Bridging Languages through Images with Deep Partial Canonical Correlation Analysis

Guy Rotman\(^1\), Ivan Vulić\(^2\) and Roi Reichart\(^1\)

\(^1\) Faculty of Industrial Engineering and Management, Technion, IIT
\(^2\) Language Technology Lab, University of Cambridge
grotman@campus.technion.ac.il
iv250@cam.ac.uk roiri@technion.ac.il

Abstract

We present a deep neural network that leverages images to improve bilingual text embeddings. Relying on bilingual image tags and descriptions, our approach conditions text embedding induction on the shared visual information for both languages, producing highly correlated bilingual embeddings. In particular, we propose a novel model based on Partial Canonical Correlation Analysis (PCCA). While the original PCCA finds linear projections of two views in order to maximize their canonical correlation conditioned on a shared third variable, we introduce a non-linear Deep PCCA (DPCCA) model, and develop a new stochastic iterative algorithm for its optimization. We evaluate PCCA and DPCCA on multilingual word similarity and cross-lingual image description retrieval. Our models outperform a large variety of previous methods, despite not having access to any visual signal during test time inference.\(^1\)

1 Introduction

Research in multi-modal semantics deals with the *grounding problem* (Harnad, 1990), motivated by evidence that many semantic concepts, irrespective of the actual language, are grounded in the perceptual system (Barsalou and Wiemer-Hastings, 2005). In particular, recent studies have shown that performance on NLP tasks can be improved by joint modeling of text and vision, with multi-modal and perceptually enhanced representation learning outperforming purely textual representations (Feng and Lapata, 2010; Kiela and Bottou, 2014; Lazaridou et al., 2015).

These findings are not surprising, and can be explained by the fact that humans understand language not only by its words, but also by their visual/perceptual context. The ability to connect vision and language has also enabled new tasks which require both visual and language understanding, such as visual question answering (Antol et al., 2015; Fukui et al., 2016; Xu and Saenko, 2016), image-to-text retrieval and text-to-image retrieval (Kiros et al., 2014; Mao et al., 2014), image caption generation (Farhadi et al., 2010; Mao et al., 2015; Vinyals et al., 2015; Xu et al., 2015), and visual sense disambiguation (Gella et al., 2016).

While the main focus is still on monolingual settings, the fact that visual data can serve as a natural bridge between languages has sparked additional interest towards *multilingual multi-modal modeling*. Such models induce bilingual multi-modal spaces based on multi-view learning (Calixto et al., 2017; Gella et al., 2017; Rajendran et al., 2016).

In this work, we propose a novel effective approach for learning bilingual text embeddings conditioned on shared visual information. This additional perceptual modality bridges the gap between languages and reveals latent connections between concepts in the multilingual setup. The shared visual information in our work takes the form of images with word-level tags or sentence-level descriptions assigned in more than one language.

We propose a deep neural architecture termed Deep Partial Canonical Correlation Analysis (DPCCA) based on the Partial CCA (PCCA) method (Rao, 1969). To the best of our knowledge, PCCA has not been used in multilingual settings before. In short, PCCA is a variant of CCA which learns maximally correlated linear projections of two views (e.g., two language-specific “text-based views”) conditioned on a shared third view (e.g.,

---

\(^1\)Our code and data are available at: https://github.com/rotmanguy/DPCCA.
We discuss the PCCA and DPCCA methods in §3 and show how they can be applied without having access to the shared images at test time inference.

PCCA inherits one disadvantageous property from CCA: both methods compute estimates for covariance matrices based on all training data. This would prevent feasible training of their deep non-linear variants, since deep neural nets (DNNs) are predominantly optimized via stochastic optimization algorithms. To resolve this major hindrance, we propose an effective optimization algorithm for DPCCA, inspired by the work of Wang et al. (2015b) on Deep CCA (DCCA) optimization.

We evaluate our DPCCA architecture on two semantic tasks: 1) multilingual word similarity and 2) cross-lingual image description retrieval. For the former, we construct and provide to the community a new Word-Image-Word (WIW) dataset containing bilingual lexicons for three languages with shared images for 5K+ concepts. WIW is used as training data for word similarity experiments, while evaluation is conducted on the standard multilingual SimLex-999 dataset (Hill et al., 2015; Leviant and Reichart, 2015).

The results reveal stable improvements over a large space of non-deep and deep CCA-style baselines in both tasks. Most importantly, 1) PCCA is overall better than other methods which do not use the additional perceptual view; 2) DPCCA outperforms PCCA, indicating the importance of non-linear transformations modeled through DNNs; 3) DPCCA outscores DCCA, again verifying the importance of conditioning multilingual text embedding induction on the shared visual view; and 4) DPCCA outperforms two recent multi-modal bilingual models which also leverage visual information (Gella et al., 2017; Rajendran et al., 2016).

2 Related Work

This work is related to two research threads: 1) multi-modal models that combine vision and language, with a focus on multilingual settings; 2) correlational multi-view models based on CCA which learn a shared vector space for multiple views.

