



HAL
open science

Deep representation design from deep kernel networks

Mingyuan Jiu, Hichem Sahbi

► **To cite this version:**

Mingyuan Jiu, Hichem Sahbi. Deep representation design from deep kernel networks. *Pattern Recognition*, 2019, 88, pp.447-457. 10.1016/j.patcog.2018.12.005 . hal-02325793

HAL Id: hal-02325793

<https://hal.science/hal-02325793>

Submitted on 18 Dec 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Deep Representation Design from Deep Kernel Networks

Mingyuan Jiu^{a,*}, Hichem Sahbi^b

^a*School of Information Engineering, Zhengzhou University, Zhengzhou, China*

^b*CNRS, LIP6 UPMC, Sorbonne University, Paris, France*

Abstract

Deep kernel learning aims at designing nonlinear combinations of multiple standard elementary kernels by training deep networks. This scheme has proven to be effective, but intractable when handling large-scale datasets especially when the depth of the trained networks increases; indeed, the complexity of evaluating these networks scales quadratically w.r.t. the size of training data and linearly w.r.t. the depth of the trained networks.

In this paper, we address the issue of efficient computation in Deep Kernel Networks (DKNs) by designing effective maps in the underlying Reproducing Kernel Hilbert Spaces (RKHS). Given a pretrained DKN, our method builds its associated Deep Map Network (DMN) whose inner product approximates the original network while being far more efficient. The design principle of our method is greedy and achieved layer-wise, by finding maps that approximate DKNs at different (input, intermediate and output) layers. This design also considers an extra fine-tuning step based on unsupervised learning, that further enhances the generalization ability of the trained DMNs. When plugged into SVMs, these DMNs turn out to be as accurate as the underlying DKNs while being at least an order of magnitude faster on large-scale datasets, as shown through extensive experiments on the challenging ImageCLEF, COREL5k benchmarks and the Banana dataset.

*Corresponding author

Email addresses: iemyjiu@zzu.edu.cn (Mingyuan Jiu), hichem.sahbi@lip6.fr (Hichem Sahbi)

Keywords: Multiple kernel learning, kernel design, deep networks, efficient computation, image annotation.

1. Introduction

Kernel design has been an active field of machine learning during the last two decades with many innovative kernel-based algorithms successfully applied to various tasks, including support vector machines (SVMs) for pattern classification and support vector regression for multivariate estimation [1, 2, 3, 4] as well as kernel-PCA for dimensionality reduction [5]. The success of these kernel-based algorithms is highly dependent on the choice of kernels; the latter are defined as symmetric and positive semi-definite functions that return similarity between data [6, 7]. Various kernels have been introduced in the literature [7] including standard elementary kernels (linear, polynomial, Gaussian, histogram intersection, etc.) as well as sophisticated ones that model more complex relationships between data [3]. However, in practice, knowing a priori which (elementary or sophisticated) kernel is suitable for a given task is not obvious and research has recently been undertaken in order to train suitable kernels for different classification tasks (see for instance [8, 9, 10, 11, 12]).

Among existing solutions, Multiple Kernel Learning (MKL) [8, 13, 14] has been popular; its principle consists in learning (sparse or convex) linear combinations of elementary kernels that maximize performances for a given classification task. Different MKL algorithms have been proposed in the literature, including constrained quadratic programming [8], second-order cone and semi-infinite linear programming [13, 15] as well as “simpleMKL” based on mixed-norm regularization [14]. Wang et al. [16] also propose an alternative algorithm with hybrid kernel alignment maximization to obtain the multiple kernel coefficients. In spite of their relative success these solutions hit two major limitations: on the one hand, the convexity of these simple linear MKL models may limit the space of possible (and also relevant) solutions. On the other hand, MKL solutions, relying on shallow kernel combinations, are less powerful (compared to their deep

28 variants) in order to capture different levels of abstractions in the learned kernel
 29 similarity. Considering these two issues, nonlinear and deep architectures have
 30 been recently proposed and turned out to be more effective: for instance, hi-
 31 erarchical multiple kernel learning is proposed in [17] where elementary kernels
 32 are embedded into acyclic directed graphs while in [18], nonlinear combination
 33 of polynomial kernels are used. Following the spirit of deep convolutional neural
 34 networks [19, 20, 21], authors in [9] adopt kernel functions as a prior knowledge
 35 for regularization. In [22], Cho and Saul propose Arc-cosine kernels that mimic
 36 the computation of large neural nets which can be used in shallow as well as deep
 37 networks. In [23], a multi-layer nonlinear MKL framework is proposed, but it is
 38 restricted to only two layers; in this solution, an exponential activation function
 39 is applied to each intermediate and output kernel combination. In [24], Jiu and
 40 Sahbi extend this method to a deeper network of more than two layers, referred
 41 to as Deep Kernel Network (DKN), using a semi-supervised setting that takes
 42 into account the topology of training and test data. In all the aforementioned
 43 MKL algorithms, the computational complexity of kernel (gram-matrix) eval-
 44 uation is a major issue that limits the applicability of these methods; indeed,
 45 considering a dataset with N samples, this complexity reaches $O(LN^2)$ with L
 46 being the depth of the deep kernel networks; this evaluation process is clearly
 47 intractable even on reasonable size datasets. Following the proposed DKNs
 48 in [24], we introduce explicit Deep Map Network (DMN) representations which
 49 substantially improve efficiency while maintaining high discrimination power
 50 of the original DKNs. The design principle of these DMNs is not only super-
 51 vised (as in [24]), but also unsupervised and this provides better approximation
 52 accuracy.

53 Existing solutions that reduce the computational complexity of evaluating
 54 kernels consider explicit maps. In this respect, different solutions have been pro-
 55 posed in the literature including: the Nyström expansion [25] which generates
 56 low-rank kernel map approximations of original gram-matrices from uniformly

57 sampled data without replacement¹, and the random Fourier sampling (pro-
58 posed by Rahimi and Recht [28] and extended to group-invariant kernel method
59 in [29]) which builds explicit features for stationary kernels using random sam-
60 pling of the Fourier spectrum. Nyström-based approximation is also studied
61 in [30] for kernel subspace learning and employed for nonlinear (kernelized) up-
62 date of the Ho-Kashyap algorithm using squared misclassification losses [31].
63 In [32], explicit feature maps for additive homogeneous kernels are given and
64 finite approximations are derived based on spectral analysis, while in [33] data-
65 independent random projections are studied for homogeneous polynomial ker-
66 nels. Other works have been undertaken including random features [34] and
67 convolutional kernel networks [35], which approximate maps of Gaussians using
68 convolutional neural networks.

69 In this paper, we propose a novel method that reduces the computational
70 complexity of DKN evaluation (and therefore SVM learning) on large datasets.
71 We address the issue of kernel map approximation for *any* deep nonlinear combi-
72 nation of elementary kernels rather than one specific type of kernels as achieved
73 in the aforementioned related work. Our solution relies on the *positive semi-*
74 *definiteness* (p.s.d) of existing elementary kernels (linear, polynomial, etc.) and
75 the *closure properties* of p.s.d with respect to different operations (including
76 product, addition and exponentiation) in order to express DKN with DMN.
77 In these closure properties, linear combinations of kernels correspond to con-
78 catenations of their respective maps, while products correspond to Kronecker
79 tensor operations, etc. As some elementary kernels² used to feed the inputs of
80 DKNs may have infinite dimensional or undefined maps, we consider new ex-
81 plicit maps that accurately approximate these elementary kernels. Considering
82 these maps as inputs, this greedy process continues layer-wise in order to find
83 all the maps of the subsequent (intermediate and output) layers. Note that the
84 contribution presented in this paper is an extension of our preliminary work

¹Bounds, on Nyström approximation and sampling, are given in [26, 27]

²such as Gaussian and Histogram Intersection.

85 in [36], but it differs at least in three aspects: first, we consider an unsuper-
 86 vised training criterion that benefits from abundant unlabeled data in order to
 87 further decrease the approximation error of the trained DMN and thereby mak-
 88 ing its generalization power as high as the underlying DKN (and also better
 89 than existing elementary and shallow kernel combinations; as shown through
 90 experiments). Furthermore, with DMNs, one may employ efficient SVM learn-
 91 ing algorithms based on stochastic gradient descent [37] on large-scale datasets,
 92 rather than usual training algorithms that rely on heavy gram-matrices and
 93 intractable quadratic programming problems. All these statements are cor-
 94 roborated through extensive experiments measuring approximation accuracy
 95 and discrimination power as well as efficiency using different image annotation
 96 benchmarks (namely ImageCLEF Photo Annotation [38] and COREL5k [39])
 97 as well as another classification task using the Banana dataset.

98 The rest of this paper is organized as follows: in Section 2 we first briefly
 99 remind DKNs, and then in Section 3 we introduce a novel method that builds
 100 their equivalent DMNs. In Section 4, we describe an unsupervised setting of
 101 our DMN design while in Section 5, we present the experimental validation
 102 of our method on image annotation tasks using ImageCLEF and COREL5k
 103 benchmarks. Finally, we conclude the paper while providing possible extensions
 104 for a future work.

105 2. Deep kernel networks at a glance

A deep kernel network [23, 24] is a multi-layered architecture that recursively
 defines nonlinear combinations of elementary kernels (linear, Gaussian, etc.).
 Let $\kappa_p^{(l)}(\cdot, \cdot)$ denote a kernel function assigned to unit p and layer l ; $\kappa_p^{(l)}$ is
 recursively defined as the output of a nonlinear activation function³ (denoted g)
 applied to a weighted combination of (input or intermediate) kernels from the

³For instance, exponential function [23].

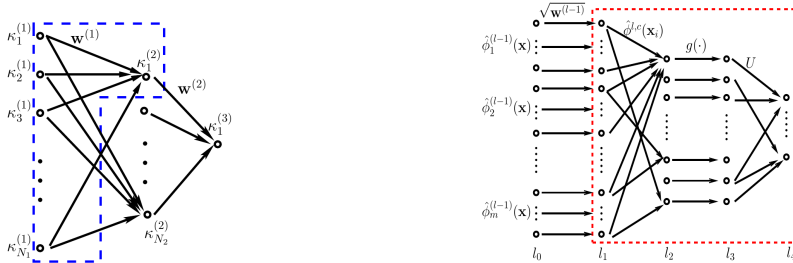


Figure 1: Left: a three-layer deep kernel network (DKN). Right: a sub-module of deep map network (DMN). The blue dash in the left figure denotes a sub-module of DKN where each node stands for a kernel. The input in the right figures are kernel maps and each unit stands for a feature.

preceding layer $(l - 1)$ as

$$\kappa_p^{(l)}(\cdot, \cdot) = g\left(\sum_q \mathbf{w}_{p,q}^{(l-1)} \kappa_q^{(l-1)}(\cdot, \cdot)\right), \quad (1)$$

106 with $\{\mathbf{w}_{p,q}^{(l-1)}\}$ being weights connecting units at layers l and $l - 1$; see the
 107 blue dashed area in Fig. 1(left). This feed-forward kernel evaluation is achieved
 108 layer-wise till reaching the final output kernel. In this recursive definition, other
 109 activation functions g can be chosen (particularly for the intermediate layers)
 110 including the hyperbolic making the learning numerically more stable while also
 111 preserving the p.s.d of the final output kernel.

