Cross-Modal Subspace Learning with Scheduled Adaptive Margin Constraints

David Semedo
NOVALINCS
Universidade NOVA de Lisboa, Portugal
df.semedo@campus.fct.unl.pt

João Magalhães
NOVALINCS
Universidade NOVA de Lisboa, Portugal
jm.magalhaes@fct.unl.pt

ABSTRACT
Cross-modal embeddings, between textual and visual modalities, aim to organise multimodal instances by their semantic correlations. State-of-the-art approaches use maximum-margin methods, based on the hinge-loss, to enforce a constant margin \( m \), to separate projections of multimodal instances from different categories. In this paper, we propose a novel scheduled adaptive maximum-margin (SAM) formulation that infers triplet-specific constraints during training, therefore organising instances by adaptively enforcing inter-category and inter-modality correlations. This is supported by a scheduled adaptive margin function, that is smoothly activated, replacing a static margin by an adaptively inferred one reflecting triplet-specific semantic correlations while accounting for the incremental learning behaviour of neural networks to enforce category cluster formation and enforcement. Experiments on widely used datasets show that our model improved upon state-of-the-art approaches, by achieving a relative improvement of up to \( \approx 12.5\% \) over the second best method, thus confirming the effectiveness of our scheduled adaptive margin formulation.

1 INTRODUCTION

Documents with both visual and textual data have very rich information that span across the two modalities. These modalities naturally co-occur, each adding a unique semantic perspective to a document instance. In this paper, we address the task of cross-modal retrieval, in which one is interested in being able to, given one modality (e.g. text), search by relevant content from the other modality (e.g. images), and vice-versa, in an unified manner. The field of cross-modal embedding learning, has been actively researched \([2, 3, 9, 17, 20, 25, 26, 32, 33]\), with the most widely used approach being representation learning, through subspace learning. The rationale is to solve the heterogeneity problem by learning a common space in which semantically equivalent instances will be structured close together. Namely, projections are learned for each modality, mapping original representation vectors to a semantically correlated space. The maximum-margin formulation, which consists of a variant of the hinge loss, has been adopted lately by most state-of-the-art approaches \([25]\). This loss function enforces a set of hinge loss constraints, over sampled triplets \((target \; instance; \; positive \; instance; \; negative \; instance)\). Namely, it enforces image and text instances of the same category to be close, and instances of different categories to be far apart by at least a fixed margin \( m \). These correlations are then grounded in statistical \([20, 28, 33]\) and temporal correlations \([22]\).

In this paper we propose an adaptive neural structuring cross-modal subspace learning model (SAM), that dynamically organises instances on the new subspace according to their semantic similarity and inter-category correlations. In particular the two main novelties of the proposed method are:

- **Adaptive margin constraints**: we part ways with state-of-the-art methods based on the hinge-loss function with a...
constant margin $m$ between different categories, and introduce a novel adaptive margin function $f_m(\cdot)$ that infers the margin constraints during training.

- **Scheduled activation of adaptive margins**: by considering the incremental learning behaviour of neural architectures [5], we propose a novel scheduled learning algorithm that progressively increases the parameters’ degrees of freedom to allow a shift from coarse-grain (fixed margin $m$) to fine-grain (adaptive margins $f_m(\cdot)$) training, as the model converges to a stable solution. Figure 1 illustrates this shift from epoch 1 to epoch $t$.

These contributions stem from the fact that the hinge-loss function does not adapt the constraints imposed by looking at the current subspace organisation, (e.g. clusters formed), at each training epoch $t$. We posit that semantic information used for subspace structuring should be directly incorporated in the ranking loss formulation, instead of adding extra terms to the main loss function. At the same time, the loss function should adapt the constraints imposed, at each training epoch, according to the current subspace structure and enforce semantic clusters formation, i.e. promote grouping of instances of the same semantic category.

In summary, we formulate an adaptive maximum-margin model, which dynamically adapts subspace structuring constraints over triplets, by jointly using semantic similarity and category clusters enforcement rules to obtain an effective semantic subspace organisation. Experiments on three cross-modal retrieval benchmark datasets, where we compare our method with a considerable number of existing methods, reveal that our model is highly effective, outperforming state-of-the-art works.

## 2 RELATED WORK

### Cross-modal subspace learning

Learning cross-modal embeddings, between visual and textual data, has been an active research topic [2, 3, 17, 20, 25, 33]. In a pioneering work [20], Canonical Correlation Analysis [8] (CCA) was used to learn linear projections for each modality, by learning a set of canonical coefficients, that define a subspace where modalities are maximally correlated. This approach was extended for the multi-label scenario, by using label embeddings to establish correspondences between instances [18]. A multi-view kernel CCA formulation is proposed in [4], where a joint space for visual, textual and semantic information is learned.

