \rightarrow \mathcal{S}$ maps dense latent image and text representations into the sparse image and text vector space $\mathcal{S} = \mathcal{R}_{>0}^{|V|}$:

$$\mathbf{s}_{\mathcal{C}} = g_\psi(\mathbf{z}_{\mathcal{C}}) \quad \text{and} \quad \mathbf{s}_{\mathcal{I}} = g_\psi(\mathbf{z}_{\mathcal{I}}). \tag{4}$$

The multimodal sparse projection head comprises four steps. First, we project the d-dimensional dense vector $\mathbf{z}$ to an ω-dimensional dense vector: $\mathbf{z}_1 = \mathbf{W}_1\mathbf{z}$, where $\mathbf{W}_1 \in \mathcal{R}^{\omega \times d}$, $\mathbf{z} \in \mathcal{R}^d$, and $\mathbf{z}_1 \in \mathcal{R}^\omega$. Second, we apply layer normalization:

$$\mathbf{z}_2 = \frac{\mathbf{z}_1 - \mathbb{E}[\mathbf{z}_1]}{\sqrt{Var[\mathbf{z}_1] + \epsilon}} \cdot \gamma + \beta, \tag{5}$$

where $\mathbb{E}[\mathbf{z}_1]$ and $Var[\mathbf{z}_1]$ are the expectation and variance of $\mathbf{z}_1$, γ and β are learnable affine transformation parameters, and $\mathbf{z}_2 \in \mathcal{R}^\omega$. Third, we project z_2 to the vocabulary space $\mathcal{S} = \mathcal{R}_{>0}^{|V|}$: $\mathbf{s} = \mathbf{W}_2\mathbf{z}_2$, where $\mathbf{W}_2 \in \mathcal{R}^{|V| \times \omega}$, $\mathbf{z}_2 \in \mathcal{R}^\omega$, and $\mathbf{s} \in \mathcal{R}^{|V|}$. $\mathbf{W}_2$ is initialized with vocabulary embeddings similar to the transformer-masked language model. Fourth, we remove negative weights and apply a logarithmic transformation to the positive weights: $\mathbf{s} = \log_e(1 + \max(0, \mathbf{s}))$, where $\mathbf{s} \in \mathcal{R}_{>0}^{|V|}$. The resulting $|V|$-dimensional sparse vector is aligned with the vocabulary, and each dimension represents the weight of the corresponding vocabulary entry. This projection head is similar to the MLM head employed in previous work [6, 33].

Probabilistic expansion control. Without any intervention, training the projection module with a standard contrastive loss could lead to high-dimension co-activation and semantic deviation as defined previously. This phenomenon affects the efficiency of an

Algorithm 1 Multimodal LSR training with probabilistic expansion control

```
Input: image-caption pair ($\mathbf{x}_\mathcal{I}$, $\mathbf{x}_\mathcal{C}$), caption encoder $f_\phi^\mathcal{C}$, image encoder $f_\theta^\mathcal{I}$, sparse projection head $g_\psi$, loss function $\mathcal{L}$, and expansion rate function $f_{\text{incr}}$.

$p_i^v \leftarrow 1 - df_i^v$
$p_c \leftarrow 0$

for epoch do
    for batch do
        $\mathbf{z}_\mathcal{C} \leftarrow f_\phi^\mathcal{C}(\mathbf{x}_\mathcal{C}), \;\; \mathbf{z}_\mathcal{I} \leftarrow f_\theta^\mathcal{I}(\mathbf{x}_\mathcal{I})$
        $\mathbf{s}_\mathcal{C} \leftarrow g_\psi(\mathbf{z}_\mathcal{C}), \;\; \mathbf{s}_\mathcal{I} \leftarrow g_\psi(\mathbf{z}_\mathcal{I})$
        $\mathcal{E}_\mathcal{C} \sim$ Ber($p_c$), $\;\; \mathcal{E}_i^v \sim$ Ber($p_i^v$)
        $\bar{\mathbf{s}}_\mathcal{C} \leftarrow \text{EXPAND}(\mathbf{x}_\mathcal{C}, \mathbf{s}_\mathcal{C}, \mathcal{E}_\mathcal{C}, \mathcal{E}_i^v)$
        $\mathcal{L} \leftarrow \mathcal{L}(\bar{\mathbf{s}}_\mathcal{C}, \mathbf{s}_\mathcal{I}, \mathbf{z}_\mathcal{I}, \mathbf{z}_\mathcal{C})$
    end for
    $p_c \leftarrow f_{incr}(p_c), \;\; p_i^v \leftarrow f_{incr}(p_i^v)$
end for

function EXPAND($\mathbf{x}_\mathcal{C}$, $\mathbf{s}_\mathcal{C}$, $\mathcal{E}_\mathcal{C}$, $\mathcal{E}_i^v$)
    for $0 \leq i <$ batch_size do
        for $0 \leq k < |V|$ do
            if $v_k \notin \mathbf{x}_\mathcal{C}$ then
                $\mathbf{s}_{\mathcal{C}i,k} \leftarrow \mathbf{s}_{\mathcal{C}i,k} \cdot \mathcal{E}_\mathcal{C} \cdot e_k^v$
            else
                $\mathbf{s}_{\mathcal{C}i,k} \leftarrow \mathbf{s}_{\mathcal{C}i,k} \cdot \mathcal{E}_k^v$
            end if
        end for
    end for
    return $\mathbf{s}_\mathcal{C}$
end function
```

inverted index and the interpretability of the outputs. To mitigate this problem, we propose a single-step training algorithm with probabilistic lexical expansion control. It is described in Algorithm 1.