112 For a given classification task, the weights $\{\mathbf{w}_{p,q}^{(l-1)}\}$ are trained discrimina-
 113 tively [23, 24] using a max margin SVM criterion which aims at minimizing a
 114 regularized hinge loss on top of the learned DKN. This results into an SVM
 115 optimization problem which is solved in its dual form by backpropagating the
 116 gradient of that form w.r.t. the output kernel using the chain rule [19]. Then,
 117 the weights connecting layers in the DKN are updated using gradient descent.
 118 Variants of this optimization criterion, leveraging both labeled and unlabeled
 119 data (following a semi-supervised and laplacian setting) makes it possible to
 120 train better DKN as detailed in [24].

121 3. Deep map networks

122 In this section, we introduce a novel method that finds for any given DKN,
123 its associated DMN; the proposed method proceeds layer-wise by finding ex-
124 plicit maps that best fit the original kernels in the DKN. As shown later in
125 experiments, this process delivers highly efficient DMNs, while being compara-
126 bly accurate w.r.t. their underlying DKNs. Later in Section 4, we introduce an
127 extension that further enhances the approximation quality of our DMN; starting
128 from the initial weights of the DMN, we update these weights by minimizing
129 the difference between inner products of the maps in the DMN and the original
130 kernels in the DKN. The strength of this extension also resides in its unsuper-
131 vised setting which makes it possible to learn from abundant unlabeled sets.

132 Considering all the elementary (input) kernels in the DKN as positive semi-
133 definite and resulting from the closure of the p.s.d w.r.t. different operations (in-
134 cluding sum, product, exponential and hyperbolic activation functions), all the
135 intermediate and output kernels $\{\kappa_p^{(l)}\}_{l,p}$ will also be p.s.d. Each $\kappa_p^{(l)}(\mathbf{x}, \mathbf{x}')$ can
136 therefore be written as an inner product of kernel maps as $\langle \phi_p^l(\mathbf{x}), \phi_p^l(\mathbf{x}') \rangle$, with
137 $\phi_p^l : \mathcal{X} \rightarrow \mathcal{H}$ being a mapping from the input space \mathcal{X} to a high dimensional space
138 \mathcal{H} . As the explicit form of ϕ_p^l is not necessarily explicit (known), our goal is to de-
139 sign an approximated mapping $\hat{\phi}_p^l$ that guarantees $\kappa_p^{(l)}(\mathbf{x}, \mathbf{x}') \simeq \langle \hat{\phi}_p^l(\mathbf{x}), \hat{\phi}_p^l(\mathbf{x}') \rangle$.
140 When these approximated mappings through different layers are known, the
141 resulting DMN provides deep kernel representations from the input data.

142 3.1. Input layer maps

143 In order to fully benefit from DMNs, the maps of the elementary kernels,
144 that feed these DMNs, should be explicitly known. As discussed earlier, dif-
145 ferent kernels have different maps; for linear and polynomial, their maps are
146 straightforward and can be easily defined. However, for other more powerful
147 and discriminating kernels, such as the Gaussian and the histogram intersection
148 (HI), their maps are either infinite dimensional or unknown. In this subsection,
149 the definitions of exact and approximate explicit maps are shown for different
150 kernels (including polynomial and HI).

Exact polynomial kernel map. A polynomial kernel defined as $\kappa_p^{(1)}(\mathbf{x}, \mathbf{x}') = \langle \mathbf{x}, \mathbf{x}' \rangle^{n+1}$ can be expressed as $\kappa_p^{(1)}(\mathbf{x}, \mathbf{x}') = \langle \mathbf{x} \otimes^n \mathbf{x}, \mathbf{x}' \otimes^n \mathbf{x}' \rangle$, with \otimes^n standing for the Kronecker tensor product applied n times; this tensor product on two matrices A (of size $m \times n$) and B (of size $p \times q$) results into a block matrix (of size $mp \times nq$) as:

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \dots & a_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & \dots & a_{mn}\mathbf{B} \end{bmatrix}. \quad (2)$$

151 Hence, it is easy to see that the exact explicit map for a polynomial kernel is
 152 $\hat{\phi}_p^{(1)}(\mathbf{x}) = \mathbf{x} \otimes^n \mathbf{x}$.

Approximate HI kernel map. The approximate explicit maps of HI can be obtained using vector quantization. Given two vectors \mathbf{x} and \mathbf{x}' of dimension s , the HI on \mathbf{x}, \mathbf{x}' is defined as $\kappa_p^{(1)}(\mathbf{x}, \mathbf{x}') = \sum_{d=1}^s \min(\mathbf{x}^d, \mathbf{x}'^d)$ (with \mathbf{x}^d being the value of d^{th} dimension of \mathbf{x}). Considering $\mathbf{x} = (\mathbf{x}^1, \dots, \mathbf{x}^s)^\top \in \mathcal{X}$, each dimension \mathbf{x}^d of \mathbf{x} is mapped to

$$\psi(\mathbf{x}^d) = 2^0 + 2^1 + \dots + 2^{k(\mathbf{x}^d)}, \quad (3)$$

153 where $k(\mathbf{x}^d) = \left\lfloor Q \frac{\mathbf{x}^d - \ell_d}{u_d - \ell_d} \right\rfloor$ and $\lfloor z \rfloor$ stands for the largest integer not greater
 154 than $z \in \mathbb{R}$, $Q \in \mathbb{N}^+$ is a predefined quantization, $\ell_d = \min_{\mathbf{x}} \{\mathbf{x}^d : \mathbf{x} \in \mathcal{X}\}$ and
 155 $u_d = \max_{\mathbf{x}} \{\mathbf{x}^d : \mathbf{x} \in \mathcal{X}\}$. In the above definition, $\psi(\cdot)$ is a ‘‘decimal-to-unary’’
 156 map; for instance, 1 is mapped to 1, 2 is mapped to 11, 3 to 111, and so on. In
 157 the following, $\psi(\mathbf{x}^d)$ is rewritten as a vector of Q dimensions, and its first $k(\mathbf{x}^d)$
 158 dimensions are set to 1 and the remaining are set to 0 [40].

Proposition 1. *Given any \mathbf{x}, \mathbf{x}' in \mathcal{X} , for sufficiently large Q , the inner product $\langle \hat{\phi}_p^{(1)}(\mathbf{x}), \hat{\phi}_p^{(1)}(\mathbf{x}') \rangle$ approximates the histogram intersection kernel $\kappa_p^{(1)}(\mathbf{x}, \mathbf{x}')$, where*

$$\hat{\phi}_p^{(1)}(\mathbf{x}) = \left(\psi(\mathbf{x}^1)^\top \sqrt{\frac{u_1 - \ell_1}{Q}}, \sqrt{\ell_1}, \dots, \psi(\mathbf{x}^s)^\top \sqrt{\frac{u_s - \ell_s}{Q}}, \sqrt{\ell_s} \right)^\top \quad (4)$$

159 *is the approximate kernel map and $\psi(\mathbf{x}^d)^\top$ stands for the transpose of $\psi(\mathbf{x}^d)$.*

160 *Proof.* The proof is given in the Appendix A. □

161 **Approximate Gaussian kernel map.** As the explicit map of the Gaussian
 162 kernel is infinite dimensional, we consider instead an approximate explicit map
 163 of that kernel using eigen decomposition (ED) as shown in Eqs. (6), (5) with
 164 $l = 1$ (see Section 3.2). This ED is not restricted to the Gaussian kernel and
 165 can also be extended to other kernels whose exact explicit maps are difficult to
 166 obtain.

167 3.2. Intermediate/output layer maps

168 Given the explicit map of each elementary kernel at the input layer, our goal is
 169 to design the maps of the subsequent layers. Since the map of each layer depends
 170 on its preceding layers, this goal is achieved layer-wise using a greedy process.
 171 As intermediate/output kernels in the DKN are defined as linear combinations
 172 of kernels in the preceding (input or intermediate) layers followed by nonlinear
 173 activations, we mainly focus on how to approximate the maps of these activation
 174 functions in the DMN; in this section, we assume that weights $\{\mathbf{w}_{p,q}^{(l)}\}$ connecting
 175 different layers are already known resulting from the initial setting of the DKN
 176 (see again Section 2).

Proposition 2. *Let $\mathcal{S} = \{\mathbf{x}_i\}_{i=1}^N$ be a subset of N samples of \mathcal{X} , and let \mathbf{K}_p^l
 be a gram-matrix whose entries are defined on \mathcal{S} . Let $\mathbf{U}_p^{(l)} = \boldsymbol{\alpha}\boldsymbol{\Lambda}^{-1/2}$ with $\boldsymbol{\alpha}$,
 $\boldsymbol{\Lambda}$ being respectively the matrices of eigenvectors and eigenvalues obtained by
 solving*

$$\mathbf{K}_p^l \boldsymbol{\alpha} = \boldsymbol{\alpha} \boldsymbol{\Lambda}. \quad (5)$$