Lately, neural methods have proved to be highly effective at learning non-linear projections that capture complex non-linear correlations. The loss function definition is generally the core component, for which several variants have been proposed. Deep Canonical Correlation Analysis (DCCA) was adopted in [33] to match images and text, using non-linear projections. DCCA is a non-linear version of CCA that exploits the fact that the CCA objective function can be formulated based on a matrix trace norm, allowing for gradient-based optimisation. In [3] a neural architecture, the Correlation Autoencoder (Corr-AE), with two uni-modal autoencoders (one per modality) is used, enforcing correlation between learned hidden-representations. Instead of solely focusing in pairwise visual-textual correlations, in [16] extra constraints are added over inter-modal sample relations. Recently, in [17] authors model intra and inter-modality correlations, to unveil complex and fine-grain modality interactions. An adversarial approach is proposed in [25], where a common subspace is learned by a mini-max game between a feature projector and a modality classifier. An effective approach, common to several state-of-the-art approaches is triplet ranking loss [21], in which different triplet mining strategies may be devised.

### Subspace structuring constraints

Apart from maximising correlation between different modalities, additional constraints are usually added to the global loss function. In [10] center-loss is used to minimise intra-category invariance, under a metric learning approach. A successful approach consists of combining intra-modality semantic category and inter-modality pairwise similarity constraints [17, 25, 28]. Such constraints are commonly enforced over sampled triplets. In [28] structure-preserving (hinge loss based) constraints with fixed margins, are used to push semantically similar instances closer to each other. In this paper, we follow a similar intuition, but instead we adaptively change the margin during training to enforce a per-category cluster formation and preservation.

### Maximum-margin learning

To organise data by their semantic correlations, ranking loss is a widely adopted approach for cross-modal subspace learning due to its effectiveness [25, 31]. A set of similarity constraints are formulated under the hinge-loss, enforcing the similarity of positive instances to be far apart from similarity of negative ones, by at least a margin $m$. In state-of-the-art works, this margin is fixed with a constant value for all categories. In fact, this corresponds to a relaxation of the subspace structuring problem, in which the embedding’s semantic similarities are neglected, thus possibly sacrificing optimal data organisation. Following this line of reasoning, Li et al. [14] replace the margin by the mean per joint error function, and in [31] the margin is replaced by the correlation of categories in the original feature space. We depart from the above methods by proposing an adaptive maximum-margin formulation that infers margin values during training.

## 3 CROSS-MODAL SUBSPACE STRUCTURING

### 3.1 Definitions

Let $C$ be a corpus of multimodal instances, where without loss of generality, the visual (images) and textual modalities are considered. Each instance $d^i \in C$ is defined as $d^i = (x^i_D, x^i_T, l^i)$, where $x^i_D \in \mathbb{R}^{D_V}$ and $x^i_T \in \mathbb{R}^{D_T}$ are the instance’s image $d^i_D$ and text $d^i_T$ feature representations, respectively, and $l^i \in L$ the instance category. $L$ is the set of semantic categories. Let $* \in \{V, T\}$ on the remainder of this paper, to avoid notational clumping.

In cross-modal subspace learning, the goal is to learn a subspace in which instances’ textual and visual elements, of the same semantic category, will be maximally correlated. The original feature spaces of $x^*_V$ and $x^*_T$ are dissimilar and cannot be used to perform cross retrieval, as they not only may have different dimensionality but also encode different characteristics and semantics. To this end, for each original modality space, the goal is to learn the projections:

$$P_{\theta_V}() : \mathbb{R}^{D_V} \mapsto \mathbb{R}^D \quad P_{\theta_T}() : \mathbb{R}^{D_T} \mapsto \mathbb{R}^D \quad (1)$$

mapping images $x^*_V$ and texts $x^*_T$ to a common cross-modal subspace, with dimensionality $D$. Similarity between two $\ell_2$ normalised
3.2 Adaptive subspace learning

Modality projections into cross-modal subspaces must capture both inter-category and inter-modality correlations in that subspace. To this end, the cross-modal subspace learning problem is commonly formulated using a maximum-margin learning approach, by imposing a set of constraints over pairwise instance’s similarity, on the target subspace [17, 22, 25, 28, 34].

For an anchor instance $x^a_i$, such constraints enforce the similarity of positive instances $s(x^a_i, x^n_j)$, i.e. sharing one category $l \in L$, to be higher than the similarity of negative samples $s(x^n_i, x^n_j)$, i.e. not sharing a category, by at least a margin $m$. This constraint is formulated as:

$$s(x^a_i, x^n_j) > s(x^n_i, x^n_j) + m. \quad (2)$$

The constraint above is then enforced over each pair of instances, resulting in a considerable large set of constraints. For training, such constraints are then relaxed using the hinge loss [7].