We define a Bernoulli random variable $\mathcal{E} \sim Ber(p), \; p \in [0, 1]$ and use it to control textual query expansion. We consider a caption-level and a word-level expansion. The *caption-level expansion* is controlled by the random variable $\mathcal{E}_\mathcal{C} \sim Ber(p_\mathcal{C})$. If $\mathcal{E}_\mathcal{C} = 1$ the expansion is allowed, while $\mathcal{E}_\mathcal{C} = 0$ means the expansion is not allowed. Analogously, the *word-level expansion*, or the expansion to the i-th word in the vocabulary, is regulated by the random variable $\mathcal{E}_i^v \sim Ber(p_i^v)$.

The parameters $p_\mathcal{C}$ and p_i^v define the likelihood of caption-level and word-level expansion within a given training epoch. During training, we initially set the caption-level expansion probability, $p_\mathcal{C}$, to zero. This initial value prevents the expansion of textual queries, forcing the model to project images onto relevant tokens belonging to the captions they were paired with. This approach facilitates the meaningful projection of dense vectors onto relevant words in the vocabulary. However, it adversely impacts retrieval effectiveness, as the model cannot expand queries. As a consequence, the model's ability to handle semantic matching is limited. To gradually relax this constraint, we use a scheduler that incrementally increases the value of p after each epoch until it reaches a maximum value of one in the final epoch. In each epoch, we sample the values of $\mathcal{E}$ per batch and enforce expansion terms to be zero when $\mathcal{E}_\mathcal{C}$ equals zero. Similarly, for word-level expansion, we initialize the expansion probability of the i-th word p_i^v to $1 - df_i^v$ where df_i^v is the normalized document frequency of vocabulary element v_i in the caption collection $\mathcal{C}$. This setting discourages the expansion of more frequent terms because they are less meaningful and can hinder the efficiency of query processing algorithms. We relax each p_i^v after every epoch, ensuring that it reaches a maximum value of one at the conclusion of the training process. The expansion rate increase after each epoch is defined as follows:

$$f_{\text{incr}}(p) = \begin{cases} p + \frac{1}{\textit{\# epochs}}, & \text{for caption-level expansion} \\ p + \frac{df_i^v}{\textit{\# epochs}}, & \text{for word-level expansion.} \end{cases} \tag{6}$$

4.2 Training loss

We train our Dense2Sparse using a loss that represents a weighted sum of a bidirectional loss and a sparse regularization parameter. The bidirectional loss is based on the following one-directional loss:

$$\ell^{(\mathcal{A}\to\mathcal{B})} = -\left(\frac{\exp(\mathbf{z}_{\mathcal{A}}^{\intercal}\mathbf{z}_{\mathcal{B}}/\tau)}{\sum_{\mathcal{I}*}\exp(\mathbf{z}_{\mathcal{A}}^{\intercal}\mathbf{z}_{\mathcal{I}*}/\tau)}\right)\log_2\left(\text{SoftMax}[\mathbf{s}_{\mathcal{A}}^{\intercal}\mathbf{s}_{\mathcal{B}}]\right),$$

where $\mathbf{s}_{\mathcal{A}} \in \mathcal{R}_{>0}^{|V|}$ and $\mathbf{s}_{\mathcal{B}} \in \mathcal{R}_{>0}^{|V|}$ are sparse vectors, $\mathbf{z}_{\mathcal{A}} \in \mathcal{R}^d$ and $\mathbf{z}_{\mathcal{B}} \in \mathcal{R}^d$ are dense vectors, and $\tau \in \mathcal{R}_{>0}$ is a temperature parameter.

The resulting loss is formalized to capture both bidirectional losses and sparse regularization. The overall loss $\mathcal{L}$ is defined as:

$$\mathcal{L} = (1-\lambda)\underbrace{[\ell^{(\mathcal{I}\to\mathcal{C})} + \ell^{(\mathcal{C}\to\mathcal{I})}]}_{\textit{bidirectional loss}} + \lambda\underbrace{\eta[L_1(\mathbf{s}_{\mathcal{I}}) + L_1(\mathbf{s}_{\mathcal{C}})]}_{\textit{sparse regularization parameter}}, \tag{7}$$

where $\ell^{(\mathcal{I}\to\mathcal{C})}$ is an image-to-caption loss, $\ell^{(\mathcal{C}\to\mathcal{I})}$ is a caption-to-image loss; $\lambda = [0,1]$ is a scalar weight, $\eta = [0,1]$ is a sparsity regularization parameter, and $L_1(\mathbf{x}) = \|\mathbf{x}\|_1$ is L_1 regularization. It is worth noting that the loss utilizes dense scores for supervision, a strategy found to be more effective than using ground truth labels.