Considering $\|\cdot\|_2$ as the ℓ_2 (matrix) norm and $\hat{\mathbf{K}}_p^l$ as the gram-matrix associated
 to $\{\langle \hat{\phi}_p^{(l)}(\mathbf{x}), \hat{\phi}_p^{(l)}(\mathbf{x}') \rangle\}_{\mathbf{x}, \mathbf{x}' \in \mathcal{S}}$ with

$$\hat{\phi}_p^{(l)}(\mathbf{x})^\top = \left(g(\langle \hat{\phi}_p^{l,c}(\mathbf{x}), \hat{\phi}_p^{l,c}(\mathbf{x}_1) \rangle) \dots g(\langle \hat{\phi}_p^{l,c}(\mathbf{x}), \hat{\phi}_p^{l,c}(\mathbf{x}_N) \rangle) \right) \mathbf{U}_p^{(l)} \quad (6)$$

and

$$\hat{\phi}_p^{l,c}(\mathbf{x}) = \left(\sqrt{\mathbf{w}_{p,1}^{(l-1)}} \hat{\phi}_1^{(l-1)}(\mathbf{x})^\top \dots \sqrt{\mathbf{w}_{p,n_{l-1}}^{(l-1)}} \hat{\phi}_{n_{l-1}}^{(l-1)}(\mathbf{x})^\top \right)^\top, \quad (7)$$

then the following property is satisfied

$$\|\hat{\mathbf{K}}_p^l - \mathbf{K}_p^l\|_2 = 0. \quad (8)$$

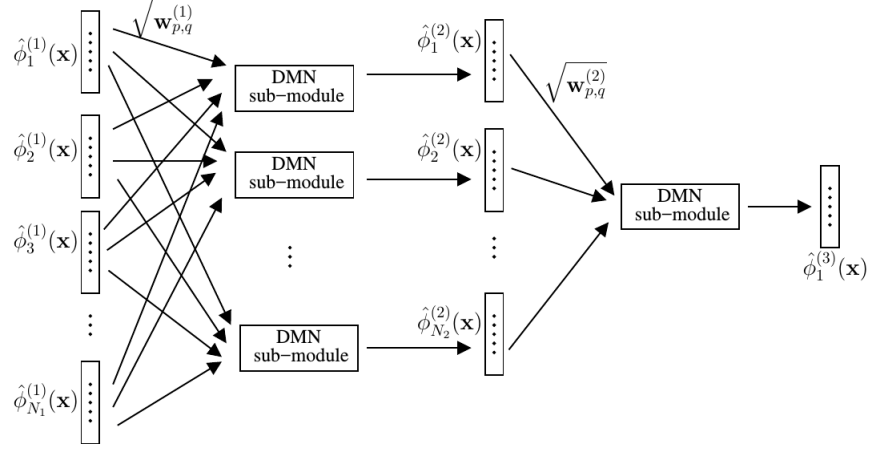


Figure 2: The flowchart of DMN for a three-layer DKN, as shown in Fig. 1(left).

177 *Proof.* The proof is given in the Appendix B. □

178 Note that for any samples \mathbf{x}, \mathbf{x}' taken out of \mathcal{S} (but with similar distribution
 179 as \mathcal{S}), it is clear (as also observed in our experiments) that $|\langle \hat{\phi}_p^{(l)}(\mathbf{x}), \hat{\phi}_p^{(l)}(\mathbf{x}') \rangle -$
 180 $\kappa_p^{(l)}(\mathbf{x}, \mathbf{x}')| \rightsquigarrow 0$ as N and the number of eigenvectors used in $\{\mathbf{U}_p^{(l)}\}$ increase.

181 3.3. Network design

182 We incrementally expand each layer l in the DKN into three sub-layers in the
 183 underlying DMN in order to design the map $\hat{\phi}_p^{(l)}$. The first sub-layer provides the
 184 products between weights $\{(\mathbf{w}_{p,q}^{(l-1)})^{1/2}\}_q$ and the preceding maps $\{\hat{\phi}_q^{(l-1)}(\mathbf{x})\}_q$
 185 resulting into the intermediate map $\hat{\phi}_p^{l,c}(\mathbf{x})$ as shown in Eq. (7). Afterwards, we
 186 feed this map $\hat{\phi}_p^{l,c}(\mathbf{x})$ to Eq. (6) in two steps: (i) in the second sub-layer, inner
 187 products are achieved between $\hat{\phi}_p^{l,c}(\mathbf{x})$ and parameters $\{\hat{\phi}_p^{l,c}(\mathbf{x}_i)\}_{i=1}^N$ followed by
 188 the activations $\{g(\cdot)\}_{i=1}^N$ (with g being the hyperbolic excepting the final layer
 189 in the DKN which uses the exponential); (ii) in the third sub-layer, the explicit
 190 map $\hat{\phi}_p^{(l)}$ is obtained as the product of $\{g(\cdot)\}_{i=1}^N$ and weights $\mathbf{U}_p^{(l)}$. Fig. 1(right)
 191 shows a sub-module of DMN with sub-layers and Fig. 2 provides the flowchart
 192 of DMN for a three-layer DKN. Similarly, all the subsequent layers in the DMN

193 are designed by processing the DKN layer-wise.⁴

194 4. Enhancing DMN Parameters

195 So far the design principle of our method (shown in Section 3.3 and Fig. 1)
 196 seeks to find explicit maps whose inner products approximate the original ker-
 197 nel values. This is achieved by expanding each layer in the DKN into three
 198 sub-layers in the DMN with parameters fixed to $\{\hat{\phi}_p^{l,c}(\mathbf{x}_i)\}_i$ and $\mathbf{U}_p^{(l)}$. In spite
 199 of being efficient and also effective w.r.t. the DKN (see experiments), the re-
 200 sulting DMN can be further improved when re-training and fine-tuning these
 201 parameters as shown subsequently.

The purpose of the proposed unsupervised algorithm is to further reduce
 the approximation error between the kernel values from DKN and the inner
 product of kernel maps from DMN. Let $\mathcal{S}' \subset \mathcal{X}$ be a subset drawn from the
 same distribution as \mathcal{S} and define \mathcal{P} as a subset of pairs taken from $\mathcal{S}' \times \mathcal{S}'$. Our
 goal is to optimize maps of DMN using the following unsupervised criterion

$$E = \sum_{(\mathbf{x}, \mathbf{x}') \in \mathcal{P}} \frac{1}{2} \|\hat{\phi}_1^{(L)}(\mathbf{x})^\top \hat{\phi}_1^{(L)}(\mathbf{x}') - \kappa_1^{(L)}(\mathbf{x}, \mathbf{x}')\|^2, \quad (9)$$

202 where $\kappa_1^{(L)}(\mathbf{x}, \mathbf{x}')$ corresponds to the kernel value obtained using the DKN and
 203 $\hat{\phi}_1^{(L)}(\mathbf{x}), \hat{\phi}_1^{(L)}(\mathbf{x}')$ are the underlying (unknown) kernel maps; initially, only
 204 $\{\hat{\phi}_p^{(1)}(\mathbf{x}), \hat{\phi}_p^{(1)}(\mathbf{x}')\}_{(\mathbf{x}, \mathbf{x}') \in \mathcal{P}}$ are known according to the procedure shown in Sec-
 205 tion 3.1.

206 Considering the initial setting of DMN parameters (i.e., $\{\hat{\phi}_p^{l,c}(\mathbf{x}_i)\}_i$ and $\mathbf{U}_p^{(l)}$),
 207 the learning process of this DMN relies on backpropagation [19]. The latter finds
 208 the best parameters by minimizing the objective function (E) following an “end-
 209 to-end” framework where the gradients of E are given using the chain rule; we
 210 firstly compute the gradients of the loss function E w.r.t. final kernel maps,

⁴As the goal, in this paper, is to build approximate deep kernel maps for a given (fixed) deep kernel network, the weights \mathbf{w} between different layers remain fixed (as shown in Eq. (7)). However, they can also be jointly learned using gradient descent, but this is out of the main scope of this paper.

211 then we backpropagate them through the DMN in order to obtain the gradients
 212 w.r.t. the parameters of DMN, finally we average them over training pairs to
 213 obtain the descent direction and update DMN parameters.

Starting from the derivative of E w.r.t. $\hat{\phi}_1^{(L)}(\mathbf{x})$

$$\frac{\partial E}{\hat{\phi}_1^{(L)}(\mathbf{x})} = (\hat{\phi}_1^{(L)}(\mathbf{x})^\top \hat{\phi}_1^{(L)}(\mathbf{x}') - \kappa_1^{(L)}(\mathbf{x}, \mathbf{x}')) \hat{\phi}_1^{(L)}(\mathbf{x}'), \quad (10)$$

we obtain the gradients w.r.t. different layers $l = L, \dots, 1$ and units $p = 1, \dots, n_l$. As the construction of DMN is achieved layer-wise (see again Section 3.3), we show below the backpropagation procedure for a module (shown in Fig. 1, right). Given the derivatives of E w.r.t. $\hat{\phi}_p^{(l)}(\mathbf{x})$ in layer l , we evaluate the derivatives w.r.t. $\hat{\phi}_q^{(l-1)}(\mathbf{x})$ in layer $(l-1)$. The derivative w.r.t. $\hat{\phi}_p^{(l)}(\mathbf{x})$ is backpropagated to $\kappa_p^{(l)}$ in Eq. (6) by

$$\frac{\partial E}{\partial \kappa_p^{(l)}(\mathbf{x}, \mathbf{x}_i)} = \left(\frac{\partial E}{\partial \hat{\phi}_p^{(l)}(\mathbf{x})} \right)^\top [\mathbf{U}_p^{(l)}]_i^\top, \quad (11)$$

here $[\cdot]_i$ stands for the i -th row of a matrix. Considering $\kappa_p^{(l)}(\mathbf{x}, \mathbf{x}_i) = g(f_p^{(l)}(\mathbf{x}, \mathbf{x}_i))$, with $f_p^{(l)}(\mathbf{x}, \mathbf{x}_i) = \langle \hat{\phi}_p^{l,c}(\mathbf{x}), \hat{\phi}_p^{l,c}(\mathbf{x}_i) \rangle$, we obtain

$$\frac{\partial E}{\partial f_p^{(l)}(\mathbf{x}, \mathbf{x}_i)} = g'(f_p^{(l)}(\mathbf{x}, \mathbf{x}_i)) \frac{\partial E}{\partial \kappa_p^{(l)}(\mathbf{x}, \mathbf{x}_i)}, \quad (12)$$

where $g'(\cdot)$ is the derivative of the nonlinear activation function; for instance, $g'(\cdot) = 1 - \tanh(\cdot)^2$ for the tangent hyperbolic and $g'(\cdot) = g(\cdot)$ for the exponential. By accumulating the derivatives from each term $f_p^{(l)}(\mathbf{x}, \mathbf{x}_i)$, we obtain

$$\frac{\partial E}{\hat{\phi}_p^{l,c}(\mathbf{x})} = \sum_{i=1}^N \hat{\phi}_p^{l,c}(\mathbf{x}_i) \frac{\partial E}{\partial f_p^{(l)}(\mathbf{x}, \mathbf{x}_i)}, \quad (13)$$

Finally, we get the derivatives w.r.t. $\hat{\phi}_q^{(l-1)}(\mathbf{x})$ for layer $(l-1)$ in Eq. (7) by

$$\frac{\partial E}{\partial \hat{\phi}_q^{(l-1)}(\mathbf{x})} = \sqrt{\mathbf{w}_{p,q}^{(l-1)}} \text{Frag}\left(\frac{\partial E}{\hat{\phi}_p^{l,c}(\mathbf{x})}\right)_q, \quad (14)$$

214 where $\text{Frag}\left(\frac{\partial E}{\hat{\phi}_p^{l,c}(\mathbf{x})}\right)_q$ stands for the fragment of derivatives corresponding to the
 215 kernel maps of the unit q at layer $(l-1)$ in the DKN.