Multi-Modal Modeling in Multilingual Settings

Research in cognitive science suggests that human meaning representations are grounded in our perceptual system and sensori-motor experience (Harnad, 1990; Lakoff and Johnson, 1999; Louwerse, 2011). Visual context serves as a useful cross-lingual grounding signal (Bruni et al., 2014; Glavaš et al., 2017) due to its language invariance, even enabling the induction of word-level bilingual semantic spaces solely through tagged images obtained from the Web (Bergsma and Van Durme, 2011; Kiela et al., 2015). Vulić et al. (2016) combine text embeddings with visual features via simple techniques of concatenation and averaging to obtain bilingual multi-modal representations, with noted improvements over text-only embeddings on word similarity and bilingual lexicon extraction. However, similar to the monolingual model of Kiela and Bottou (2014), their models lack the training phase, and require the visual signal at test time.

Recent work from Gella et al. (2017) exploits visual content as a bridge between multiple languages by optimizing a contrastive loss function. Furthermore, Rajendran et al. (2016) extend the work of Chandar et al. (2016) and propose to use a pivot representation in multimodal multilingual setups, with English representations serving as the pivot. While these works learn shared multimodal multilingual vector spaces, we demonstrate improved performance with our models (see §7).

Finally, although not directly comparable, recent work in neural machine translation has constructed models that can translate image descriptions by additionally relying on visual features of the image provided (Calixto and Liu, 2017; Elliott et al., 2015; Hitschler et al., 2016; Huang et al., 2016; Nakayama and Nishida, 2017, inter alia).

Correlational Models

CCA-based techniques support multiple views on related data: e.g., when coupled with a bilingual dictionary, input monolingual word embeddings for two different languages can be seen as two views of the same latent semantic signal. Recently, CCA-based models for bilingual text embedding induction were proposed. These models rely on the basic CCA model (Chandar et al., 2016; Faruqui and Dyer, 2014), its deep variant (Lu et al., 2015), and a CCA extension which supports more than two views (Funaki and Nakayama, 2015; Rastogi et al., 2015). In this work, we propose to use (D)PCCA, which organically supports our setup: it conditions the two (textual) views on a shared (visual) view.

CCA-based methods (including PCCA) require the estimation of covariance matrices over all training data (Kessy et al., 2017). This hinders the use of DNNs with these models, as DNNs are typically trained via stochastic optimization over mini-
batches on very large training sets. To address this limitation, various optimization methods for Deep CCA were proposed. Andrew et al. (2013) use L-BFGS (Byrd et al., 1995) over all training samples, while Arora and Livescu (2013) and Yan and Mikolajczyk (2015) train with large batches. However, these methods suffer from high memory complexity with unstable numerical computations.

Wang et al. (2015b) have recently proposed a stochastic approach for CCA and DCCA which copes well with small and large batch sizes while preserving high model performance. They use orthogonal iterations to estimate a moving average of the covariance matrices, which improves memory consumption. Therefore, we base our novel optimization algorithm for DPCCA on this approach.

3 Methodology: Deep Partial CCA

Given two image descriptions \(x\) and \(y\) in two languages and an image \(z\) that they refer to, the task is to learn a shared bilingual space such that similar descriptions obtain similar representations in the induced space. The image \(z\) serves as a shared third view on the textual data during training. The representation model is then utilized in cross-lingual and monolingual tasks. In this paper we focus on the more realistic scenario where no relevant visual content is available at test time. For this goal we propose a novel Deep Partial CCA (DPCCA) framework.

In what follows, we first review the CCA model and its deep variant: DCCA. We then introduce our DPCCA architecture, and describe our new stochastic optimization algorithm for DPCCA.

3.1 CCA and Deep CCA

DCCA (Andrew et al., 2013) extends CCA by learning non-linear (instead of linear) transformations of features contained in the input matrices \(X \in \mathbb{R}^{D_x \times N}\) and \(Y \in \mathbb{R}^{D_y \times N}\), where \(D_x\) and \(D_y\) are input vector dimensionsality, and \(N\) is the number of input items. Since CCA is a special case of the non-linear DCCA (see below), we here briefly outline the more general DCCA model.

The DCCA architecture is illustrated in Figure 1a. Non-linear transformations are achieved through two DNNs \(f : \mathbb{R}^{D_x \times N} \rightarrow \mathbb{R}^{D'_x \times N}\) and \(g : \mathbb{R}^{D_y \times N} \rightarrow \mathbb{R}^{D'_y \times N}\) for \(X\) and \(Y\). \(D'_x\) and \(D'_y\) are the output dimensionalities. A final linear layer is added to resemble the linear CCA projection.