5 Experiments and Results

5.1 Experimental setup

Datasets. We trained and evaluated our models on two widely used datasets for text-image retrieval: MSCOCO [27] and Flickr30k [41]. Each image in the two datasets is paired with five short captions (with some exceptions). We re-used the splits from [17] for training, evaluating, and testing. The splits on MSCOCO have 113.2k pairs for training, and 5k pairs for each validation/test set. Flickr30 is smaller with 29.8k/1k/1k for train, validation, test splits respectively. The best model is selected based on the validation set and evaluated on the test set.

Evaluation metrics. To evaluate model performance and effectiveness, we report R@k where $k = \{1, 5\}$, and MRR@10 using the *ir_measures* [32] library.

Implementation and training details. The caption and image dense vectors of BLIP [22] and ALBEF [23] models are pre-computed with checkpoints from the larvis library [21]. We train our models to convert from dense vectors to sparse vectors on a single A100 GPU with a batch size of 512 for 200 epochs. The training takes around 2 hours and only uses up to around 10 GB of GPU memory. We set the temperature τ to 0.001 and experiment with sparse regularization weights $\eta \in [1e-5, 1e-2]$.

Table 1: The effectiveness of sparsified models (D2S) and baselines. ($^\dagger p < 0.05$ *with paired two-tailed t-test comparing D2S to the dense model with Bonferroni correction*)

Model	**MSCOCO (5k)**				**Flickr30k (1k)**			
	R@1↑	R@5↑	MRR@10↑	FLOPs↓	R@1↑	R@5↑	MRR@10↑	FLOPs↓
T2I Dense Retrieval								
COOKIE [47]	46.6	75.2	-	-	68.3	91.1	-	-
COTS (5.3M) [29]	50.5	77.6	-	-	75.2	93.6	-	-
ALBEF [23]	53.1	79.3	64.3	-	79.1	94.9	86.6	-
BLIP [22]	**57.3**	**81.8**	**67.8**	-	**83.2**	**96.7**	**89.3**	-
T2I Sparse Retrieval								
VisualSparta	45.1	73.0	-	-	57.1	82.6	-	-
STAIR (zero-shot)	41.1	56.4	-	-	66.6	88.7	-	-
LexLIP (4.3M)	51.9	78.3	-	-	76.7	93.7	-	-
LexLIP (14.3M)	53.2	79.1	-	-	78.4	94.6	-	-
D2S (ALBEF, $\eta = 1e-3$)	49.6†	77.7†	61.4†	18.7	74.2†	93.8†	82.6†	21.7
D2S (ALBEF, $\eta = 1e-5$)	50.7†	78.2†	62.4†	74.2	75.4†	94.3†	83.6†	64.3
D2S (BLIP, $\eta = 1e-3$)	51.8†	79.3†	63.4†	**11.5**	77.1†	94.6†	84.6†	**9.9**
D2S (BLIP, $\eta = 1e-5$)	**54.5**†	**80.6**†	**65.6**†	78.4	**79.8**†	**95.9**†	**86.7**†	39.5

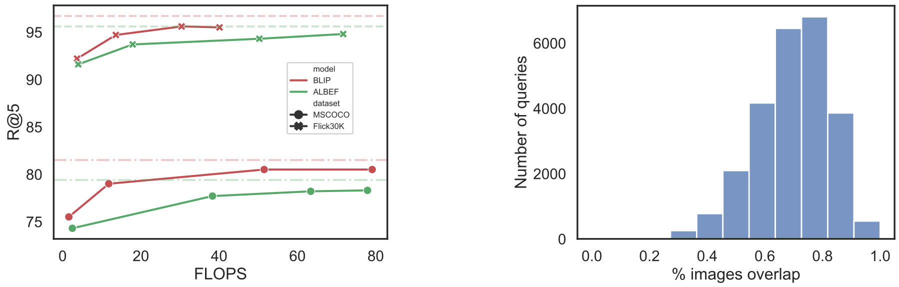
5.2 Results and discussion

RQ1: How effective and efficient is the proposed method for converting dense to sparse? We trained various Dense2Sparse models (D2S) using our proposed training method with different sparse regularization weights ranging from $1e-5$ to $1e-2$. Figure 2a illustrates the effectiveness and efficiency of these variations, with detailed results presented in Table 1. Firstly, we observe that increasing the sparse regularization weight enhances model efficiency (reduced FLOPs) but reduces its effectiveness (lower Recall and MRR). On the MSCOCO dataset, our most efficient sparse BLIP model ($\eta = 1e-2$) achieves a R@1 of 47.2 and MRR@10 of 58.5 with the lowest FLOPs value of 1.6. Relaxing the regularization weight to $1e-3$ results in an approximately 10% increase in R@1 to 51.8 and a similar rise in MRR@10 to 63.4, albeit at the expense of around 7 times higher FLOPs (less efficient).

Further relaxing the sparse regularization gradually brings the sparsified model's effectiveness closer to the original dense model, while reducing the efficiency. The most effective sparsified BLIP model with $\eta = 1e-5$ performs competitively with the original dense version (54.5 vs. 57.3) and outperforms other dense baselines.