The gradients of the loss function E w.r.t. $\mathbf{U}_p^{(l)}$ and $\hat{\phi}_p^{l,c}(\mathbf{x}_i)$ are then given as

$$\Delta \mathbf{U}_p^{(l)} = (\kappa_p^{(l)}(\mathbf{x}, \mathbf{x}_1) \dots \kappa_p^{(l)}(\mathbf{x}, \mathbf{x}_N))^\top \left(\frac{\partial E}{\partial \hat{\phi}_p^{(l)}(\mathbf{x})} \right)^\top \quad (15)$$

$$\Delta \hat{\phi}_p^{l,c}(\mathbf{x}_i) = \frac{\partial E}{\partial f_p^{(l)}(\mathbf{x}, \mathbf{x}_i)} \hat{\phi}_p^{l,c}(\mathbf{x}). \quad (16)$$

216 Error backpropagation is achieved layer-wise from the final to the input layer;
 217 the increments of $\{\hat{\phi}_p^{l,c}(\mathbf{x}_i)\}_{i=1}^N$ and $\mathbf{U}_p^{(l)}$ are obtained by Eq. (16) and Eq. (15).
 218 Gradient descent with a step η (see experiments) is performed to update the
 219 parameters of DMN. The whole learning procedure is shown in Algorithm 1.

220 As described earlier, an initial DMN is firstly set using the training set \mathcal{S} ,
 221 then sample pairs in \mathcal{P} are randomly selected from \mathcal{S}' to further enhance the
 222 parameters of the new (fine-tuned) DMN. As a result, the fine-tuned DMN en-
 223 ables us to obtain a better approximation of the original DKN on large datasets
 224 while being highly efficient as shown through the following experiments in image
 225 annotation.

226 5. Experiments

227 In this section, we compare the performance of the proposed DMN w.r.t. its
 228 underlying DKN in three aspects: i) discrimination power, ii) relative approxi-
 229 mation error between DMN and DKN and iii) also efficiency. The targeted task
 230 is image annotation; given a picture, the goal is to predict a list of keywords that
 231 best describes the visual content of that image. We consider two challenging
 232 and widely used annotation benchmarks: ImageCLEF [38], COREL5k [39], and
 233 also Banana (see details below). For three sets, we learn – highly competitive –
 234 3-layer DKNs using the setting in [24] and we plug these DKNs into SVMs in
 235 order to achieve classification and annotation.

The discrimination power of the learned DMN and DKN networks is measured following the protocol defined by challenge organizers and data providers (see [38] for ImageCLEF and [39] for COREL5k; see also extra details below).

Input: Fixed $\{\mathbf{w}_{p,q}^{(l-1)}\}$ ($l = 2, \dots, L$),

A set of sample pairs \mathcal{P} ,

Kernel maps $\{\hat{\phi}_p^{(1)}(\mathbf{x})\}_{\mathbf{x} \in \mathcal{S}'}$ at the input layer,

Output kernel values $\{\kappa_1^{(L)}(\mathbf{x}, \mathbf{x}')\}_{(\mathbf{x}, \mathbf{x}') \in \mathcal{P}}$.

Initialization: $\{\hat{\phi}_p^{l,c}(\mathbf{x}_i)\}_{i=1}^N$ and $\{\mathbf{U}_p^{(l)}\}$, $p = \{1, \dots, n_l\}$, learning rate η .

Output: Optimal (updated) $\{\hat{\phi}_p^{l,c}(\mathbf{x}_i)\}_{i=1}^N$ and $\{\mathbf{U}_p^{(l)}\}$.

repeat

for each pair $(\mathbf{x}, \mathbf{x}') \in \mathcal{P}$ **do**

Forward $(\hat{\phi}_p^{(1)}(\mathbf{x}), \hat{\phi}_p^{(1)}(\mathbf{x}'))$ through DMN to obtain

$(\hat{\phi}_1^{(L)}(\mathbf{x}), \hat{\phi}_1^{(L)}(\mathbf{x}'))$ by Eqs. (7), (6);

Compute the loss by Eq. (9);

Compute the gradients $\frac{\partial E}{\partial \hat{\phi}_1^{(L)}(\mathbf{x})}$ by Eq. (10);

for $l = L : 2$ **do**

Backward the gradients $\frac{\partial E}{\partial \hat{\phi}_1^{(L)}(\mathbf{x})}$ by Eqs. (11)-(14);

Compute $(\Delta U_p^{(l)})_{\mathbf{x}}$ and $(\Delta \hat{\phi}_p^{l,c}(\mathbf{x}_i))_{\mathbf{x}}$ by Eq. (15) and (16);

Compute the gradients from $\hat{\phi}_p^{(L)}(\mathbf{x}')$: $(\Delta U_p^{(l)})_{\mathbf{x}'}$ and

$(\Delta \hat{\phi}_p^{l,c}(\mathbf{x}_i))_{\mathbf{x}'}$;

Average both gradients: $\Delta U_p^{(l)} \leftarrow \frac{1}{2}((\Delta U_p^{(l)})_{\mathbf{x}} + (\Delta U_p^{(l)})_{\mathbf{x}'})$;

$\Delta \hat{\phi}_p^{l,c}(\mathbf{x}_i) \leftarrow \frac{1}{2}((\Delta \hat{\phi}_p^{l,c}(\mathbf{x}_i))_{\mathbf{x}} + (\Delta \hat{\phi}_p^{l,c}(\mathbf{x}_i))_{\mathbf{x}'})$;

Update these parameters by gradient descents;

$U_p^{(l)} \leftarrow U_p^{(l)} - \eta \Delta U_p^{(l)}$;

$\hat{\phi}_p^{l,c}(\mathbf{x}_i) \leftarrow \hat{\phi}_p^{l,c}(\mathbf{x}_i) - \eta \Delta \hat{\phi}_p^{l,c}(\mathbf{x}_i)$;

end

end

until Convergence;

Algorithm 1: Unsupervised DMN learning algorithm

The relative approximation error (RE) of a given DMN w.r.t. its underlying DKN is measured (on a given set $\mathcal{T} \subset \mathcal{X}$) as

$$\text{RE} = \frac{1}{|\mathcal{T}|^2} \sum_{\mathbf{x}, \mathbf{x}' \in \mathcal{T}} \frac{|\langle \hat{\phi}_1^{(3)}(\mathbf{x}), \hat{\phi}_1^{(3)}(\mathbf{x}') \rangle - \kappa_1^{(3)}(\mathbf{x}, \mathbf{x}')|}{|\langle \hat{\phi}_1^{(3)}(\mathbf{x}), \hat{\phi}_1^{(3)}(\mathbf{x}') \rangle| + |\kappa_1^{(3)}(\mathbf{x}, \mathbf{x}')|} \times 100\%, \quad (17)$$

236 In the remainder of this section, we show different evaluation measures (discrim-
 237 ination power, RE and efficiency) on ImageCLEF and COREL5k benchmarks;
 238 note that efficiency was measured on a Mac OS with Intel Core i5 processors.

239 5.1. ImageCLEF benchmark

240 The ImageCLEF Photo Annotation benchmark [38] includes more than 250k
 241 (training, dev and test) images belonging to 95 different concepts. As ground
 242 truth is available (released) only on the dev set (with 1,000 images), we learn
 243 DKNs and SVMs [24] using only the dev set; the latter is split into two sub-
 244 sets: the first one used for DKN+SVM training while the other one for SVM
 245 testing. Given a concept and a test image, the decision about whether that
 246 concept is present in that test image depends on the score of a classifier; the
 247 latter corresponds to a “one-versus-all” SVM that returns a positive score if
 248 the concept is present in the test image and a negative score otherwise. We
 249 employ the LIBSVM library [41] in order to train each SVM independently, and
 250 we obtain the optimal trade-off parameter C_k ($\in [2^{-10}, \dots, 2^{10}]$) using 3-fold
 251 cross-validation on the training set. The discrimination power of DKN and
 252 DMN (when combined with SVMs) is evaluated using the F-measure (defined
 253 as harmonic means of recalls and precisions) both at the concept and the image
 254 levels (resp. denoted MF-C and MF-S) as well as the Mean Average Precision
 255 (MAP) [38]; high values of these measures imply better performances.

256 In order to feed the inputs of DKN, we consider a combination of 10 visual
 257 features (provided by the ImageCLEF challenge organizers) and 4 elementary
 258 kernels (i.e. linear, polynomial with 2 orders, Gaussian⁵ and histogram inter-
 259 section) and we train a three-layer DKN with 40 input and 80 hidden units in

⁵with a scale hyper-parameter set to be average Euclidean distance between data samples and their neighbors.

260 a supervised way following the scheme in [24]; the only difference w.r.t. [24]
 261 resides in the hyperbolic tangent activation function which is used to provide a
 262 better numerical stability and convergence when training DKN.

263 **Initial DMNs.** Assuming the weights $\{\mathbf{w}_{p,q}^{(l-1)}\}$ of three-layer DKN known, we
 264 build its equivalent DMN (referred to as initial DMN) as shown in Section 3.
 265 In these experiments, we consider two random samplings of the subset \mathcal{S} – from
 266 the dev set with $|\mathcal{S}| = 500$ and $|\mathcal{S}| = 1000$ – in order to build the initial DMN
 267 (see Section 3 and Eqs. (7), (6)). According to Table 1, we observe that the
 268 performance of the initial DMN – with $|\mathcal{S}| = 500$ – slightly degrades compared
 269 to its underlying DKN; indeed, MF-S and MF-C decrease by 1.3 and 2.6 pts
 270 respectively while MAP decreases by 6.0 pts. With $|\mathcal{S}| = 1000$ performances of
 271 the initial DMN is clearly improved compared to the one with $|\mathcal{S}| = 500$; we
 272 obtain a slight gain in MF-S and comparable performance in MF-C. We also
 273 provide a comparison of the discrimination power of initial DMN against shallow
 274 DKN (i.e two-layer DKN) using a supervised setting; Table 1 clearly shows the
 275 superiority of initial DMN (when $|\mathcal{S}| = 1000$). The relative approximation
 276 error (RE) of the two initial DMNs (i.e., with $|\mathcal{S}| = 500$ and $|\mathcal{S}| = 1000$) are
 277 also shown in Table 2; we evaluate these REs on \mathcal{T} with a cardinality ranging
 278 from 2,000 to 10,000 samples. From these results, we observe that REs are
 279 comparably low on small sets; indeed, with $|\mathcal{T}| = 2,000$, the obtained REs are
 280 equal to 0.94% when $|\mathcal{S}| = 500$ and 0.95% when $|\mathcal{S}| = 1000$. Higher REs are
 281 obtained on larger \mathcal{T} and this clearly motivates the importance of fine-tuning
 282 in order to make REs (and thereby performances) of the learned DMN stable
 283 (and close to the underlying DKN).