The goal is to project the features of \(X\) and \(Y\) into a shared \(L\)-dimensional \((1 \leq L \leq \min(D'_x, D'_y))\) space such that the canonical correlation of the final outputs \(F(X) = W^T f(X)\) and \(G(Y) = V^T g(Y)\) is maximized. \(W \in \mathbb{R}^{D'_x \times L}\) and \(V \in \mathbb{R}^{D'_y \times L}\) are projection matrices: they project the final outputs of the DNNs to the shared space. \(W_f\) and \(V_g\) (the parameters of \(f\) and \(g\)) and the projection matrices are the model parameters: \(W_F = \{W_f, W\}; V_G = \{V_g, V\}\).²

Formally, the DCCA objective can be written as:

\[
\max_{W_F, V_G} \text{Tr}(\Sigma_{FG})
\]

so that \(\hat{\Sigma}_{FF} = \hat{\Sigma}_{GG} = I.\)

\[
\hat{\Sigma}_{FG} \equiv \frac{1}{N-1} F(X)G(Y)^T
\]

is the estimation of the cross-covariance matrix of the outputs, and \(\hat{\Sigma}_{FF} \equiv \frac{1}{N} F(X)F(X)^T, \hat{\Sigma}_{GG} \equiv \frac{1}{N} G(Y)G(Y)^T\) are the estimations of the auto-covariance matrices of the outputs.³ Further, following Wang et al. (2015b), the optimal solution of Eq. (1) is equivalent to the optimal solution of the following:

\[
\min_{W_F, V_G} \frac{1}{N-1} \| F(X) - G(Y) \|^2_F
\]

s.t. \(\hat{\Sigma}_{FF} = \hat{\Sigma}_{GG} = I.\)

The main disadvantage of DCCA is its inability to support more than two views, and to learn conditioned on an additional shared view, which is why we introduce Deep Partial CCA.

3.2 New Model: Deep Partial CCA

Figure 1b illustrates the architecture of DPCCA. The training data now consists of triplets \((x_i, y_i, z_i)_{i=1}^N\) from three views, forming the columns of \(X, Y\) and \(Z\), where \(x_i \in \mathbb{R}^{D_x}, y_i \in \mathbb{R}^{D_y}, z_i \in \mathbb{R}^{D_z}\) for \(i = 1, \ldots, N\). The objective is to maximize the canonical correlation of the first two views \(X\) and \(Y\) conditioned on the shared third variable \(Z\). Following Rao (1969)’s work on Partial CCA, we first consider two multivariate linear multiple regression models:

\[
F(X) = AZ + F(X|Z),
\]

\[
G(Y) = BZ + G(Y|Z).
\]

²For notational simplicity, we assume \(f(X)\) and \(g(Y)\) have zero-means, otherwise it is possible to centralize them at the final layer of each network to the same effect.

³The CCA model can be seen as a special (linear) case of the more general DCCA model. The basic CCA objective can be recovered from the DCCA objective by simply setting \(D'_x = D_x, D'_y = D_y\) and \(f(X) = id_x, g(Y) = id_y; id\) is the identity mapping.
A, B ∈ \mathbb{R}^{L \times D_z}$ are matrices of coefficients, and $F(X|Z), G(Y|Z) \in \mathbb{R}^{L \times N}$ are normal random error matrices: residuals. We then minimize the mean-squared error regression criterion:

$$
\min_A \frac{1}{N-1} \| F(X) - AZ \|^2_F, \quad (5)
$$

$$
\min_B \frac{1}{N-1} \| G(Y) - BZ \|^2_F. \quad (6)
$$

After obtaining the optimal solutions for the coefficients, $A$ and $B$, the residuals are as follows:

$$
F(X|Z) = F(X) - \hat{A}Z = F(X) - \hat{\Sigma}_FZ^{-1}Z. \quad (7)
$$

$G(Y|Z)$ is computed in the analogous manner, now relying on $G(Y)$ and $\hat{B}Z$. $\hat{\Sigma}_Z = \frac{1}{N-1}SZ^T$ refers to the covariance matrix estimator of $S'$ and $Z$, where $(S', S) \in \{(F, F(X)), (G, G(Y)), (Z, Z)\}$.\(^4\)

The canonical correlation between the residual matrices $F(X|Z)$ and $G(Y|Z)$ is referred to as the partial canonical correlation. The Deep PCCA objective can be obtained by replacing $F(X)$ and $G(Y)$ with their residuals in Eq. (2):

$$
\min_{W, v_G} \frac{1}{N-1} \| F(X|Z) - G(Y|Z) \|^2_F \quad \text{s.t. } \hat{\Sigma}_{F|F} = \hat{\Sigma}_{G|G} = I. \quad (8)
$$

The computation of the conditional covariance matrix $\hat{\Sigma}_{F|F|Z}$ can be formulated as follows:

$$
\hat{\Sigma}_{F|F|Z} = \frac{1}{N-1} (F(X|Z)F(X|Z)^T - \hat{\Sigma}_{F|F} - \hat{\Sigma}_{F|G}Z^{-1}Z\hat{\Sigma}_{F|G}Z^{-1}Z\hat{\Sigma}_{F|F}). \quad (9)
$$

The other conditional covariance matrix $\hat{\Sigma}_{G|G|Z}$ is again computed in the analogous manner, replacing $F$ with $G$ and $X$ with $Y$.\(^5\)

While the (D)PCCA objective is computed over the residuals, after the network is trained (using multilingual texts and corresponding images) we can compute the representations of $F(X)$ and $G(Y)$ at test time without having access to images (see the network structure in Figure 1b). This heuristic enables the use of DPCCA in a real-life scenario in which images are unavailable at test time, and its encouraging results are demonstrated in §7.