Additionally, we observe a diminishing gap between dense and sparsified models as we assess recalls at higher cutoff positions, such as $R@5$ and $R@10$. Similar trends are observed across different datasets, including Flickr30k and MSCOCO, as well as among different dense models, including BLIP and ALBEF. This indicates the broad applicability of our proposed approach to diverse datasets and models.

RQ2: How does our sparsified model compare to state-of-the-art multi-modal LSR models? In this research question, we compare our sparsified models with existing LSR baselines, namely Visual Sparta, STAIR, and LexLIP. Currently, neither the code nor the checkpoints for these baselines are publicly available. Therefore, we rely on the numbers reported in their respective papers for comparison, excluding the FLOPs.



(a) Efficiency vs. effectiveness of sparsified models

(b) Fraction of overlapping images in the top-10 by sparsified and dense BLIP model.

Fig. 2: Sparisified models compared to original dense models.

STAIR and LexLIP are two of the most recent multimodal LSR approaches, both trained on large datasets, with STAIR utilizing 1 billion internal text-image pairs. In contrast, our proposed method leverages pretrained dense retrieval models to efficiently learn a lightweight sparse projection for converting dense vectors to sparse vectors.

The effectiveness of our methods and the baselines on MSCOCO and Flickr30k is presented in Table 1. Notably, our efficient model, D2S(BLIP, $\eta = 1e - 3$), performs competitively with LexLIP trained on 4.3 million text-image pairs at R@1. Its R@5 is slightly better than LexLIP (4.3M) and comparable to the LexLIP model trained on 14.3 million pairs. With a lower sparse regularization, our D2S(BLIP, η=1e$-$5) model significantly outperforms all baselines on both MSCOCO and Flickr30k. On MSCOCO, its R@1 is 21%, 5%, and 2.8% higher than the R@1 of Visual Sparta, LexLIP (4.3M), and LexLIP (14.3M), respectively. All our models outperform Visual Sparta and STAIR, although this comparison with STAIR uses a zero-shot setting, because we lack access to their code and checkpoints for fine-tuning STAIR further with in-domain data.

We kept the dense encoders frozen, so the effectiveness of our sparsified models is inherently bounded by the dense results. Our sparsified ALBEF models, for example, exhibit slightly lower overall effectiveness since their corresponding dense performance is lower than that of BLIP's dense scores. Nonetheless, our sparsified ALBEF models are also comparable with LexLIP variants.

RQ3: Does the proposed training method help address the dimension co-activation and semantic deviation issues? As discussed in Section 3, high co-activation increases posting list length, impacting inverted index efficiency. We examine this impact by analyzing FLOPs alongside model effectiveness metrics. Table 1 presents results for models trained with our method and three baseline variants, with fixed expansion rates of 0 and 1 in the first two baselines. The third baseline ($exp = c$) explores the influence of word-level expansion control, excluding it from our training method.

At an expansion rate of zero, models project the caption's dense vector only onto terms from the caption, with all other projections forced to zero. The image projector must then learn to align the image vector with terms in the paired captions. Conversely, setting exp to 1 gives the model the freedom to project onto any output vectors, making it more inclined toward dimension co-activation.

Table 2: The dimension co-activation effect of Dense2Sparse (D2S) variations.

Model (D2S variations)	**MSCOCO (5k)**				**Flickr30k (1k)**			
	R@1↑	R@5↑	MRR@10↑	FLOPs↓	R@1↑	R@5↑	MRR@10↑	FLOPs↓
(BLIP, $\eta = 1e-3$, $exp = 0$)	45.5	73.0	57.3	2.8	68.9	89.5	77.8	3.0
(BLIP, $\eta = 1e-3$, $exp = 1$)	53.4	80.0	64.6	49.1	79.5	95.5	86.4	50.3
(BLIP, $\eta = 1e-3$, $exp = c$)	51.9	79.0	63.4	11.8	77.3	94.7	84.8	13.6
(BLIP, $\eta = 1e-3$, $exp = c + w$)	51.8	79.3	63.4	11.5	77.1	94.6	84.6	9.9
(BLIP, $\eta = 1e-5$, $exp = 0$)	47.2	74.4	58.8	3.2	72.3	91.8	80.7	3.5
(BLIP, $\eta = 1e-5$, $exp = 1$)	55.9	81.3	66.8	343	81.4	96.0	87.7	213
(BLIP, $\eta = 1e-5$, $exp = c$)	54.7	80.5	65.8	79.1	79.9	95.5	86.7	40.1
(BLIP, $\eta = 1e-5$, $exp = c + w$)	54.5	80.6	65.6	78.4	79.8	95.9	86.7	39.5
(ALBEF, $\eta = 1e-3$, $exp = 0$)	43.8	71.8	55.7	2.5	65.8	88.3	75.4	3.0
(ALBEF, $\eta = 1e-3$, $exp = 1$)	50.9	78.4	62.5	68.2	75.7	94.2	83.8	61.9
(ALBEF, $\eta = 1e-3$, $exp = c$)	49.7	77.7	61.5	38.3	74.6	93.7	82.8	17.9
(ALBEF, $\eta = 1e-3$, $exp = c + w$)	49.6	77.7	61.4	18.7	74.2	93.8	82.6	21.7
(ALBEF, $\eta = 1e-5$, $exp = 0$)	45.9	73.9	83.0	3.4	68.1	90.0	77.6	3.2
(ALBEF, $\eta = 1e-5$, $exp = 1$)	52.4	78.7	63.7	283	77.2	94.6	84.8	210
(ALBEF, $\eta = 1e-5$, $exp = c$)	51.2	78.3	62.8	77.9	76.4	94.8	84.0	71.7
(ALBEF, $\eta = 1e-5$, $exp = c + w$)	50.7	78.2	62.4	74.2	75.4	94.3	83.6	64.3