284 **Fine-tuned DMNs.** In order to fine-tune the parameters of DMN, we use
 285 the learning procedure presented in Section 4. We consider an unlabeled set
 286 \mathcal{S}' (with $|\mathcal{S}'|$ ranging from 1,000 to 4,000) and we randomly sample 100,000
 287 pairs from $\mathcal{S}' \times \mathcal{S}'$ to minimize criterion (9) using gradient descent with a step-
 288 size empirically set to 10^{-6} , a mini-batch size equal to 200 and a max number
 289 of iterations set to 5,000. Fig. 3 shows the evolution of the approximation
 290 loss (9) as the learning process iterates; we clearly observe that 100,000 pairs

Framework	MF-S	MF-C	MAP
2-layer DKN	44.96	25.77	53.95
3-layer DKN	46.23	30.00	55.73
Initial DMN ($ \mathcal{S} = 500$)	44.92	27.39	49.75
Fine-tuned DMN ($ \mathcal{S}' = 2000$)	45.05	27.51	49.80
Fine-tuned DMN ($ \mathcal{S}' = 3000$)	44.94	27.40	49.80
Fine-tuned DMN ($ \mathcal{S}' = 4000$)	45.06	27.44	49.79
Initial DMN ($ \mathcal{S} = 1000$)	47.73	29.40	53.15
Fine-tuned DMN ($ \mathcal{S}' = 2000$)	47.79	29.68	52.89
Fine-tuned DMN ($ \mathcal{S}' = 3000$)	47.95	29.80	53.32
Fine-tuned DMN ($ \mathcal{S}' = 4000$)	47.70	29.30	53.33

Table 1: The discrimination power (in %) of different DMNs w.r.t. the underlying DKN; in these experiments, two initial DMNs are designed using 500 and 1000 samples.

Configuration	$ \mathcal{S}' $	2K	3K	4K	5K	6K	7K	8K	9K	10K
Initial DMN ($ \mathcal{S} = 500$)	-	0.94	1.25	1.41	1.51	1.58	1.62	1.66	1.69	1.71
	500	0.89	1.19	1.35	1.45	1.52	1.57	1.60	1.63	1.65
	1000	0.89	1.20	1.36	1.46	1.53	1.58	1.61	1.64	1.66
Fine-tuned DMN	2000	0.42	0.46	0.50	0.52	0.54	0.56	0.57	0.58	0.59
	3000	0.52	0.47	0.47	0.47	0.47	0.47	0.48	0.48	0.48
	4000	0.60	0.51	0.49	0.47	0.47	0.46	0.46	0.46	0.46
Initial DMN ($ \mathcal{S} = 1000$)	-	0.95	1.27	1.44	1.54	1.62	1.67	1.70	1.74	1.76
	1000	0.89	1.21	1.38	1.48	1.55	1.60	1.64	1.67	1.69
	2000	0.37	0.41	0.44	0.46	0.48	0.49	0.50	0.51	0.52
Fine-tuned DMN	3000	0.46	0.43	0.43	0.43	0.44	0.44	0.44	0.45	0.45
	4000	0.54	0.48	0.46	0.45	0.45	0.44	0.44	0.44	0.44

Table 2: Relative errors of initial and fine-tuned DMNs w.r.t. the DKN for different dataset cardinalities $|\mathcal{T}|$ (ranging from 2K to 10K) and when two different initializations are employed.

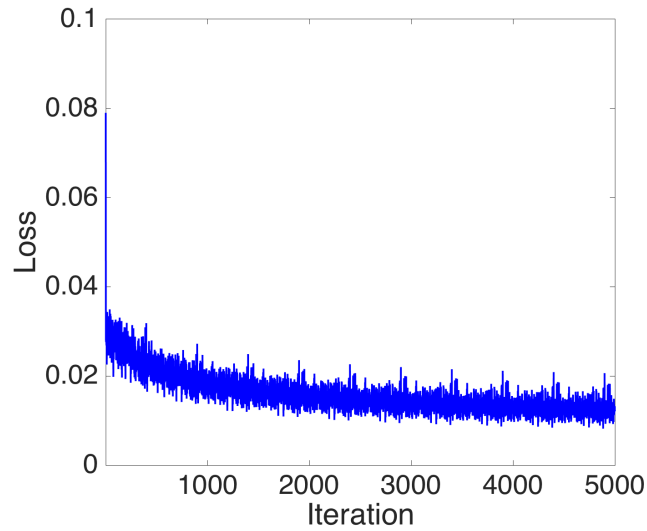


Figure 3: This figure shows the loss criterion in Eq. (9) as the learning iterates when $|\mathcal{S}| = 500$ and $|\mathcal{S}'| = 2000$.

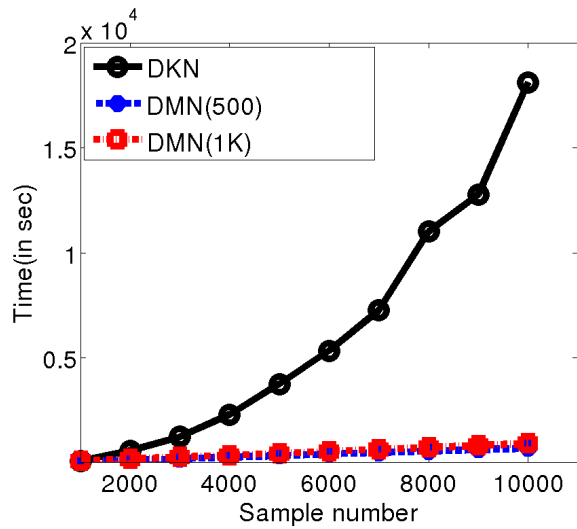


Figure 4: This figure shows a comparison of processing time between two different DMNs and their underlying DKN as $|\mathcal{T}|$ increases (with $|\mathcal{S}| = 500$ and $|\mathcal{S}| = 1000$) on ImageCLEF dataset.

	$ \mathcal{T} $	50K	100K
3-layer DKN	Time	40.4 hrs	160.3 hrs
Fine-tuned DMN	Time	1.1 hrs	2.4 hrs
$ \mathcal{S} = 500, \mathcal{S}' = 4000$	RE	0.46%	0.46%
Fine-tuned DMN	Time	1.3 hrs	2.8 hrs
$ \mathcal{S} = 1000, \mathcal{S}' = 4000$	RE	0.45%	0.45%

Table 3: This table shows a comparison of processing time and relative errors between the DKN and the fine-tuned DMN on 50K and 100K images of ImageCLEF. “hrs” stands for “hours”.

291 are already sufficient to train accurate DMNs. With an extra fine tuning step
 292 (shown subsequently), the accuracy of these DMNs is further improved.

293 As shown in Table 1, we observe that the discrimination power of different
 294 DMNs remains stable (with a slight gain in MF-S when $|\mathcal{S}'| = 3000$) w.r.t. their
 295 underlying DKNs, and this naturally follows the noticeably small REs of the
 296 fine-tuned DMNs (see Table 2). The latter are further positively impacted when
 297 $|\mathcal{S}'|$ becomes larger; for instance, when increasing $|\mathcal{S}'|$ from 1,000 to 4,000, the
 298 RE decreases significantly (particularly when $|\mathcal{T}| = 10,000$). Moreover, and in
 299 contrast to the initial DMNs, the fine-tuned DMNs are less sensitive to $|\mathcal{T}|$ as
 300 shown through the observed REs which remain stable w.r.t. $|\mathcal{T}|$.

301 Finally, we measure the gain in efficiency obtained with DMNs against
 302 DKNs. From Fig. 4, we observe that DMN is (at least) an order of magnitude
 303 faster compared to its DKN; for instance, with 10,000 samples, DKN requires
 304 more than 15,000 seconds in order to compute kernel values while DMN requires
 305 less than 1,000 seconds. Table 3 also provides a comparison of efficiency and
 306 RE on much larger sets (resp. 50K and 100K) randomly sampled from the (un-
 307 labeled) training set of ImageCLEF; a significant improvement in efficiency is
 308 observed. In other words, the complexity of evaluating DMNs is linear while
 309 for DKN it is quadratic. These results clearly corroborate the fact that the
 310 proposed DMNs are as effective as DKNs while being highly efficient especially

311 on large scale datasets.

312 5.2. COREL5k benchmark

313 The COREL5k database introduced in [39] is another benchmark which is
314 widely used for image annotation. In this database, 4,999 images are collected
315 and a vocabulary of 200 keywords is used for annotation. This set is split into
316 two parts; the first one includes 4,500 images for training and the second one
317 499 images for testing. As for ImageCLEF, the task is again to assign a list of
318 keywords for each image in the test set.

319 Each image in COREL5k is described using 15 types of INRIA features
320 [42] including: GIST features, 6 color histograms for RGB, HSV, LAB in two
321 spatial layouts, 8 bag-of-features based on SIFT and robust hue descriptors
322 in two spatial layouts. Following the standard protocol defined on COREL5k
323 [39], each test image is annotated with up to 5 keywords and performances
324 (discrimination power of image classification/annotation) are measured by the
325 mean precision and recall over keywords (referred to as \mathbf{P} and \mathbf{R} respectively)
326 as well as the number of keywords with non-zero recall value (denoted \mathbf{N}_+);
327 again, higher values of these measures imply better performances.

328 As in ImageCLEF (see section 5.1), we use 4 elementary kernels for each
329 feature: linear, order two polynomial, RBF (with a scale parameter set to the
330 average distance between data) and histogram intersection; in total, we use
331 60 different elementary kernels as inputs to the 3-layer DKN. We also use the
332 same DKN architecture on COREL5k with a slight difference in the number of
333 units in the hidden layers (equal to 120 instead of 80 in ImageCLEF). Again,
334 the weights of DKN are learned using the semi-supervised learning procedure
335 presented in [24] where the similarity between images \mathbf{x} and \mathbf{x}' is computed by
336 the heat kernel $S(\mathbf{x}, \mathbf{x}') = \exp \frac{-\|\mathbf{x}-\mathbf{x}'\|^2}{4t}$, where t is the mean distance between
337 data samples and their neighbors. An ensemble of “one-versus-all” SVM clas-
338 sifiers is trained on top of DKN for each category using LIBSVM [41] and the
339 trade-off parameter C_k is also chosen by 3-fold cross-validation on the training
340 set. The average decision score from all the classifiers is taken as a final score

Framework	R	P	N₊
3-layer DKN	37.65	25.49	158
Initial DMN ($ \mathcal{S} = 500$)	31.30	18.67	155
Fine-tuned DMN ($ \mathcal{S}' = 2000$)	31.34	18.54	155
Fine-tuned DMN ($ \mathcal{S}' = 3000$)	31.62	18.43	153
Fine-tuned DMN ($ \mathcal{S}' = 4000$)	31.18	19.04	155
Fine-tuned DMN ($ \mathcal{S}' = 4999$)	31.65	19.13	157
Initial DMN ($ \mathcal{S} = 700$)	32.31	19.39	155
Fine-tuned DMN ($ \mathcal{S}' = 2000$)	32.57	19.82	157
Fine-tuned DMN ($ \mathcal{S}' = 3000$)	33.05	20.88	159
Fine-tuned DMN ($ \mathcal{S}' = 4000$)	33.08	20.40	158
Fine-tuned DMN ($ \mathcal{S}' = 4999$)	33.30	20.18	158

Table 4: The discrimination power of different DMNs w.r.t. the underlying DKN on COREL5k; in these experiments, two initial DMNs are designed using 500 and 700 samples.