3.2.1 Static maximum-margin formulation. We start by formulating a loss function $L_i$, under this framework, by imposing maximum-margin constraints over the two modality directions (image $\rightarrow$ text and text $\rightarrow$ image), thus simultaneously capturing inter-modality and inter-category correlations. Namely, at every training epoch $t$, given triplets of the form $(x^a_i, x^n_j, x^n_k)$, where $x^a_i$ and $x^n_j$ stand for positive and negative instances, respectively, w.r.t. an anchor $x^a_i$, we compute the model loss,

$$L(t, \theta) = \sum_{i,j} \max(0, m - s(x^a_i, x^n_j) - s(x^a_i, x^n_k)) +$$

$$\max(0, m - s(x^n_i, x^n_j) - s(x^n_i, x^n_k)),$$

$$s(x^a_i, x^n_j) > s(x^n_i, x^n_j) + m. \quad (3)$$

where $m$ denotes the margin and $\theta$ the model parameters. Note that unlike other cross-modal subspace learning works [17, 25, 31], the positive instance $x^n_j$ from each triplet is only the opposite modality of the same instance $d^i$, i.e. $x^n_j = x^n_i$ or $x^n_j = x^n_k$. A negative sampling strategy is then applied to mine triplets that respect these conditions.

3.2.2 Adaptive maximum-margin formulation. The maximum-margin formulation defined in eq. 3 assumes that any two instances from different categories are equally correlated. This is reflected by the adoption of a constant margin $m$.

Inspired by maximum-margin structured SVM [24] formulation, we propose to (1) incorporate inter-category semantic correlations into the subspace structuring and (2) guide the projection learning algorithm, at each epoch, with structure preserving constraints that are derived from the current state of the subspace. To achieve this, we propose an adaptive margin formulation, defined by a non-negative margin function $f_m(d^a, d^n, t)$, where $d^a$ and $d^n$ correspond to semantically different instances (i.e. belong to different categories) and $t$ denotes the current epoch of the subspace training algorithm. The margin constraints, for every instance pair, at epoch $t$, are then reformulated as:

$$s(x^a_i, x^n_j) > s(x^n_i, x^n_j) + f_m(d^a, d^n, t). \quad (4)$$
The rationale enclosed in this formulation is that for each pair of instances of different categories, \( f_m(t) \) outputs a margin that encodes the degree of separation between the considered categories. On every epoch \( t \), the margin is linked to the pairwise correlation of the instances’ original feature vectors and current subspace structure. Accordingly, the adaptive subspace learning loss function \( \mathcal{L}_{SAM} \), at epoch \( t \) becomes:

\[
\mathcal{L}_{SAM}(t, \theta) = \sum_{p,n} \max(0, f_m(d^a, d^n, t) - s(x^a_T, x^n_T) + s(x^a_V, x^n_V)) + \sum_{p,n} \max(0, f_m(d^a, d^n, t) - s(x^a_T, x^n_V) + s(x^a_V, x^n_T)).
\]

(5)

Similarly to maximum margin methods, this formulation guides the model towards incorporating semantic information, regarding intra-category pairwise correlations. However, we observe that the difference of the similarities between positive and negative instances are, on average, inter-category specific. Therefore, we account for the current subspace organisation (at each epoch \( t \)), to decide what should be the magnitude of the margin, i.e. \( f_m \).

### 3.3 Scheduled activation of adaptive margins

For neural subspace learning, in the first gradient updates, the subspace organisation is expected to be highly volatile, constantly changing at each epoch. It follows that for neural networks trained using stochastic gradient descent, it is not trivial to estimate beforehand when (i.e. at each epoch) is the model about to converge. Thus, we propose an approximation strategy that imposes a hard (i.e. a static high magnitude) margin on all triplet constraints on the first few epochs. This allows the model to find an initial coarse organisation of the subspace. Then, as the number of epochs progress, the static constraints give way to triplet specific constraints, that better capture the fine-grain interactions among instances.

Inspired by adaptive strategies for neural network training, such as the Adam [12] optimiser, which schedules different learning rates, we propose a smoothed scheduled shift function from static to an adaptive maximum-margin formulation, as the training algorithm converges (Figure 3). To this end, a scheduled adaptive margin function \( f_m \) is defined as:

\[
f_m(d^a, d^n, t) = \alpha(t) \cdot f_{am}(d^a, d^n, t) + (1 - \alpha(t)) \cdot m
\]

s.t. \( \alpha(t) = \frac{1}{1 + e^{-k(t-f_a n_p)}} \).