In Table 2, rows with ($exp = 0$) show models with no expansion, resulting in remarkably low FLOPs, with each query averaging 2 to 3 overlapping terms with each document. However, disabling expansion reduces the model's ability for semantic matching, leading to modest effectiveness (45–47R@1 on MSCOCO and 68–72R@1 on Flickr30k with varying sparsity). Enabling non-regulated expansion ($exp = 1$) significantly improves model effectiveness (50–55 R@1 on MSCOCO and 75–79R@1 on Flickr30k with various regularization weights). However, this improvement comes at the cost of substantially increased FLOP scores, sometimes by up to 100 times, making sparsified vectors very computationally expensive. Ultimately, the resulting models behave like dense models, which is an undesired effect.

Our training method, which incorporates expansion control at the caption and word levels, is designed to gradually transition from one extreme ($exp = 0$) to the other ($exp = 1$). During training, we allow a likelihood of expansion, which increases progressively to over time. However, we also introduce random elements, represented by a random variable, to remind the model to remain faithful to the original captions/images.

The results, displayed in rows labeled with $exp = c + w$, demonstrate that our approach strikes a better balance between efficiency and effectiveness. It achieves competitive levels of effectiveness compared to models with $exp = 1$ while requiring only half or a third of the computational operations (FLOPs). For example, on MSCOCO with the BLIP model, Dense2Sparse ($\eta = 1e - 3$) achieves a performance of 51.8 R@1 (compared to 53.4 when $exp = 1$) with just 11.8 FLOPs, making it four times more efficient than the $exp = 1$ baseline. With the same setting, our method achieves 14% higher R@1 and 11% higher MRR@10 than the baseline with no expansion ($exp = 0$). Compared to the baseline without word-level expansion control, no significant differences are observed in terms of efficiency and effectiveness. Thus, caption-level expansion control alone seems sufficient for achieving reasonable efficiency and effectiveness. Similar results are noted across various settings, datasets, and dense models.

Table 3: Semantic deviation on different Dense2Sparse (D2S) variations. ($^\dagger p < 0.01$ *with paired two-tailed t-test comparing exp=c to exp=1*)

Model (D2S variations)	**MSCOCO (5k)**		**Flickr30k (1k)**	
	Exact@20	Semantic@20	Exact@20	Semantic@20
(BLIP, $\eta = 1e-5$, $exp = c$)	20.0†	60.1†	18.3†	58.0†
(BLIP, $\eta = 1e-5$, $exp = 1$)	6.9	48.5	3.2	40.7
(BLIP, $\eta = 1e-3$, $exp = c$)	25.0†	63.2†	23.1†	60.6†
(BLIP, $\eta = 1e-3$, $exp = 1$)	2.5	42.0	2.2	41.1
(ALBEF, $\eta = 1e-5$, $exp = c$)	20.5†	61.0†	19.2†	59.8†
(ALBEF, $\eta = 1e-5$, $exp = 1$)	5.6	43.5	1.2	40.5
(ALBEF, $\eta = 1e-3$, $exp = c$)	15.1†	51.3†	19.6†	56.4†
(ALBEF, $\eta = 1e-3$, $exp = 1$)	1.6	40.6	1.3	41.5

Table 4: Examples of semantic deviation. We show the top-10 terms per model.

Caption, Image	**D2S** ($\eta = 1e-3$, **exp=c**)	**D2S** ($\eta = 1e-3$, **exp=1.0**)
A man with a red helmet on a small moped on a dirt road	dirt, mo, motor, motorcycle, bike, red, riding, features, soldier, ##oot	, accent " yourself natural may while officer english ac
	mountain mountains bike bee dirt mo red path ##oot person man riding bicycle	accent ship natural de crown yourself " ra now wild
A women smiling really big while holding a Wii remote.	lady woman smile women remote laughing wii smiling video controller	, kai called forces rush lee war oil like ##h
	smile after green woman smiling sweater remote lady wii her	tall kai forces oil rush met war college thus there
A couple of dogs sitting in the front seats of a car.	dogs dog car backseat seat couple vehicle sitting two puppy	, electric stood forest national master help arts fc -
	dog car dogs puppy out vehicle pup inside early open	stood forest national electric master twice grant men para yet

Sparse representations contain interpretable output dimensions aligned with a vocabulary. However, training a D2S model without our expansion regulation leads to semantic deviation, turning vocabulary terms into non-interpretable latent dimensions. We assess this effect using Exact@k and Semantic@k metrics (defined in Section 3), reporting results in Table 3 and providing qualitative examples in Table 4.