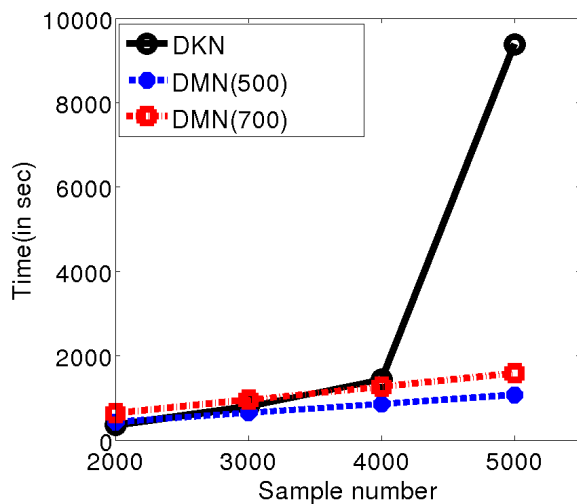


Figure 5: Comparison of processing time between two approximated DMNs (with $|\mathcal{S}| = 500$ and $|\mathcal{S}| = 700$) and their underlying DKN as $|\mathcal{T}|$ increases on COREL5k dataset.

Framework	$ \mathcal{S}' $	2K	3K	4K	4999
Initial DMN $ \mathcal{S} = 500$	-	2.45	2.41	2.35	2.26
	500	1.22	1.28	1.32	1.37
	1000	1.23	1.35	1.40	1.42
Fine-tuned DMN	2000	1.12	1.15	1.19	1.19
	3000	1.14	1.12	1.13	1.12
	4000	1.18	1.14	1.11	1.10
	4999	1.18	1.14	1.12	1.10
Initial DMN $ \mathcal{S} = 700$	-	2.43	2.39	2.33	2.24
	700	1.30	1.42	1.48	1.51
	1000	1.22	1.35	1.42	1.44
Fine-tuned DMN	2000	1.09	1.13	1.17	1.18
	3000	1.11	1.10	1.11	1.11
	4000	1.16	1.12	1.09	1.08
	4999	1.16	1.12	1.10	1.08

Table 5: Relative errors of initial and fine-tuned DMNs (w.r.t. the underlying DKN) on COREL5k as $|\mathcal{T}|$ increases (with values ranging from 2K to 4999)

Method	Learned Input feat.	context	R	P	N₊
CRM [43]	no	no	19	16	107
InfNet [44]	no	no	24	17	112
JEC-15 [45]	no	yes	33	28	140
TagPop σ ML [42]	no	yes	42	33	160
wTKML [46]	no	yes	42	21	173
LDMKL [47]	no	yes	44	29	179
CNN-R [48]	yes	yes	41.3	32.0	166
3-layer DKN+SVM [49]	no	no	37.7	25.5	158
Init. DMN+SVM ($ \mathcal{S} = 700$)	no	no	32.3	19.3	155
FT DMN+SVM ($ \mathcal{S} = 700$)	no	no	33.1	20.9	159
Init. DMN+SVM ($ \mathcal{S} = 1200$)	no	no	34.0	20.9	162
FT DMN+SVM ($ \mathcal{S} = 1200$)	no	no	34.7	21.0	168
ResNet[50] + SVM	yes	no	34.5	21.8	161
3-layer DKN+SVM [49]	yes	no	42.6	24.9	180
Init. DMN+SVM ($ \mathcal{S} = 700$)	yes	no	36.1	21.7	166
FT DMN+SVM ($ \mathcal{S} = 700$)	yes	no	36.8	22.4	165
Init. DMN+SVM ($ \mathcal{S} = 1000$)	yes	no	37.4	21.6	162
FT DMN+SVM ($ \mathcal{S} = 1000$)	yes	no	37.7	22.3	164
Init. DMN+SVM ($ \mathcal{S} = 1200$)	yes	no	37.8	23.2	167
FT DMN+SVM ($ \mathcal{S} = 1200$)	yes	no	38.9	23.2	169

Table 6: Extra comparison of the proposed DMN w.r.t different settings as well as the related work. In these experiments, $|\mathcal{S}'| = 3000$ and different $|\mathcal{S}|$ are used. In this table, FT stands for Fine-Tuned.

341 for a given category. In order to avoid the severe imbalanced class distribu-
 342 tions in SVM training, we adopt a sampling strategy that randomly selects a
 343 subset of negative samples whose cardinality is equal to the number of positive
 344 training samples. Hence, each classifier is learned using all the positive data
 345 and a random subset of negative data. The discrimination power of the learned
 346 DKNs+SVMs is shown in Table 4.

347 **Initial and fine-tuned DMNs.** Assuming the weights $\{\mathbf{w}_{p,q}^{(l-1)}\}$ of DKN
 348 known (learned), we build the initial DMN as shown in Section 3. We consider
 349 two random samplings of the subset \mathcal{S} – from the training set with $|\mathcal{S}| = 500$
 350 and $|\mathcal{S}| = 700$ – in order to build the initial DMN. We also use the learning
 351 procedure presented in Section 4 in order to fine-tune the parameter of the
 352 DMN. We consider an unlabeled set \mathcal{S}' which includes up to 4,999 samples (i.e.
 353 the whole COREL5k set); again we randomly sample 100,000 pairs in order
 354 to minimize the criterion in Eq. (9) using gradient descent with a step-size
 355 empirically set to 10^{-6} , a mini-batch size equal to 200 and a max number of
 356 iterations set to 5,000.

357 According to Table 4, we observe that the performances of the initial DMNs
 358 (\mathbf{R} , \mathbf{P} and \mathbf{N}_+) again degrade compared to their underlying DKNs as a result of
 359 the high RE of these DMNs. This degradation in performances is also amplified
 360 by the scarceness of training data for SVM learning in COREL5k (in contrast to
 361 ImageCLEF) especially when the RE is relatively large (see Table. 5). However,
 362 the discrimination power is improved when more data are used to design these
 363 DMNs (i.e., with $|\mathcal{S}| = 700$ and also $|\mathcal{S}| = 1200$ in Table 6). Furthermore,
 364 fine-tuning DMNs reduces the RE as $|\mathcal{S}'|$ increases, and makes RE stable even
 365 with a relatively large $|\mathcal{T}|$, so RE (on COREL5k) behaves similarly compared
 366 to ImageCLEF. Finally, Fig. 5 shows a comparison of processing time between
 367 DMN and DKN. It is easy to see that when $|\mathcal{T}|$ is small, the processing times of
 368 DKN and DMN are comparable. However, when $|\mathcal{T}|$ reaches large values (e.g.,
 369 $|\mathcal{T}| = 4999$), DMN becomes an order of magnitude faster than its underlying
 370 DKN while maintaining a comparable accuracy.

371 **Extra comparisons.** We further compare the performance of DMNs against

372 two generative methods (i.e., CRM [43] and infNet [44]) and several discrimina-
 373 tive methods (i.e., JEC-15 [45], TagPop σ ML [42], wTKML [46], LDMKL [47],
 374 CNN-R [48]) for image annotation. CRM [43] and infNet [44] learn optimal joint
 375 probability distributions between features and semantic labels while JEC-15 [45]
 376 and TagPop σ ML [42] (based on KNN) define classification criteria for images
 377 by weighting labels of their neighbors. wTKML [46] and LDMKL [47] are the
 378 most closely related (kernel) methods; wTKML learns explicit and transductive
 379 kernel maps using a priori knowledge taken from the statistical (semantic and ge-
 380 ometric) dependencies between classes while LDMKL combines Laplacian SVM
 381 with deep kernel networks using an “end-to-end” learning framework. CNN-
 382 R [48] adopts convolutional neural networks for annotation, that combines deep
 383 features from Caffe-Net with word2vec embedding. As introduced in the liter-
 384 ature, these related methods leverage different sources of contexts and a priori
 385 knowledge while our method does not.

386 In our experiments (see Table 6), we use four elementary kernels (linear,
 387 polynomial, RBF and HI) combined with different features as inputs to the
 388 designed DKN and DMN networks: “handcrafted features” including GIST
 389 and SIFT and “learned features” taken from ResNet [50] (pretrained on the
 390 ImageNet) which is a very deep architecture consisting of 152 layers; the 2048
 391 dimensional features of the last pooling layer are used in our annotation task.
 392 Using all these elementary kernels and features, we first train a DKN in a
 393 supervised way according to [49], then we design and fine-tune its associated
 394 DMNs with $|\mathcal{S}| = 700$ and $|\mathcal{S}'| = 3000$ (as done in Table. 4).

395 From the results shown in Table 6, first, we observe that the use of ResNet
 396 features as inputs to our DMN framework provides a clear gain compared to the
 397 use of handcrafted features. Second, fine-tuning DMNs bring a clear gain com-
 398 pared to the initial DMNs as well as ResNet. Our DKN (and its DMN variant)
 399 can even catch (and sometimes outperform) the aforementioned related work
 400 which again relies on different contextual clues, in contrast to our method. We
 401 believe that considering context will further enhance the performance of DKNs
 402 and their associated DMNs, but this is out of the main scope of this paper and

403 will be investigated as a future work.

404 Finally, Fig. 6 shows examples of annotation results, on the test set, ob-
 405 tained using the learned DMNs and the underlying DKNs on ImageCLEF and
 406 COREL5k datasets. From these figures, DMNs behave similarly, w.r.t. DKNs,
 407 with an extra advantage of being computationally more efficient especially on
 408 COREL5k (as shown in Table 7); whereas the computational complexity of
 409 DKN evaluation scales linearly w.r.t. the number of support vectors (which is
 410 an order of magnitude larger on COREL5k w.r.t. ImageCLEF: 4,500 versus
 411 500), the computational complexity of DMN evaluation grows slowly and re-
 412 mains globally stable w.r.t. the number of support vectors (which is again an
 413 order of magnitude larger on COREL5k). These results are also consistent with
 414 those already shown in Fig. 4 and Fig. 5.