(6)

where the \( \alpha(t) \) is a scheduler function, defined as a compressed sigmoid, that gradually activates the adaptive margin, according to the current epoch \( t \). The \( \alpha(t) \) function is defined by a smoothing term \( k \), the total number of epochs \( n_p \) and an activation factor \( f_a \in [0, 1] \). Figure 3 illustrates how each parameter is used to define \( \alpha(t) \).

### 3.4 Adaptive margin function definition

In this section we describe how the adaptive margin function \( f_{am}(d^a, d^n, t) \) is materialised. We formulate \( f_{am} \) such that it implements an adaptive margin, encoding: a) the semantic correlation estimated from original modality features – between instances from different categories, and b) cluster formation enforcement, for each semantic category, according to the epoch \( t \) of the algorithm. Figure 2 illustrates \( f_{am} \) components. In particular, we define the adaptive margin function as

\[
f_{am}(d^a, d^n, t) = \lambda \cdot f_{ms}(d^a, d^n) + (1 - \lambda) \cdot f_{mc}(d^a, d^n, t),
\]

(7)

where \( f_{ms} \) quantifies semantic correlation, and \( f_{mc} \) the similarity between category clusters at epoch \( t \), of two instances \( d^a \) and \( d^n \). The parameter \( \lambda \) models the trade-off between the two components.

**Semantic inter-category pairwise correlations.** From a semantic standpoint, pairwise correlations across categories, will be different (e.g. instances from category sky are expected to be more correlated with instances from clouds than from flowers). In our neural subspace structuring model, the function \( f_{ms} \) accounts for such semantic correlations by evaluating similarity on each modality original spaces. The function \( f_{ms} \) is then defined as:

\[
f_{ms}(d^a, d^n) = \frac{||x^a_T - x^n_T||_2 + ||x^a_V - x^n_V||_2}{2}.
\]

(8)

From the definition, \( f_{ms} \) averages the semantic similarity of both visual and textual modalities, extracted from the modalities’ original feature space. The output of this function is normalized to \([0, 1]\).

**Category cluster formation and preservation.** Given a randomly initialised neural network model, it can converge to different local optima, thus resulting in different subspace organisation. From this observation, we pose that for near convergence epochs, it is important to restrict model updates, preserving currently formed category clusters and forcing instances to move towards their category cluster. As a generalization, the centroid of a given category \( l \) is computed as:

\[
P_{\theta_c}(l, t) = \frac{1}{|\{x^a_\gamma : \gamma = l\}|} \sum_{x^a_\gamma \in \{x^a_\gamma : \gamma = l\}} \mathcal{P}_{\theta}(x^a_\gamma, t).
\]

(9)

To materialise the described behaviour, we rely on the cosine distance \( d \) to define \( f_{mc} \) as:

\[
f_{mc}(d^a, d^n, t) = \frac{1}{2} \left[ d(P_{Vc}(l^n, t), P_{Vc}(l^n, t)) + d(P_{Vc}(l^a, t), P_{Vc}(l^a, t)) \right],
\]

(10)

where for a given category \( l \), \( P_{Vc}(l, t) \) and \( P_{Tc}(l, t) \) denote the centroid of the visual and textual projections, at epoch \( t \). \( d \) stands for the cosine distance \( 1 - s(\cdot, \cdot) \), with \( s \) being normalised a priori to \([0, 1]\) range. Essentially, given a pair of instances, \( f_{mc} \) evaluates the distance between the corresponding category centroids, for
We jointly learn both the cross-modal projections while adaptively performing neural subspace structuring, by minimising the function:

$$\arg \min_{\theta_V, \theta_T} L_{SAM}(\theta_V, \theta_T)$$

(11)

where $L_{SAM}$ adaptively organizes instances according to their inter-category and inter-modal correlations. Pseudocode is illustrated in Algorithm 1. A stochastic sampling strategy is adopted, in which to evaluate $L_{SAM}(\theta_V, \theta_T)$, negative samples are sampled directly from mini-batches. At each epoch, all samples are seen by the network. This approach severely reduces the model complexity, while still achieving convergence. The whole model is then optimised using Stochastic Gradient Descent.

Algorithm 1: Pseudocode for SAM optimisation.