Uncontrolled models (with $exp = 1$) exhibit lower Exact@20 and Semantic@20 than our expansion-controlled models ($exp = c$). In the top 20 terms of uncontrolled models, only one or none are in the original captions, while controlled models generate

Table 5: Correlation between dense and different variations of Dense2Sparse (D2S).

Model (D2S variations)	**MSCOCO (5k)**			**Flickr30k (1k)**		
	ρ-R@1↑	ρ-R@5↑	ρ-MRR@10↑	ρ-R@1↑	ρ-R@5↑	ρ-MRR@10↑
(BLIP, $\eta = 1e - 2$)	61.0	65.7	72.3	54.7	55.0	63.9
(BLIP, $\eta = 1e - 3$)	74.0	76.9	83.8	66.2	65.5	73.6
(BLIP, $\eta = 1e - 4$)	79.7	82.1	88.2	71.6	72.8	79.3
(BLIP, $\eta = 1e - 5$)	81.2	83.8	89.2	74.3	74.0	81.1
(ALBEF, $\eta = 1e - 2$)	64.4	68.7	75.5	57.7	57.0	67.5
(ALBEF, $\eta = 1e - 3$)	73.1	76.7	83.5	68.8	69.0	77.2
(ALBEF, $\eta = 1e - 4$)	78.1	80.7	87.2	73.2	74.6	81.3
(ALBEF, $\eta = 1e - 5$)	78.2	81.3	87.3	74.2	72.5	82.0

3 to 5 caption terms. The low Semantic@20 of the uncontrolled models also suggests low relatedness of output terms to the caption terms. This implication could be further supported by the examples demonstrated in Table 4. Uncontrolled models generate random terms, while our method produces terms that more faithfully reflect captions and images. Most top-10 terms from our method are relevant to the input, including a mix of original terms and synonyms (e.g., “dog” vs. “puppy”, “car” vs. “vehicle”).

RQ4: Is the sparsified model faithful to the dense model? This research question aims to analyze the faithfulness of sparsified models to their original dense models. We report in Table 5 the Pearson correlation calculated for various effectiveness metrics of dense and sparsified queries. The results show that the correlation between sparsified and dense models is consistently positive and tends to increase as we relax the sparse regularization. Furthermore, as we consider higher cutoff values (R@1, R@5, MRR@10), the correlation tends to increase as the performance gap between the two systems narrows. Manually comparing the top-10 ranked images of the most differing queries, we find that while the two models rank top-10 images differently, there are a lot of common images (including the golden image) that look equally relevant to the query. Figure 2b shows that a high ratio (average: 70%) of the top-10 images appear in both dense and sparse ranking lists. This analysis shows that the sparsified model is reasonably faithful to the dense model, suggesting that the sparse output terms could potentially be used for studying the semantics of dense vectors.

5.3 Retrieval latency of dense and sparsified models

We discussed the average FLOPs of sparsified models for retrieval efficiency. We now present query throughput and retrieval latency results in Table 6. Using Faiss [15] and PISA [31, 35] on a single-threaded AMD Genoa 9654 CPU, the dense BLIP model with Faiss HNSW is exceptionally fast, outperforming D2S models with PISA. D2S models with query expansion (*exp=c*) are slower due to high FLOPs and possibly LSR known limitations [34]. Removing expansion terms (*exp=0*) improves latency (FLOPs similar to DistilSPLADE [6, 7]) but is still approximately 30× slower than dense retrieval. To balance efficiency and effectiveness of D2S, we propose using the inverted index with original query terms for retrieval, followed by re-scoring with expansion terms. With our simple iterative implementation, this approach proves effective, especially

Table 6: Retrieval latency (CPU - 1 thread) of D2S models on 123k MSCOCO images.

Model		**Throughput** (q/s)			**Latency** (ms)		
	FLOPS	@10	@100	@1000	@10	@100	@1000
Dense (BLIP, HNSW, Faiss)	-	13277	9739	7447	0.08	0.10	0.14
D2S (BLIP, $\eta = 1e-3$, exp=c, PISA)	11.5	6	5	5	156.60	183.42	193.46
D2S (BLIP, $\eta = 1e-3$, exp=0, PISA)	2.8	449	284	160	2.23	3.52	6.25
No Expansion >> Expansion	-	369	120	18	2.70	8.31	54.05
D2S (BLIP, $\eta = 1e-5$, exp=c, PISA)	78.4	<1	<1	<1	>300	>600	>700
D2S (BLIP, $\eta = 1e-5$, exp=0, PISA)	3.2	230	146	90	4.34	6.85	11.04
No Expansion >> Expansion	-	189	70	11	5.30	14.37	86.66
D2S (BLIP, HNSW, Faiss)	-	262	262	256	3.82	3.82	3.90

for retrieving fewer images per query. Surprisingly, indexing D2S models with Faiss HNSW competes well with PISA, particularly at higher cut-off values (100, 1000).

6 Conclusion

We have focused on the problem of efficiently transforming a pretrained dense retrieval model into a sparse model. We show that training a projection layer on top of dense vectors with the standard contrastive learning technique leads to the problems of dimension co-activation and semantic deviation. To mitigate these issues, we propose a training algorithm that uses a Bernoulli random variable to control the term expansion. Our experiments show that our Dense2Sparse sparsified model trained with the proposed algorithm suffers less from those issues. In addition, our sparsified models perform competitively to the state-of-the-art multi-modal LSR, while being faithful to the original dense models.