**Model Variants** We consider two DPCCA variants: 1) in DPCCA Variant A, the shared view $Z$ is kept fixed; 2) DPCCA Variant B also optimizes over $Z$, as illustrated in Figure 1b. Variant A may be seen as a special case of Variant B.\(^6\)

Variant B learns a non-linear function of the shared variable, $H(Z) = U^Th(Z)$, during training, where $h: \mathbb{R}^{D_x \times N} \to \mathbb{R}^{D_i \times N}$ is a DNN having the same architecture as $f$ and $g$. $U \in \mathbb{R}^{D_i \times L}$ is the final linear layer of $H$, such that overall, the additional parameters of the model are $U_H = \{U_h, U\}$. Instead of assuming a linear connection between $F(X)$ and $G(Y)$ to $Z$, as in Variant A, we now assume that the linear connection takes place with $H(Z)$. This assumption

---

\(^4\)A small value $\epsilon > 0$ is added to the main diagonal of the covariance estimators for numerical stability.

\(^5\)The original PCCA objective can be recovered by setting $D_x = D_y = D_y$ and $f(X) = id_X$, $g(Y) = id_Y$.

\(^6\)For Variant A, in order for $Z$ to be on the same range of values as in $F$ and $G$, we pass it through the activation function of the network, $Z = \sigma(Z)$. Due to space constraints we discuss DPCCA Variant A in the supplementary material only.
changes Eq. (3) and Eq. (4) to:⁷

\[ F(X) = A' \cdot H(Z) + F(X|H(Z)), \quad (10) \]
\[ G(Y) = B' \cdot H(Z) + G(Y|H(Z)). \quad (11) \]

4 DPCCA: Optimization Algorithm

Training deep variants of CCA-style multi-view models is non-trivial due to estimation on the entire training set related to whitening constraints (i.e., the orthogonality of covariance matrices). To overcome this issue, Wang et al. (2015b) proposed a stochastic optimization algorithm for DCCA via non-linear orthogonal iterations (DCCA_NOI). Relying on the solution for DCCA (§4.1), we develop a new optimization algorithm for DPCCA in §4.2.

4.1 Optimization of DCCA

The DCCA optimization from Wang et al. (2015b), fully provided in Algorithm 1, relies on three key steps. First, the estimation of the covariance matrices in the form of \( \hat{\Sigma}_{FF} \) at time \( t \) is calculated by a moving average over the minibatches:

\[
\hat{\Sigma}_{FF} = \rho \hat{\Sigma}_{FF,t-1} + (1 - \rho)(\frac{|b_t|}{N-1})^{-1} F(X_{b_t}) F(X_{b_t})^T. \quad (12)
\]

\( b_t \) is the minibatch at time \( t \), \( X_{b_t} \) is the current input matrix at time \( t \), and \( \rho \in [0, 1] \) controls the ratio between the overall covariance estimation and the covariance estimation of the current minibatch.⁸ This step eliminates the need of estimating the covariances over all training data, as well as the inherent bias when the estimate relies only on the current minibatch.

Second, the DCCA_NOI algorithm forces the whitening constraints to hold by performing an explicit matrix transformation in the form of:

\[
F(X_{b_t}) = \hat{\Sigma}_{FF}^{-\frac{1}{2}} F(X_{b_t}). \quad (13)
\]

According to Horn et al. (1988), if \( \rho = 0 \):

\[
(\frac{|b_t|}{N-1})^{-1} F(X_{b_t}) F(X_{b_t})^T = I. \quad (14)
\]

Finally, in order to optimize the DCCA objective (see Eq. (2)), the weights of the two DNNs are decoupled: i.e., the objective is disassembled into two separate mean-squared error objectives. Instead of trying to bring \( F(X_{b_t}) \) and \( G(Y_{b_t}) \) closer in one gradient descent step, two steps are performed: one of the views is fixed, and a gradient step over the other is performed, and so on, iteratively. The final objective functions at each time step are:

\[
\min_{W_F} \frac{1}{|b_t|} \| F(X_{b_t}) - G(Y_{b_t}) \|^2_F. \quad (15)
\]
\[
\min_{V_G} \frac{1}{|b_t|} \| G(Y_{b_t}) - F(X_{b_t}) \|^2_F. \quad (16)
\]

Wang et al. (2015b) show that the projection matrices \( W \) and \( V \) converge to the exact solutions of CCA as \( t \to \infty \) when considering linear CCA.

4.2 Optimization of DPCCA

Our DPCCA optimization is based on the DCCA_NOI algorithm with several adjustments. Besides the requirement to obtain the sample covariances \( \hat{\Sigma}_{FF} \) and \( \hat{\Sigma}_{GG} \), when calculating the conditional variables \( F(X|Z) \), \( G(Y|Z) \), \( \hat{\Sigma}_{FF|Z} \) and \( \hat{\Sigma}_{GG|Z} \), we additionally have to obtain the stochastic estimators \( \hat{\Sigma}_{FZ} \), \( \hat{\Sigma}_{GZ} \) and \( \hat{\Sigma}_{ZZ} \). To this end, we use the moving average estimation from Eq. (12). Next, we define the whitening transformation on the residuals:

Algorithm 1 The non-linear orthogonal iterations (NOI) algorithm for DCCA (DCCA_NOI)

Input: Data matrices \( X \in \mathbb{R}^{D_x \times N}, Y \in \mathbb{R}^{D_y \times N} \), time constant \( \rho \), learning rate \( \eta \).

initialization: Initialize weights (\( W_F, V_G \)). Randomly choose a minibatch (\( X_{b_0}, Y_{b_0} \)). Initialize covariances:

\[
\hat{\Sigma}_{FF} = \frac{1}{|b_0|} F(X_{b_0}) F(X_{b_0})^T \]
\[
\hat{\Sigma}_{GG} = \frac{1}{|b_0|} G(Y_{b_0}) G(Y_{b_0})^T \]

for \( t = 1, 2, \ldots, n \) do

Randomly choose a minibatch (\( X_{b_t}, Y_{b_t} \)).