Dataset	Framework	time (in sec)
ImageCLEF	DKN	0.68
	Fine-tuned DMN ($ \mathcal{S} = 500$)	0.57
	Fine-tuned DMN ($ \mathcal{S} = 1000$)	0.95
COREL5k	DKN	10.39
	Fine-tuned DMN ($ \mathcal{S} = 500$)	1.22
	Fine-tuned DMN ($ \mathcal{S} = 700$)	1.58
	Fine-tuned DMN ($ \mathcal{S} = 1000$)	2.51
	Fine-tuned DMN ($ \mathcal{S} = 1200$)	3.67

Table 7: Comparison of the average processing time per test image (excluding feature extraction) on ImageCLEF and COREL5k datasets.

415 5.3. Banana dataset

416 In Sections 5.1 and 5.2, we studied the efficiency and the effectiveness of
 417 the proposed method in image annotation. In this section, we further inves-
 418 tigate the applicability of our method to another classification problem using
 419 the Banana dataset [51]. The latter differs from ImageCLEF and COREL5k



Figure 6: Examples of annotation results using DKNs and their "Fine-tuned" DMN variants on ImageCLEF (top) and COREL5k (bottom). "GT" stands for ground-truth keywords and the symbol "*" stands for the presence of a keyword in a given test image.

Method	LDKL [51]	SDMKL [49]	Initial DMNs	Fine-tuned DMNs
Accuracy	88.67	91.07	90.21	91.05

Table 8: This table shows the comparison in accuracy of LDKL, kernel-based semi-supervised learning of 3-layer deep kernel network (SDMKL), initial DMNs and fine-tuned DMNs on Banana dataset.

420 in that (i) it corresponds to vectors of measurements⁶ rather than images, and
 421 (ii) the labels associated to these measurements are exclusive. The Banana
 422 dataset corresponds to a binary classification problem, which contains 1,000
 423 samples for training and 4,300 for testing. Following the standard experimental
 424 protocol [51], we first evaluate 21 RBF kernels (with scale factors ranging from
 425 $2^{-10}M$ to $2^{10}M$; here M is the average Euclidean distance between samples
 426 and their neighbors), then we train a three-layer deep kernel network (referred

⁶Similar to the spirit of industrial scenarios where data are collected using specific instruments.

427 to as SDMKL [49]) using a semi-supervised setting. In these experiments, we
428 set the SVM trade-off parameter C_k to 0.25 using cross validation on a random
429 subset containing 20% of training data. We also consider another baseline for
430 comparison referred to as LDKL (Local Deep Kernel Learning) [51]. The accu-
431 racy of LDKL and SDMKL as well as our DMNs (designed and fine-tuned as
432 described in Sections 3, 4) are shown in Tab. 8. From this table, we observe that
433 the performances of the initial DMNs are close to their original DKNs and the
434 fine-tuned DMNs further improve these performances, and this clearly validates
435 the applicability of our method to different classification problems.

436 6. Conclusion

437 In this paper we introduce a novel method that transforms deep kernel net-
438 works (DKNs) into highly efficient deep map networks (DMNs). DKNs, as
439 nonlinear and multi-layered combinations of standard kernels, are highly effec-
440 tive but computationally very demanding especially when handling large-scale
441 problems. In order to reduce the computational complexity of DKNs, the pro-
442 posed method defines a DMN architecture layer-wise by expressing positive
443 semi-definite kernels in different (input, intermediate, and output) layers of
444 DKN as inner products involving explicit maps. As also theoretically shown,
445 these maps are either exactly designed for input kernels (including linear and
446 polynomial) or tightly approximated (for intermediate and output kernels in
447 DKN) with at least an order of magnitude gain both in kernel and SVM eval-
448 uation while maintaining a comparable classification accuracy. An extra fine-
449 tuning step makes it possible to further enhance the accuracy of DMNs; this
450 step, totally unsupervised, benefits from large unlabeled sets in order to further
451 minimize the difference between the inner products of the designed maps and
452 the original DKNs in the associated Reproducing Kernel Hilbert Space. Exten-
453 sive experiments on the challenging ImageCLEF and COREL5k benchmarks for
454 image annotation as well as the Banana dataset, clearly demonstrate the effec-
455 tiveness of DMNs and their high efficiency. As a future work, we are currently

456 studying the memory burden of DMNs that may arise as a function of their
 457 (high) dimensionality and depth; this issue will be tackled by coupling DMNs
 458 with auto-encoders.

459 Appendix A: Proof of Proposition 1

Proof. For pair of samples $(\mathbf{x}_i, \mathbf{x}_j), \forall i, j \in \{1, \dots, m\}$, we have:

$$\begin{aligned} \langle \hat{\phi}_p^{(1)}(\mathbf{x}_i), \hat{\phi}_p^{(1)}(\mathbf{x}_j) \rangle &= \langle \psi(\mathbf{x}_i^1), \psi(\mathbf{x}_j^1) \rangle \left(\frac{u_1 - \ell_1}{Q} \right) + \ell_1 + \dots \\ &+ \langle \psi(\mathbf{x}_i^s), \psi(\mathbf{x}_j^s) \rangle \left(\frac{u_s - \ell_s}{Q} \right) + \ell_s \end{aligned} \quad (18)$$

It is easy to see that $\forall d \in \{1, \dots, s\}, \langle \psi(\mathbf{x}_i^d), \psi(\mathbf{x}_j^d) \rangle = \min(k(\mathbf{x}_i^d), k(\mathbf{x}_j^d))$. By replacing in Eq. (18)

$$\begin{aligned} & \left| \langle \hat{\phi}_p^{(1)}(\mathbf{x}_i), \hat{\phi}_p^{(1)}(\mathbf{x}_j) \rangle - \kappa_p^{(1)}(\mathbf{x}_i, \mathbf{x}_j) \right| \\ &= \left| \sum_{d=1}^s \min(k(\mathbf{x}_i^d), k(\mathbf{x}_j^d)) \frac{u_d - \ell_d}{Q} + \ell_d - \min(\mathbf{x}_i^d, \mathbf{x}_j^d) \right| \\ &\leq \sum_{d=1}^s \left| \left[Q \frac{\min(\mathbf{x}_i^d, \mathbf{x}_j^d) - \ell_d}{u_d - \ell_d} \right] \frac{u_d - \ell_d}{Q} + \ell_d - \min(\mathbf{x}_i^d, \mathbf{x}_j^d) \right| \\ &= \sum_{d=1}^s \frac{u_d - \ell_d}{Q} \left| \left[Q \frac{\min(\mathbf{x}_i^d, \mathbf{x}_j^d) - \ell_d}{u_d - \ell_d} \right] - Q \frac{\min(\mathbf{x}_i^d, \mathbf{x}_j^d) - \ell_d}{u_d - \ell_d} \right| \\ &\leq \frac{1}{Q} \sum_{d=1}^s u_d - \ell_d, \quad (\text{as } ||z| - z| \leq 1), \end{aligned} \quad (19)$$

460 as Q increases, $\left| \langle \hat{\phi}_p^{(1)}(\mathbf{x}_i), \hat{\phi}_p^{(1)}(\mathbf{x}_j) \rangle - \kappa_p^{(1)}(\mathbf{x}_i, \mathbf{x}_j) \right| \rightsquigarrow 0$. □

461 Appendix B: Proof of Proposition 2

462 *Proof.* Let's proceed layer-wise by induction; for $l = 1$ (and following Section
 463 3.1), the initial kernel maps $\{\hat{\phi}_p^{(1)}(\cdot)\}$ are designed to satisfy $\hat{\phi}_p^{(1)}(\mathbf{x})^\top \hat{\phi}_p^{(1)}(\mathbf{x}') =$
 464 $\kappa_p^{(1)}(\mathbf{x}, \mathbf{x}')$.

Now provided that $\hat{\phi}_q^{(l-1)}(\mathbf{x})^\top \hat{\phi}_q^{(l-1)}(\mathbf{x}') = \kappa_q^{(l-1)}(\mathbf{x}, \mathbf{x}')$, the property to show is $\hat{\phi}_p^{(l)}(\mathbf{x})^\top \hat{\phi}_p^{(l)}(\mathbf{x}') = \kappa_p^{(l)}(\mathbf{x}, \mathbf{x}')$, $\forall \mathbf{x} \in \mathcal{S}$. Following (7) we have

$$\begin{aligned} \langle \hat{\phi}_p^{l,c}(\mathbf{x}), \hat{\phi}_p^{l,c}(\mathbf{x}') \rangle &= \sum_{q=1}^{n_{l-1}} \mathbf{w}_{p,q}^{(l-1)} \hat{\phi}_q^{(l-1)}(\mathbf{x})^\top \hat{\phi}_q^{(l-1)}(\mathbf{x}') \\ &= \sum_{q=1}^{n_{l-1}} \mathbf{w}_{p,q}^{(l-1)} \kappa_q^{(l-1)}(\mathbf{x}, \mathbf{x}'), \end{aligned} \quad (20)$$

the second equality results from the hypothesis of induction. By plugging (20) into (6), we obtain

$$\hat{\phi}_p^{(l)}(\mathbf{x})^\top = (\kappa_p^{(l)}(\mathbf{x}, \mathbf{x}_1), \dots, \kappa_p^{(l)}(\mathbf{x}, \mathbf{x}_N)) \mathbf{U}_p^{(l)}, \quad (21)$$

and equivalently $\hat{\mathbf{K}}_p^l = \mathbf{K}_p^l \mathbf{U}_p^{(l)} \mathbf{U}_p^{(l)\top} \mathbf{K}_p^l$. Hence,

$$\begin{aligned} \|\hat{\mathbf{K}}_p^l - \mathbf{K}_p^l\|_2 &= \|\mathbf{K}_p^l \boldsymbol{\alpha} \boldsymbol{\Lambda}^{-1/2} \boldsymbol{\Lambda}^{-1/2} \boldsymbol{\alpha}^\top \mathbf{K}_p^l - \mathbf{K}_p^l\|_2 \\ &= \|\boldsymbol{\alpha} \boldsymbol{\Lambda} \boldsymbol{\Lambda}^{-1} \boldsymbol{\alpha}^\top \mathbf{K}_p^l - \mathbf{K}_p^l\|_2 \\ &= \|\boldsymbol{\alpha} \boldsymbol{\alpha}^\top \mathbf{K}_p^l - \mathbf{K}_p^l\|_2 \\ &= \|\mathbf{K}_p^l - \mathbf{K}_p^l\|_2 = 0 \end{aligned} \quad (22)$$

465 which also results from Eq. (5) and the orthogonality of eigenvectors in $\boldsymbol{\alpha}$. \square

466 **Acknowledgement**

467 This work was supported by a grant from National Natural Science Founda-
468 tion of China (No. 61806180) and in part by a grant from the research agency
469 ANR (Agence Nationale de la Recherche) under the MLVIS project (ANR-11-
470 BS02-0017).