**Initialization:** Corpus $C = \{d_1, \ldots, d^m\}$ of multimodal instances, with $d^i = (x^i_1, x^i_2, \ldots, x^i_T)$; initialise network weights: $\theta_V, \theta_T$; hyperparameters: $\lambda, k, f_a$, subspace dimensionality $D$, learning rate $\eta$, mini-batch size $b$;

1. **repeat until convergence:**
   2. **for** $t$ epochs **do**
   3. Sample mini-batch to create triplets of the form $(x^i_1, x^j_2, x^k_2)$ and $(x^j_1, x^i_2, x^k_2)$;
   4. Update $\theta_V$ and $\theta_T$ through BP, with stochastic gradients, using $\alpha(t)$:
   5. $\theta_V \leftarrow \theta_V - \eta \cdot \nabla_{\theta_V} \frac{1}{b} L_{SAM}$;
   6. $\theta_T \leftarrow \theta_T - \eta \cdot \nabla_{\theta_T} \frac{1}{b} L_{SAM}$;
   7. Update the weight of the adaptive margin:
   8. $\alpha(t + 1) \leftarrow \frac{1}{1 + e^{-k(\theta_T - \theta^1)}}$
   9. **end for**
10. **return** projection networks, $P_{\theta_V}(\cdot)$ and $P_{\theta_T}(\cdot)$.

Both visual and textual projections. Grounding the margin on $f_{mc}$ simultaneously enforces cluster formation and preservation. This is achieved because during training, the function $f_{mc}$ will simultaneously attempt to preserve the current subspace organisation and push hard aligned projections towards the corresponding category centroid.

### 3.5 Neural architecture

To learn projections $P_{\theta_V}(\cdot)$ and $P_{\theta_T}(\cdot)$, we consider a two decoupled network architecture, to learn non-linear mappings, predominant across multiple state-of-the-art works [2, 3, 15, 25, 33]. The networks are jointly trained by a common loss function $L_{SAM}$. For each modality, a feedforward network $f_a$ maps original modality representations onto $S$, comprising 2 fully connected layers (with dimensions 1024 and $D$, respectively) and $\tanh$ non-linearities. For semantically rich image feature representation, each $x^i_1$ is obtained from a pre-trained CNN. Each $x^i_T$ is represented as a bag-of-words vector. Then, $f_a$ takes its corresponding modality as input: $d^i_f$. RGB image for $f_v$ and $d^i_T$ bag-of-words text representation for $f_T$. Figure 2 depicts the full architecture.

### 3.6 Training and inference

We jointly learn both the cross-modal projections $P_{\theta_V}(\cdot)$ and $P_{\theta_T}(\cdot)$, while adaptively performing neural subspace structuring, by minimising the function:

### 4 EVALUATION

#### 4.1 Datasets

We evaluate our proposed methods in three widely used cross-modal retrieval benchmark datasets. **Wikipedia** [20]. Comprised by a total of 2,866 visual-textual pairs, extracted from Wikipedia’s “featured articles”, where each article is accompanied by a single image. Each article is annotated with 10 semantic categories. We split the dataset following [3, 16, 20], with 2,173 instances for training, 231 for validation, and 462 for testing. **NUS-WIDE** [1]. The NUS-WIDE dataset is comprised by a total of 269,648 instances (images and corresponding tags), from the Flickr network, annotated with one or more categories from a total of 81 distinct semantic categories. For comparison, we follow the protocol of Peng et al. [17]; only instance pairs that belong to a single category are kept and the instances from the 10 categories with more instances are chosen. This results in more than 60,000 instances. Splits are created following [17], resulting in 23,661 instances for testing, 5,000 for validation and the remaining for training. **NUS-WIDE-10K** is a subset of NUS-WIDE created by strictly following the protocol of [3]; the 10 categories with more instances are chosen, and for each category, 1000 instances are sampled. Only pairs that belong to a single category are considered. Three splits, equally balanced w.r.t. to number of instances per category, are sampled randomly: 8,000 instances for training, 1,000 for validation and 1,000 for testing. **Pascal Sentence** [19]. Comprised by 1,000 visual-textual pairs, from the 2008 PASCAL development kit, categorised within 20 categories, with instances evenly distributed across categories. We follow [3, 16] and randomly and evenly split the dataset with 800 instances for training, 100 for validation and 100 for testing.

#### 4.2 Methodology

We evaluate the retrieval performance using mean Average Precision (mAP), which is the standard evaluation metric for cross-modal retrieval [3, 17, 20, 25, 27, 33]. We follow [11, 17, 20, 35] and compute mAP for all the retrieved results. For mAP, an instance is relevant if it has the same category. Two tasks are evaluated: 1) Image-to-Text retrieval ($I \rightarrow T$) and 2) Text-to-Image ($T \rightarrow I$) retrieval. Core parameters of SAM are analysed to assess their impact in the performance. Each mAP result reported of our method corresponds to the average of 5 runs.