Update covariances:

\[
\hat{\Sigma}_{FF} \leftarrow \rho \hat{\Sigma}_{FF,t-1} + (1 - \rho)(\frac{|b_t|}{N-1})^{-1} F(X_{b_t}) F(X_{b_t})^T \]
\[
\hat{\Sigma}_{GG} \leftarrow \rho \hat{\Sigma}_{GG,t-1} + (1 - \rho)(\frac{|b_t|}{N-1})^{-1} G(Y_{b_t}) G(Y_{b_t})^T \]

Fix \( G(Y_{b_t}) = \sum_{b_t}^2 G(Y_{b_t}) \), and compute \( \nabla W_F \) with respect to:

\[
\min_{W_F} \frac{1}{|b_t|} \| F(X_{b_t}) - G(Y_{b_t}) \|^2_F.
\]

Update parameters:

\[
W_F \leftarrow W_F - \eta \nabla W_F
\]

Fix \( F(X_{b_t}) = \sum_{b_t}^2 F(X_{b_t}) \), and compute \( \nabla V_G \) with respect to:

\[
\min_{V_G} \frac{1}{|b_t|} \| G(Y_{b_t}) - F(X_{b_t}) \|^2_F.
\]

Update parameters:

\[
V_G \leftarrow V_G - \eta \nabla V_G
\]

end for

Output: \( (W_F, V_G) \)
As before, the whitening constraints hold when \( \rho = 0 \). From here, we derive our two final objective functions over the residuals at time \( t \):

\[
\min_{W_F} \frac{1}{|b_t|} \left\| F(X_{b_t}|Z_{b_t}) - G(Y_{b_t}|Z_{b_t}) \right\|^2_f. \tag{19}
\]

\[
\min_{V_G} \frac{1}{|b_t|} \left\| G(Y_{b_t}|Z_{b_t}) - F(X_{b_t}|Z_{b_t}) \right\|^2_f. \tag{20}
\]

Equivalently to Eq. (15)-(16) that replace Eq. (2), Eq. (19)-(20) replace Eq. (8) by performing stochastic, decoupled, and unconstrained steps. As our algorithm performs CCA over the residuals, we gain the same guarantees as Wang et al. (2015b), now for the projection matrices of the residuals.

Algorithm 2 shows the full optimization procedure for the more complex DPCCA Variant B. The full algorithm for Variant A is provided in the supplementary material. The main difference is that with Variant B we replace \( Z \) with \( H(Z) \) in all equations where it appears, and we optimize over \( \hat{U}_H \) along with \( W_F \) and \( V_G \) in Eq. (19) and Eq. (20), respectively.

5 Tasks and Data

Cross-lingual Image Description Retrieval

The cross-lingual image description retrieval task is formulated as follows: taking an image description as a query in the source language, the system has to retrieve a set of relevant descriptions in the target language which describe the same image. Our evaluation assumes a single-best scenario, where only a single target description is relevant for each query. In addition, in our setup, images are not available during inference: retrieval is performed based solely on text queries. This enables a fair comparison between our model and many baseline models that cannot represent images and text in a shared space. Moreover, it allows us to test our model in the realistic setup where images are not available at test time. To avoid the use of images at retrieval time with DPCCA, we perform the retrieval on \( F(X) \) and \( G(Y) \), rather than on \( F(X|Z) \) and \( G(Y|Z) \) (see §3.2).

We use the Multi30K dataset (Elliott et al., 2016), originated from Flickr30K (Young et al., 2014) that is comprised of Flickr images described with 1-5 English descriptions per image. Multi30K adds two English versions of the Multi30K dataset (Elliott et al., 2016), namely the English descriptions per image. Multi30K adds the 999 English (EN)

\[
F(X_{b_t}|Z_{b_t}) = \Sigma^{-\frac{1}{2}}_{F_{F_{\mathcal{G}}|Z}} F(X_{b_t}|Z_{b_t}), \tag{17}
\]

\[
G(Y_{b_t}|Z_{b_t}) = \Sigma^{-\frac{1}{2}}_{G_{F_{\mathcal{G}}}|Z} G(Y_{b_t}|Z_{b_t}). \tag{18}
\]

Algorithm 2 The non-linear orthogonal iterations (NOI) algorithm for DPCCA Variant B

Input: Data matrices \( X \in \mathbb{R}^{D_x \times N}, Y \in \mathbb{R}^{D_y \times N}, Z \in \mathbb{R}^{D_z \times N} \), time constant \( \rho \), learning rate \( \eta \).