471 **References**

472 **References**

- 473 [1] B. Caputo, C. Wallraven, M. Nilsback, Object categorization via local
474 kernels, in: Proceedings of the 17th International Conference on Pattern
475 Recognition (ICPR), Vol. 2, 2004, pp. 132–135.
- 476 [2] S. Lyu, Mercer kernels for object recognition with local features, in: 2005
477 IEEE Computer Society Conference on Computer Vision and Pattern
478 Recognition (CVPR), Vol. 2, 2005, pp. 223–229.
- 479 [3] K. Grauman, T. Darrell, The pyramid match kernel: Efficient learning with
480 sets of features, Journal of Machine Learning Research 8 (2007) 725–760.

- 481 [4] X. Qi, Y. Han, Incorporating multiple svms for automatic image annota-
482 tion, *Pattern Recognition* 40 (2) (2007) 728–741.
- 483 [5] K. Q. Weinberger, F. Sha, L. K. Saul, Learning a kernel matrix for non-
484 linear dimensionality reduction, in: *International Conference on Machine*
485 *Learning (ICML)*, 2014, pp. 839–846.
- 486 [6] V. Vapnik, *Statistical learning theory*, Wiley, New York, 1998.
- 487 [7] J. Shawe-Taylor, N. Cristianini, *Kernel methods for pattern analysis*, Cam-
488 bridge University Press, 2004.
- 489 [8] G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, M. I. Jordan, Learn-
490 ing the kernel matrix with semi-definite programming, *Journal of Machine*
491 *Learning Research* 5 (2004) 27–72.
- 492 [9] K. Yu, W. Xu, Y. Gong, Deep learning with kernel regularization for vi-
493 sual recognition, in: *Advances in Neural Information Processing Systems*
494 *(NIPS)*, 2008, pp. 1889–1896.
- 495 [10] C. Corinna, M. Mehryar, R. Afshin, Two-stage learning kernel algorithms,
496 in: *International Conference on Machine Learning (ICML)*, 2010, pp. 239–
497 246.
- 498 [11] H. Sahbi, J.-Y. Audibert, R. Keriven, Context-dependent kernels for ob-
499 ject classification, *IEEE Transactions on Pattern Analysis and Machine*
500 *Intelligence* 33 (2011) 699–708.
- 501 [12] H. Sahbi, X. Li, Context-based support vector machines for interconnected
502 image annotation, in: *Asian Conference on Computer Vision (ACCV)*,
503 2011, pp. 214–227.
- 504 [13] F. Bach, G. Lanckriet, M. Jordan, Multiple kernel learning, conic duality,
505 and the smo algorithm, in: *International Conference on Machine Learning*
506 *(ICML)*, 2004, pp. 1–6.

- 507 [14] A. Rakotomamonjy, F. Bach, C. S., G. Yves, Simplemkl, Journal of Ma-
508 chine Learning Research 9 (2008) 2491–2521.
- 509 [15] S. Sonnenburg, G. Rätsch, C. Schafer, B. Schölkopf, Large scale multiple
510 kernel learning, Journal of Machine Learning Research 7 (2006) 1531–1565.
- 511 [16] Y. Wang, X. Liu, Y. Dou, Q. Lv, Y. Lu, Multiple kernel learning with
512 hybrid kernel alignment maximization, Pattern Recognition 70 (2017) 104–
513 111.
- 514 [17] F. Bach, Exploring large feature spaces with hierarchical multiple kernel
515 learning, in: Advances in Neural Information Processing Systems (NIPS),
516 2009, pp. 1–9.
- 517 [18] C. Cortes, M. Mohri, A. Rostamizadeh, Learning non-linear combinations
518 of kernels, in: Advances in Neural Information Processing Systems (NIPS),
519 2009, pp. 1–9.
- 520 [19] Y. LeCun, L. Botto, Y. Bengio, P. Haffner, Gradient-based learning applied
521 to document recognition, Proceedings of IEEE 86 (11) (1998) 2278–2324.
- 522 [20] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with
523 deep convolutional neural networks, in: Advances in Neural Information
524 Processing Systems (NIPS), Vol. 60, 2012, pp. 1097–1105.
- 525 [21] C. Farabet, C. Couprie, L. Najman, Y. LeCun, Learning hierarchical fea-
526 tures for scene labeling, IEEE Transactions on Pattern Analysis and Ma-
527 chine Intelligence 35 (8) (2013) 1915–1929.
- 528 [22] Y. Cho, L. Saul, Kernel methods for deep learning, in: Advances in Neural
529 Information Processing Systems (NIPS), Vol. 28 (1), 2009, pp. 342–350.
- 530 [23] J. Zhuang, I. Tsang, S. Hoi, Two-layer multiple kernel learning, in: Inter-
531 national Conference on Machine Learning (ICML), 2011, pp. 909–917.

- 532 [24] M. Jiu, H. Sahbi, Semi supervised deep kernel design for image annotation,
533 in: 2015 IEEE International Conference on Acoustics, Speech and Signal
534 Processing (ICASSP), 2015, pp. 1156–1160.
- 535 [25] C. Williams, M. Seeger, Using the nyström method to speed up kernel
536 machines, in: Advances in Neural Information Processing Systems (NIPS),
537 2001, pp. 682–688.
- 538 [26] P. Drineas, M. Mahoney, On the nyström method for approximating a gram
539 matrix for improved kernel-based learning, *Journal of Machine Learning*
540 *Research* 6 (2005) 2153–2175.
- 541 [27] S. Kumar, M. Mohri, A. Talwalkar, Sampling methods for the nyström
542 method, *Journal of Machine Learning Research* 13 (1) (2012) 981–1006.
- 543 [28] A. Rahimi, B. Recht, Random features for large-scale kernel machines, in:
544 *Advances in Neural Information Processing Systems (NIPS)*, Vol. 20, 2007,
545 pp. 1177–1184.
- 546 [29] F. Li, C. Ionescu, C. Sminchisescu, Random fourier approximations for
547 skewed multiplicative histogram kernels, in: *DAGM conference Pattern*
548 *Recognition*, 2010, pp. 262–271.
- 549 [30] A. Iosifidis, M. Gabbouj, Nyström-based approximate kernel subspace
550 learning, *Pattern Recognition* 57 (2016) 190–197.
- 551 [31] C. Zhu, D. Gao, Improved multi-kernel classification machine with nyström
552 approximation technique, *Pattern Recognition* 48 (2015) 1490–1509.
- 553 [32] A. Vedaldi, A. Zisserman, Efficient additive kernels via explicit feature
554 maps, *IEEE Transactions on Pattern Analysis and Machine Intelligence*
555 34 (3) (2012) 480–492.
- 556 [33] D. López-Sánchez, A. G. Arrieta, J. M. Corchado, Data-independent ran-
557 dom projections from the feature-space of the homogeneous polynomial
558 kernel, *Pattern Recognition* 82 (2018) 130–146.

- 559 [34] P. Huang, L. Deng, M. Hasegawa-Johnson, X. He, Random features for
560 kernel deep convex network, in: 2013 IEEE International Conference on
561 Acoustics, Speech and Signal Processing (ICASSP), 2013, pp. 3143–3147.
- 562 [35] J. Mairal, P. Koniusz, Z. Harchaoui, C. Schmid, Convolutional kernel net-
563 works, in: Advances in Neural Information Processing Systems (NIPS),
564 2014, pp. 2627–2635.
- 565 [36] M. Jiu, H. Sahbi, Deep kernel map networks for image annotation, in: 2016
566 IEEE International Conference on Acoustics, Speech and Signal Processing
567 (ICASSP), 2016, pp. 1571–1575.
- 568 [37] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, C.-J. Lin, Liblinear: A
569 library for large linear classification, *Journal of Machine Learning Research*
570 9 (2008) 1871–1874.
- 571 [38] M. Villegas, R. Paredes, B. Thomee, Overview of the imageclef 2013 scal-
572 able concept image annotation subtask, in: CLEF 2013 Evaluation Labs
573 and Workshop, 2013.
- 574 [39] P. Duygulu, K. Barnard, N. de Freitas, D. Forsyth, Object recognition as
575 machine translation: Learning a lexicon for a fixed image vocabulary, in:
576 European Conference on Computer Vision (ECCV), 2002, pp. 97–112.
- 577 [40] H. Sahbi, Imageclef annotation with explicit context-aware kernel maps,
578 *International Journal of Multimedia Information Retrieval* 4 (2015) 113–
579 128.
- 580 [41] C.-C. Chang, C.-J. Lin, Libsvm: A library for support vector machines,
581 *ACM Transactions on Intelligent Systems and Technology* 2 (2011) 1–27.
- 582 [42] M. Guillaumin, T. Mensink, J. Verbeek, C. Schmid, Tagprop: Discrimina-
583 tive metric learning in nearest neighbor models for image auto-annotation,
584 in: 2009 IEEE 12th International Conference on Computer Vision (ICCV),
585 2009, pp. 309–316.

- 586 [43] V. Lavrenko, R. Manmatha, J. Jeon, A model for learning the semantics of
587 pictures, in: *Advances in Neural Information Processing Systems (NIPS)*,
588 2003, pp. 553–560.
- 589 [44] D. Metzler, R. Manmatha, A inference network approach to image retrieval,
590 in: *International Conference on Image and Video Retrieval (CIVR)*, 2004,
591 pp. 42–50.
- 592 [45] A. Makadia, V. Pavlovic, S. Kumar, A new baseline for image annotation,
593 in: *European Conference on Computer Vision (ECCV)*, 2008, pp. 316–329.
- 594 [46] P. Vo, H. Sahbi, Transductive kernel map learning and its application to
595 image annotation, in: *The British Machine Vision Conference (BMVC)*,
596 2012, pp. 1–12.
- 597 [47] D. Zhang, M. Islam, G. Lu, A review on automatic image annotation tech-
598 niques, *Journal of the China Society for Scientific and Technical Informa-*
599 *tion* 45 (1) (2013) 346–362.
- 600 [48] V. N. Murthy, S. Maji, R. Manmatha, Automatic image annotation using
601 deep learning representations, in: *International Conference on Multimedia*
602 *Retrieval*, 2015, pp. 603–606.
- 603 [49] M. Jiu, H. Sahbi, Nonlinear deep kernel learning for image annotation,
604 *IEEE Transactions on Image Processing* 26 (4) (2017) 1820–1832.
- 605 [50] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recog-
606 nition, in: *IEEE Computer Society Conference on Computer Vision and*
607 *Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- 608 [51] C. Jose, P. Goyal, P. Aggrwal, M. Varma, Local deep kernel learning for ef-
609 ficient non-linear svm prediction, in: *International Conference on Machine*
610 *Learning (ICML)*, 2013, pp. 486–494.