#### 4.3 Implementation details

Networks are jointly trained using Stochastic Gradient Descent, with 0.9 Nesterov momentum, and a learning rate $\eta = 5 \times 10^{-3}$, with a decay of $1 \times 10^{-6}$. The model with lowest validation error is kept. Mini-batch size is set to 200 for all datasets, and the total number of epochs is set to 100. The projections dimension is set to $D = 200$ and the margin $m = 1.0$. For each neuron, $\tanh$ non-linearities are applied. Dropout with $p = 0.1$ is applied to the first hidden layer. Images representations are obtained by feeding each individual image through a pre-trained VGG-19 [23] convolutional network, and extracting the output of the last fully connected layer.

1 Top-10 categories: ‘person’, ‘animal’, ‘sky’, ‘window’, ‘water’, ‘flowers’, ‘food’, ‘toy’, ‘grass’, ‘clouds’.
We compare our proposed approach, SAM, with a total of 11 state-of-the-art methods for subspace learning, we obtain a relative improvement of approximately 9% compared to all methods, on both cross-modal retrieval directions. It outperforms all the compared methods, on both Cross-modal retrieval directions. We believe that the reason is that the best result was obtained with \( \lambda = 0.05, f_a = 0.9 \) and \( k = 0.1 \). Hence, in contrast to the results on the Pascal sentences dataset, the method started activating the adaptive margin near the last epochs of training. Moreover, once again, more importance was given to the cluster enforcement and preservation (95% of the weight).

Our method obtains a high mAP on both directions, but performs better on the \( T \rightarrow I \) direction. We believe that the reason is that visually, some categories have very similar content (e.g., sky vs. clouds). However, the text in this dataset correspond to tags, which due to the sparsity of BoW representation, turns out to have good discriminative properties.

### Wikipedia dataset
As with the previous datasets, our method outperforms all the compared methods. On the Wikipedia dataset, categories are very broad (e.g., Art & Architecture, Media, etc.), with texts and images of the same category being highly diverse. Therefore, in this dataset, given the small amount of instances available for training, it is harder to align modalities. As this is reflected in original feature representations, the function \( f_{mc} \), which organises instances according to semantic similarity on original features, ends up not helping structuring the space. Supporting this observation is the fact that the best result was obtained with \( \lambda = 0.05 \). The category cluster formation and preservation, enforced by function \( f_{mc} \), provides the major contribution to the effectiveness.

To further complement our evaluation, we also compare our method against CMOLRS [31], which formulated the margin as an original-feature driven margin that is fixed during training, i.e., using only a simplified version of \( f_{mc} \) factor of SAM. On the Wikipedia dataset, CMOLRS achieved a mAP@100 of 0.413 while SAM achieves a mAP@100 of 0.541. As authors only report mAP@100, we did not included it in table 1. This confirms the importance of dynamically adjusting margin values during training and of the novel cluster formation and preservation component \( f_{mc} \).

### Large-scale NUS-WIDE
To further explore the generalisation of SAM algorithm, we evaluated SAM in the large-scale full NUS-WIDE dataset. Table 2 supports the same conclusions that where

---

**Table 1: mAP performance results across different datasets. The second half of the table concern deep-learning methods.**

| Method  | Pascal Sentences | NUS-WIDE-10k | Wikipedia |
|---------|-----------------|--------------|-----------|
|         | \( I \rightarrow T \) | \( T \rightarrow I \) | \( I \rightarrow T \) | \( T \rightarrow I \) | \( I \rightarrow T \) | \( T \rightarrow I \) |
| CCA [8] | 0.203 | 0.208 | 0.206 | 0.417 | 0.406 | 0.342 |
| CFA [13] | 0.476 | 0.470 | 0.473 | 0.406 | 0.435 | 0.421 |
| KCQA [6] | 0.488 | 0.446 | 0.467 | 0.351 | 0.356 | 0.354 |
| LGCF [11] | 0.539 | 0.503 | 0.521 | 0.453 | 0.485 | 0.469 |
| JRL [35] | 0.563 | 0.505 | 0.534 | 0.466 | 0.499 | 0.483 |
| CCL [17] | 0.538 | 0.544 | 0.541 | 0.519 | 0.542 | 0.531 |
| SAM \((\alpha(t) = 1, \lambda = 1)\) | 0.586 | 0.659 | 0.659 | 0.563 | 0.594 | 0.579 |
| SAM | 0.637 | 0.643 | 0.640 | 0.563 | 0.640 | 0.637 |

(\( fc7 \)). For texts, we adopt a BoW representation, with 1000-D vocabulary size for NUS-WIDE-10k and Pascal Sentences, and 3000-D for Wikipedia.

## 4.4 Cross-modal retrieval
We compare our proposed approach, SAM, with a total of 11 state-of-the-art works, on the task of cross-modal retrieval. Namely, we compare against CCA [8], CFA [13], KCQA [6], Corr-AE [3], JRL [35], LGCF [11], DCCA [33], CMDN [16], Deep-SM [29], ACMR [25] and CCL [17].