Initialization: Initialize weights \((W_F, V_G, U_H)\). Randomly choose a minibatch \((X_{b_t}, Y_{b_t}, Z_{b_t})\).

Initialize covariances:

\[
\hat{\Sigma}_{FF} \leftarrow \frac{N}{N-1} F(X_{b_t})F(X_{b_t})^T,
\]

\[
\hat{\Sigma}_{GG} \leftarrow \frac{N}{N-1} G(Y_{b_t})G(Y_{b_t})^T,
\]

\[
\hat{\Sigma}_{HH} \leftarrow \frac{N}{N-1} H(Z_{b_t})H(Z_{b_t})^T,
\]

\[
\hat{\Sigma}_{FH} \leftarrow \frac{N}{N-1} F(X_{b_t})H(Z_{b_t})^T,
\]

\[
\hat{\Sigma}_{GH} \leftarrow \frac{N}{N-1} G(Y_{b_t})H(Z_{b_t})^T.
\]

for \( t = 1, 2, \ldots, n \) do

Randomly choose a minibatch \((X_{b_t}, Y_{b_t}, Z_{b_t})\).

Update covariances:

\[
\hat{\Sigma}_{FF} \leftarrow \rho \hat{\Sigma}_{FF} + (1 - \rho) \frac{N}{N-1} F(X_{b_t})F(X_{b_t})^T,
\]

\[
\hat{\Sigma}_{GG} \leftarrow \rho \hat{\Sigma}_{GG} + (1 - \rho) \frac{N}{N-1} G(Y_{b_t})G(Y_{b_t})^T,
\]

\[
\hat{\Sigma}_{HH} \leftarrow \rho \hat{\Sigma}_{HH} + (1 - \rho) \frac{N}{N-1} H(Z_{b_t})H(Z_{b_t})^T,
\]

\[
\hat{\Sigma}_{FH} \leftarrow \rho \hat{\Sigma}_{FH} + (1 - \rho) \frac{N}{N-1} F(X_{b_t})H(Z_{b_t})^T,
\]

\[
\hat{\Sigma}_{GH} \leftarrow \rho \hat{\Sigma}_{GH} + (1 - \rho) \frac{N}{N-1} G(Y_{b_t})H(Z_{b_t})^T.
\]

Update conditional variables:

\[
F|H \leftarrow F(X_{b_t}) - \hat{\Sigma}_{FH} \hat{\Sigma}_{HH}^{-1} H(Z_{b_t}),
\]

\[
G|H \leftarrow G(Y_{b_t}) - \hat{\Sigma}_{GH} \hat{\Sigma}_{HH}^{-1} H(Z_{b_t}),
\]

\[
\hat{\Sigma}_{FF}|H \leftarrow \hat{\Sigma}_{FF} - \hat{\Sigma}_{FH} \hat{\Sigma}_{HH}^{-1} \hat{\Sigma}_{FH}^T,
\]

\[
\hat{\Sigma}_{GG}|H \leftarrow \hat{\Sigma}_{GG} - \hat{\Sigma}_{GH} \hat{\Sigma}_{HH}^{-1} \hat{\Sigma}_{GH}^T.
\]

Fix \( \hat{G}|H = \hat{\Sigma}_{GG}|H \hat{G}|H \) and compute \( \nabla W_F, \nabla U_H \) with respect to:

\[
\min_{W_F, U_H} \frac{1}{|b_t|} \left\| F|H - \hat{G}|H \right\|^2_f.
\]

Update parameters:

\[
W_F \leftarrow W_F - \eta \nabla W_F, U_H \leftarrow U_H - \eta \nabla U_H.
\]

Fix \( F|H = \hat{\Sigma}_{FF}|H F|H \) and compute \( \nabla V_G, \nabla U_H \) with respect to:

\[
\min_{V_G, U_H} \frac{1}{|b_t|} \left\| G|H - \hat{F}|H \right\|^2_f.
\]

Update parameters:

\[
V_G \leftarrow V_G - \eta \nabla V_G, U_H \leftarrow U_H - \eta \nabla U_H.
\]

end for

Output: \((W_F, V_G, U_H)\)

Multilingual Word Similarity

The word similarity task tests the correlation between automatic and human generated word similarity scores. We evaluate with the Multilingual SimLex-999 dataset (Leviant and Reichart, 2015): the 999 English (EN) German descriptions to a total of 30,014 images: most were written independently of the English descriptions, while some are direct translations. Each image is associated with one English and one German description. We rely on the original Multi30K splits with 29,000, 1,014, and 1,000 triplets for training, validation, and test, respectively.
Table 1: WIW statistics: the number of WIW entries across POS classes in each language pair. The numbers of words per POS class are not summed to the total number of words as other (less frequent) POS tags are also represented.