Acknowledgments. We thank our anonymous reviewers for their valuable feedback. This research was supported by Ahold Delhaize, the Hybrid Intelligence Center, a 10-year program funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, project LESSEN with project number NWA.1389.20.183 of the research program NWA ORC 2020/21, which is (partly) financed by the Dutch Research Council (NWO), project IDEAS with project number VI.Vidi.223.166 of the NWO Talent Programme, which is (partly) financed by the Dutch Research Council (NWO), and the FINDHR (Fairness and Intersectional Non-Discrimination in Human Recommendation) project that received funding from the European Union's Horizon Europe research and innovation program under grant agreement No 101070212. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

Bibliography

[1] Brown, A., Xie, W., Kalogeiton, V., Zisserman, A.: Smooth-AP: Smoothing the path towards large-scale image retrieval. In: Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IX, vol. 12354, pp. 677–694, Springer (2020)

[2] Chen, C., Zhang, B., Cao, L., Shen, J., Gunter, T., Jose, A.M., Toshev, A., Shlens, J., Pang, R., Yang, Y.: STAIR: Learning sparse text and image representation in grounded tokens. arXiv preprint arXiv:2301.13081 (2023)

[3] Dai, Z., Callan, J.: Context-aware sentence/passage term importance estimation for first stage retrieval. arXiv preprint arXiv:1910.10687 (2019)

[4] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

[5] Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: VSE++: Improving visual-semantic embeddings with hard negatives. arXiv preprint arXiv:1707.05612 (2017)

[6] Formal, T., Lassance, C., Piwowarski, B., Clinchant, S.: From distillation to hard negative sampling: Making sparse neural ir models more effective. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2353–2359, SIGIR '22, Association for Computing Machinery, New York, NY, USA (2022)

[7] Formal, T., Piwowarski, B., Clinchant, S.: SPLADE: Sparse lexical and expansion model for first stage ranking. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2288–2292 (2021)

[8] Frome, A., Corrado, G., Shlens, J., Bengio, S., Dean, J., Ranzato, M., Mikolov, T.: DeViSE: A deep visual-semantic embedding model. In: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, p. 2121–2129 (2013)

[9] Gao, D., Jin, L., Chen, B., Qiu, M., Li, P., Wei, Y., Hu, Y., Wang, H.: Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2251–2260 (2020)

[10] Goei, K., Hendriksen, M., de Rijke, M.: Tackling attribute fine-grainedness in cross-modal fashion search with multi-level features. In: SIGIR 2021 Workshop on eCommerce, ACM (July 2021)

[11] Gong, Y., Wang, L., Hodosh, M., Hockenmaier, J., Lazebnik, S.: Improving image-sentence embeddings using large weakly annotated photo collections. In: European Conference on Computer Vision, pp. 529–545, Springer (2014)

[12] Gu, J., Cai, J., Joty, S.R., Niu, L., Wang, G.: Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7181–7189 (2018)

[13] Hendriksen, M., Bleeker, M., Vakulenko, S., van Noord, N., Kuiper, E., de Rijke, M.: Extending CLIP for category-to-image retrieval in e-commerce. In: European Conference on Information Retrieval, pp. 289–303, Springer (2022)

[14] Hendriksen, M., Vakulenko, S., Kuiper, E., de Rijke, M.: Scene-centric vs. object-centric image-text cross-modal retrieval: A reproducibility study. In: European Conference on Information Retrieval, pp. 68–85, Springer (2023)

[15] Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with gpus. IEEE Transactions on Big Data **7**(3), 535–547 (2019)

[16] Kamalloo, E., Thakur, N., Lassance, C., Ma, X., Yang, J.H., Lin, J.: Resources for brewing BEIR: Reproducible reference models and an official leaderboard. arXiv preprint arXiv:2306.07471 (2023)

[17] Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3128–3137 (2015)

[18] Klein, B., Lev, G., Sadeh, G., Wolf, L.: Fisher vectors derived from hybrid gaussian-laplacian mixture models for image annotation. arXiv preprint arXiv:1411.7399 (2014)

[19] Laenen, K.: Cross-modal Representation Learning for Fashion Search and Recommendation. Ph.D. thesis, KU Leuven (2022)

[20] Lee, K.H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 201–216 (2018)

[21] Li, D., Li, J., Le, H., Wang, G., Savarese, S., Hoi, S.C.: Lavis: A library for language-vision intelligence. arXiv preprint arXiv:2209.09019 (2022)

[22] Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning, pp. 12888–12900, PMLR (2022)

[23] Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. Advances in Neural Information Processing Systems **34**, 9694–9705 (2021)

[24] Lin, J., Ma, X.: A few brief notes on deepimpact, coil, and a conceptual framework for information retrieval techniques. arXiv preprint arXiv:2106.14807 (2021)

[25] Lin, S.C., Lin, J.: Densifying sparse representations for passage retrieval by representational slicing. arXiv preprint arXiv:2112.04666 (2021)