**Pascal sentences dataset.** Table 1 shows the results obtained. Our method outperforms all the compared methods, on both \( I \rightarrow T \) and \( T \rightarrow I \) settings. Namely, SAM achieved a relative improvement of \( \approx 12.5\% \), with respect to the second best performing method, CCL. CCL models intra-modality and inter-modality correlations through distinct constraints, using a strategy that balances both types of correlation constraints. These are then superseded by a ranking loss function in which a static margin is used. Instead, SAM adopts an adaptive margin formulation, in which intra and inter-modality correlations are directly modelled in a single constraint. The best result was achieved with \( \lambda = 0.25, f_a = 0.4 \) and \( k = 0.1 \). SAM started smoothly activating the adaptive margin at about half the training epochs, revealing preference for starting using \( f_{mc} \) sooner. The semantic similarity component \( f_{ms} \) plays an important role in organising the space. Notwithstanding, the component \( f_{mc} \) has revealed to be the most important one (75%), effectively guiding the subspace structuring.

**NUS-WIDE-10k dataset.** From the results on table 1, we can see that our method also achieved the best performance when compared to all methods, on both cross-modal retrieval directions. It outperformed both traditional cross-media models (top rows of table 1) and the most recent deep learning methods. With respect to the second best performing method, ACMR, which uses an adversarial approach for subspace learning, we obtain a relative improvement of \( \approx 9\% \), on the average of \( T \rightarrow I \) and \( I \rightarrow T \). This confirms the importance of moving towards an adaptive margin formulation.
we computed the average margin value, imposed to all triplets, on each epoch \( t \), on the NUSWIDE-10k dataset. Figure 4 shows the euler function \( \alpha(t) \) over each epoch (\( t \)). In this section we examine behaviour of SAM by looking at the

4.5 Scheduled adaptive margins analysis

4.5.1 Average margin values vs. scheduler function. The scheduler function \( \alpha(t) \) shifts from a high-magnitude constant margin \( m = 1 \), to the adaptive margin \( f_m \). To inspect this behaviour, we computed the average margin value, imposed to all triplets, on each epoch \( t \), on the NUSWIDE-10k dataset. Figure 4 shows the average \( f_m \) value (blue line) versus the scheduler function value \( \alpha(t) \) (green line), over the training epochs. It can be observed that at each epoch, the average margin imposed by \( f_m \) tends to be smaller. One can also observe that \( \alpha(t) \) has a sigmoidal shape.

4.5.2 Average margin values for each Category. In order to provide a deeper understanding of what the model achieves, we show in Figure 5, also on the NUSWIDE-10k dataset, the average margin values between three pairs of categories at each training epoch \( t \), and a projection of the final cross-modal subspace.

These experiments are crucial to understand the underpinnings of SAM: Figure 5 confirms that the average margin value gradually decreases during training, with triplet constraints over window-sky category having lower magnitude margins than window-grass, thus reflecting visual and textual semantic similarity as intended.

Figure 6 delves into this question and shows the average margin value per category imposed by \( f_m \), against triplets of the remaining categories, at each epoch \( t \). Given the target category \( c_1 \) of each plot, each line corresponds to a category \( c_2 \). Namely, it corresponds to the average of the margin values, imposed by \( f_m \), to triplets with the positive instance belonging to category \( c_1 \) and the negative belonging to category \( c_2 \). It is interesting to note that all margins are significantly different. In particular, categories grass and person are the ones with most homogenous margins. In contrast, categories sky and animal took full advantage of the scheduled adaptive margins and ended up with very different margins to all other categories.

4.5.3 Scheduler and \( f_{mc} \) impact. The scheduler, together with the cluster formation and enforcement \( f_{mc} \) component of the adaptive margin, are key novel components, responsible for achieving state-of-the-art performance. To confirm this, we evaluated SAM

| Methods | \( I \mapsto T \) | \( T \mapsto I \) | Avg. |
|---------|----------------|---------------|------|
| CCA [8] | 0.244          | 0.275         | 0.260|
| CFA [13] | 0.358          | 0.361         | 0.360|
| KCCA [6] | 0.348          | 0.481         | 0.415|
| LGCFL [11] | 0.512        | 0.600         | 0.356|
| JRL [35] | 0.615          | 0.592         | 0.604|
| Corr-AE [3] | 0.391        | 0.429         | 0.410|
| DCMA [33] | 0.475          | 0.500         | 0.488|
| CMDN [16] | 0.643          | 0.626         | 0.635|
| CCL [17] | 0.671          | 0.676         | 0.674|
| **SAM** | **0.701**      | **0.707**     | **0.704**|