<table>
<thead>
<tr>
<th></th>
<th>EN-DE</th>
<th>EN-IT</th>
<th>EN-RU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nouns</td>
<td>4606</td>
<td>4735</td>
<td>4106</td>
</tr>
<tr>
<td>Adjectives</td>
<td>405</td>
<td>416</td>
<td>348</td>
</tr>
<tr>
<td>Verbs</td>
<td>392</td>
<td>400</td>
<td>227</td>
</tr>
<tr>
<td>Adverbs</td>
<td>167</td>
<td>161</td>
<td>142</td>
</tr>
<tr>
<td>Prepositions</td>
<td>12</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>5598</td>
<td>5740</td>
<td>4838</td>
</tr>
</tbody>
</table>

We introduce a new dataset termed Word-Image-Word (WIW), which we use to train word-level models for the multilingual word similarity task. WIW contains three bilingual lexicons (EN-DE, EN-IT, EN-RU) with images shared between words in a lexicon entry. Each WIW entry is a triplet: an English word, its translation in DE/IT/RU, and a set of images relevant to the pair.

English words were taken from the January 2017 Wikipedia dump. After removing stop words and punctuation, we extract the 6,000 most frequent words from the cleaned corpus not present in SimLex. DE/IT/RU words were obtained semi-automatically from the EN words using Google Translate. The images are crawled from the Bing search engine using MMFeat\(^9\) (Kiela, 2016) by querying the EN words only. Following the suggestions from the study of Kiela et al. (2016), we save the top 20 images as relevant images.\(^{10}\)

Table 1 provides a summary of the WIW dataset. The dataset contains both concrete and abstract words, and words of different POS tags.\(^{11}\) This property has an influence on the image collection: similar to Kiela et al. (2014), we have noticed that images of more concrete concepts are less dispersed (see also examples from Figure 2).

6 Experimental Setup

Data Preprocessing and Embeddings For the sentence-level task, all descriptions were lower-cased and tokenized. Each sentence is represented with one vector: the average of its word embeddings. For English, we rely on 500-dimensional English skip-gram word embeddings (Mikolov et al., 2013) trained on the January 2017 Wikipedia dump with bag-of-words contexts (window size of 5). For German we use the deWaC 1.7B corpus (Baroni et al., 2009) to obtain 500-dimensional German embeddings using the same word embedding model. For word similarity, to be directly comparable to previous work, we rely on 300-dim word vectors in EN, DE, IT, and RU from Mrkšić et al. (2017).

Visual features are extracted from the penultimate layer (FC7) of the VGG-19 network (Simonyan and Zisserman, 2015), and compressed to the dimensionality of the textual inputs by a Principal Component Analysis (PCA) step. For the word similarity task, we average the visual vectors across all images of each word pair as done in, e.g., (Vulić et al., 2016), before the PCA step.

Baseline Models We consider a wide variety of multi-view CCA-based baselines. First, we compare against the original (linear) CCA model (Hotelling, 1936), and its deep non-linear extension DCCA (Andrew et al., 2013). For DCCA: 1) we rely on its improved optimization algorithm from Wang et al. (2015a) which uses a stochastic approach with large minibatches; 2) we compare against the DCCA NOI variant (Wang et al., 2015b) described by Algorithm 1, and another recent DCCA variant with the optimization algorithm based on a stochastic decorrelational loss (Chang et al., 2017) (DCCA SDL); and 3) we also test the DCCA Autoencoder model (DCCAE) (Wang et al., 2015a), which offers a trade-off between maximizing the canonical correlation of two sets of variables and finding informative features for their reconstruction.

Another baseline is Generalized CCA (GCCA) (Funaki and Nakayama, 2015; Horst, 1961; Rastogi et al., 2015): a linear model which extends CCA to

\(^9\)https://github.com/douwekiela/mmfeat.

\(^{10}\)Offensive words and images are manually cleaned.

\(^{11}\)POS tag information is taken from the NLTK toolkit for the English words.
three or more views. Unlike PCCA, GCCA does not condition two variables on the third shared one, but rather seeks to maximize the canonical correlations of all pairs of views. We also compare to Non-parametric CCA (NCCA) (Michaeli et al., 2016), and to a probabilistic variant of PCCA (PPCCA, Mukuta and Harada (2014)).

Finally, we compare with the two recent models which operate in the setup most similar to ours: 1) Bridge Correlational Networks (BCN) (Rajendran et al., 2016); and 2) Image Pivoting (IMG_PIVOT) from Gella et al. (2017). For both models, we report results only with the strongest variant based on the findings from the original papers, also verified by additional experimentation in our work. 12

Hyperparameter Tuning The hyperparameters of the different models are tuned with a grid search over the following values: \{2,3,4,5\} for number of layers, \{tanh, sigmoid, ReLU\} as the activation functions (we use the same activation function in all the layers of the same network), \{64,128,256\} for minibatch size, \{0.001,0.0001\} for learning rate, and \{128,256\} for L (the size of the output vectors). The dimensions of all mid-layers are set to the input size. We use the Adam optimizer (Kingma and Ba, 2015), with the number of epochs set to 300.

For all participating models, we report test performance of the best hyperparameter on the validation set. For word similarity, following a standard practice (Levy et al., 2015; Vulić et al., 2017) we tune all models on one half of the SimLex data and evaluate on the other half, and vice versa. The reported score is the average of the two halves. Similarity scores for all tasks were computed using the cosine similarity measure.