[26] Lin, S.C., Lin, J.: A dense representation framework for lexical and semantic matching. ACM Transactions on Information Systems **41**(4), 1–29 (2023)

[27] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, pp. 740–755, Springer (2014)

[28] Liu, C., Mao, Z., Liu, A.A., Zhang, T., Wang, B., Zhang, Y.: Focus your attention: A bidirectional focal attention network for image-text matching. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 3–11 (2019)

[29] Lu, H., Fei, N., Huo, Y., Gao, Y., Lu, Z., Wen, J.: COTS: collaborative two-stream vision-language pre-training model for cross-modal retrieval. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, pp. 15671–15680, IEEE (2022)

[30] Luccioni, A.S., Hernandez-Garcia, A.: Counting carbon: A survey of factors influencing the emissions of machine learning. arXiv preprint arXiv:2302.08476 (2023)

[31] MacAvaney, S., Macdonald, C.: A python interface to pisa! In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (2022), https://doi.org/10.1145/3477495.3531656

[32] MacAvaney, S., Macdonald, C., Ounis, I.: Streamlining evaluation with irmeasures. In: European Conference on Information Retrieval, pp. 305–310, Springer (2022)

[33] MacAvaney, S., Nardini, F.M., Perego, R., Tonellotto, N., Goharian, N., Frieder, O.: Expansion via prediction of importance with contextualization. In: Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, pp. 1573–1576 (2020)

[34] Mackenzie, J., Trotman, A., Lin, J.: Wacky weights in learned sparse representations and the revenge of score-at-a-time query evaluation. arXiv preprint arXiv:2110.11540 (2021)

[35] Mallia, A., Siedlaczek, M., Mackenzie, J., Suel, T.: PISA: performant indexes and search for academia. In: Proceedings of the Open-Source IR Replicability Challenge co-located with 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, OSIRRC@SIGIR 2019, Paris, France, July 25, 2019., pp. 50–56 (2019), URL http://ceur-ws.org/Vol-2409/docker08.pdf

[36] Messina, N., Amato, G., Esuli, A., Falchi, F., Gennaro, C., Marchand-Maillet, S.: Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **17**(4), 1–23 (2021)

[37] Nam, H., Ha, J.W., Kim, J.: Dual attention networks for multimodal reasoning and matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 299–307 (2017)

[38] Nguyen, T., Hendriksen, M., Yates, A.: Multimodal learned sparse retrieval for image suggestion. In: TREC (2023)

[39] Nguyen, T., MacAvaney, S., Yates, A.: Adapting learned sparse retrieval for long documents. arXiv preprint arXiv:2305.18494 (2023)

[40] Nguyen, T., MacAvaney, S., Yates, A.: A unified framework for learned sparse retrieval. In: European Conference on Information Retrieval, pp. 101–116, Springer (2023)

[41] Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2641–2649 (2015)

[42] Ram, O., Bezalel, L., Zicher, A., Belinkov, Y., Berant, J., Globerson, A.: What are you token about? dense retrieval as distributions over the vocabulary. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2481–2498, Association for Computational Linguistics, Toronto, Canada (Jul 2023)

[43] Sheng, S., Laenen, K., Van Gool, L., Moens, M.F.: Fine-grained cross-modal retrieval for cultural items with focal attention and hierarchical encodings. Computers **10**(9), 105 (2021)

[44] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems **30** (2017)

[45] Wang, H., Sahoo, D., Liu, C., Shu, K., Achananuparp, P., Lim, E.p., Hoi, S.C.: Cross-modal food retrieval: learning a joint embedding of food images and recipes with semantic consistency and attention mechanism. IEEE Transactions on Multimedia **24**, 2515–2525 (2021)

[46] Wang, L., Li, Y., Lazebnik, S.: Learning deep structure-preserving image-text embeddings. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5005–5013 (2016)

[47] Wen, K., Xia, J., Huang, Y., Li, L., Xu, J., Shao, J.: COOKIE: contrastive cross-modal knowledge sharing pre-training for vision-language representation. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, pp. 2188–2197, IEEE (2021)

[48] Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics **2**, 67–78 (2014)

[49] Zamani, H., Dehghani, M., Croft, W.B., Learned-Miller, E., Kamps, J.: From neural re-ranking to neural ranking: Learning a sparse representation for inverted indexing. In: Proceedings of the 27th ACM international conference on information and knowledge management, pp. 497–506 (2018)

[50] Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: Proceedings of the Machine Learning for Healthcare Conference, MLHC 2022, 5-6 August 2022, Durham, NC, USA, Proceedings of Machine Learning Research, vol. 182, pp. 2–25, PMLR (2022)

[51] Zhao, P., Xu, C., Geng, X., Shen, T., Tao, C., Ma, J., Jiang, D., et al.: LexLIP: Lexicon-bottlenecked language-image pre-training for large-scale image-text retrieval. arXiv preprint arXiv:2302.02908 (2023)

[52] Zhuge, M., Gao, D., Fan, D.P., Jin, L., Chen, B., Zhou, H., Qiu, M., Shao, L.: Kaleido-BERT: Vision-language pre-training on fashion domain. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12647–12657 (2021)