| Methods | \( I \mapsto T \) | \( T \mapsto I \) | Avg. |
|---------|----------------|---------------|------|
| CCA [8] | 0.244          | 0.275         | 0.260|
| CFA [13] | 0.358          | 0.361         | 0.360|
| KCCA [6] | 0.348          | 0.481         | 0.415|
| LGCFL [11] | 0.512        | 0.600         | 0.356|
| JRL [35] | 0.615          | 0.592         | 0.604|
| Corr-AE [3] | 0.391        | 0.429         | 0.410|
| DCMA [33] | 0.475          | 0.500         | 0.488|
| CMDN [16] | 0.643          | 0.626         | 0.635|
| CCL [17] | 0.671          | 0.676         | 0.674|
| **SAM** | **0.701**      | **0.707**     | **0.704**|

drawn from the previous analysis. It is also noticeable, that all models improved thanks to the larger training dataset.

Overview. In overall, our method has proven to be effective, out-performing previous state-of-the-art methods on all datasets. The cluster enforcement and preservation component \( f_{mc} \) proved to be crucial to achieve state-of-the-art performance. Unlike most methods, which impose extra constraints by augmenting a projection network by adding additional loss terms, our approach imposes those constraints by directly adapting the margin between instance pairs during training, thus resulting in a simpler but effective model.

By modelling the semantic inter-category pairwise correlations, our model is able to transfer semantic correlations from the original feature space directly to the common subspace. Then, by enforcing cluster formation after achieving a stable subspace organisation, our method improves significantly from state-of-the-art works.

Figure 4: Global average adaptive margin \( f_m \) over training epochs (\( t \)). The left y-axis corresponds to the \( f_m \) value and the right y-axis to the scheduling function \( \alpha(t) \) value.

Figure 5: Analysis of the margin values over each epoch (\( t \)), between three categories (left), versus final t-SNE model projections (right).
with the scheduler deactivated ($\alpha(t) = 1$) and with $f_{mc}$ disabled ($\lambda = 1$). As can be seen from Table 1, this results in a drop of performance of $\approx 8\%$, $\approx 5\%$ and $\approx 19\%$, on Pascal Sentences, NUS-WIDE-10k and Wikipedia, respectively, confirming the crucial importance of the scheduler and $f_{mc}$.

### 4.6 Analysis of activation phase

In this section we will analyse the impact of the activation phase $f_a$ and the semantic correlation vs. cluster enforcement trade-off $\lambda$ parameter. To do this, we measure the mAP score on the Pascal Sentences dataset. Namely, we evaluate $f_a \in \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ and $\lambda \in \{0.0, 0.1, 0.25, 0.75, 1.0\}$, fixing all the remaining parameters, and show the results in Figure 7. The x-axis represents the value of $f_a$ and the y-axis the mAP score obtained. Each curve corresponds to a value of $\lambda$.

The first observation is that imposing the adaptive margin too early is bad. For instance, when $f_a$ is close to zero, the method starts using the adaptive margin from the beginning of the training, resulting in low performance. This confirms our intuition that in the first training iterations, the subspace is still coarsely organised. As the parameter $f_a$ increases, we can see that the results improve significantly, reaching a performance peak on $f_a = 0.4$, for four of the five experimented values of $\lambda$. Namely, smoothly activating the adaptive margin with $f_a = 0.4$, and giving around 75% weight to $f_{mc}$ (cluster formation and preservation component) and the remaining to $f_{ms}$, leads to the best performance. For all values of $\lambda$, activating the adaptive margin too late leads to significant performance drops. This is due to the fact that by activating later, the network has more chances to overfit using a static margin. At this point, neither the cluster formation $f_{mc}$, nor the semantic correlations $f_{ms}$ components are able to improve the subspace organisation. Regarding the trade-off parameter $\lambda$, we observe the trend that cluster formation has a higher impact on achieving better performance than semantic correlation, with peak performance occurring when both components are active.

### 5 CONCLUSIONS

In this paper we described a novel method to learn cross-modal embeddings. The method introduces a scheduled activation of adaptive margins that allow for triplet specific margins. The key takeaways of the proposed method are:

- **Adaptive margin constraints**: our approach impose general constraints while training the model by adapting the margins between instance pairs. This overcomes the fact that using a unique margin for all pairs is insufficient to adequately structure the subspace.
- **Effective learning of pair-specific margins**: results show that adaptive margins deliver state-of-the-art results. This is further possible due to the pair-specific margins that are learned by the model as illustrated by experimental results.
- **Scheduled learning**: new neural-network training approach was introduced that progressively activates the adaptive margin function, through an epoch-aware scheduling strategy.

As future work, we plan to generalise adaptive ranking loss. Current results hint that the same principle can encode different constraints and thus be extended to other multimedia modelling tasks.