7 Results and Discussion

Cross-lingual Image Description Retrieval We report two standard evaluation metrics: 1) Recall at 1 (R@1) scores, and 2) the sentence-level BLEU+1 metric (Lin and Och, 2004), a variant of BLEU which smooths terms for higher-order n-grams, making it more suitable for evaluating short sentences. The scores for the retrieval task with all models are summarized in Table 2.

<table>
<thead>
<tr>
<th>Model</th>
<th>EN→DE</th>
<th>DE→EN</th>
<th>EN→DE</th>
<th>DE→EN</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCCA (Wang et al., 2015a)</td>
<td>0.619</td>
<td>0.621</td>
<td>0.750</td>
<td>0.754</td>
</tr>
<tr>
<td>DCCA (Wang et al., 2015b)</td>
<td>0.654</td>
<td>0.542</td>
<td>0.607</td>
<td>0.598</td>
</tr>
<tr>
<td>IMG_PIVOT (Gella et al., 2017)</td>
<td>0.772</td>
<td>0.763</td>
<td>0.789</td>
<td>0.781</td>
</tr>
<tr>
<td>BCN (Rajendran et al., 2016)</td>
<td>0.579</td>
<td>0.570</td>
<td>0.628</td>
<td>0.626</td>
</tr>
<tr>
<td>DPCCA (Variant A)</td>
<td>0.795</td>
<td>0.779</td>
<td>0.836</td>
<td>0.827</td>
</tr>
<tr>
<td>DPCCA (Variant B)</td>
<td>0.809</td>
<td>0.794</td>
<td>0.848</td>
<td>0.839</td>
</tr>
<tr>
<td>DCCA (NCCA) (Michaeli et al., 2016)</td>
<td>0.826</td>
<td>0.788</td>
<td>0.849</td>
<td>0.830</td>
</tr>
<tr>
<td>DCCA_NOI (Wang et al., 2015a)</td>
<td>0.812</td>
<td>0.788</td>
<td>0.849</td>
<td>0.830</td>
</tr>
<tr>
<td>DCCA_NOI (Wang et al., 2015b)</td>
<td>0.507</td>
<td>0.487</td>
<td>0.552</td>
<td>0.533</td>
</tr>
<tr>
<td>GCCA (Funaki and Nakayama, 2015)</td>
<td>0.699</td>
<td>0.690</td>
<td>0.742</td>
<td>0.743</td>
</tr>
<tr>
<td>NCCA (Michaeli et al., 2016)</td>
<td>0.157</td>
<td>0.165</td>
<td>0.205</td>
<td>0.213</td>
</tr>
<tr>
<td>PPCCA (Mukuta and Harada, 2014)</td>
<td>0.035</td>
<td>0.050</td>
<td>0.063</td>
<td>0.086</td>
</tr>
</tbody>
</table>

Table 2: Results on cross-lingual image description retrieval. NN-based models are above the dashed line. Best overall results are in bold. Best results with non-deep models are underlined.

The results clearly demonstrate the superiority of DPCCA (with a slight advantage to the more complex Variant B) and of the concatenation of their representation with that of the DCCA_NOI (strongest) baseline. Furthermore, the non-deep, linear PCCA achieves strong results: it outscores all non-deep models, as well as all deep models except from DCCA_NOI, IMG_PIVOT in one case, and its deep version: DPCCA. This emphasizes our contribution in proposing PCCA for multilingual processing with images as a cross-lingual bridge.

The results suggest that: 1) the inclusion of visual information in the training process helps the retrieval task even without such information during inference. DPCCA outscores all DCCA variants (either alone or through a concatenation with the DCCA_NOI representation), and PCCA outscores the original two-view CCA model; and 2) deep, non-linear architectures are useful: our DPCCA outperforms the linear PCCA model.

We also note clear improvements over the two recent models which also rely on visual information: IMG_PIVOT and BCN. The gain over IMG_PIVOT is observed despite the fact that IMG_PIVOT is a more complex multi-modal model which relies on RNNs, and is tailored to sentence-level tasks. Finally, the scores from Table 2 suggest that improved performance can be achieved by an ensemble model, that is, a simple concatenation of DPCCA (B) and DCCA_NOI.

Multilingual Word Similarity The results, presented as standard Spearman’s rank correlation scores, are summarized in Table 3: we present fine-grained results over different POS classes for EN and DE, and compare them to the results from

---

12 More details about preprocessing and baselines (including all links to their code), are in the supplementary material. We use original readily available implementations of all baselines whenever this is possible, and our in-house implementations for baselines for which no code is provided by the original authors.
Table 3: Results on EN and DE SimLex-999 (POS-based evaluation). All scores are Spearman’s rank correlations. INIT_EMB refers to initial pre-trained monolingual word embeddings (see §6).

<table>
<thead>
<tr>
<th>Model</th>
<th>EN-DE</th>
<th>EN-IT</th>
<th>EN-RU</th>
</tr>
</thead>
<tbody>
<tr>
<td>DPCCA (Variant A)</td>
<td>0.640</td>
<td>0.369</td>
<td>0.430</td>
</tr>
<tr>
<td>DPCCA (Variant B)</td>
<td>0.626</td>
<td>0.316</td>
<td>0.369</td>
</tr>
<tr>
<td>DCCA_NOI (Wang et al., 2015b)</td>
<td>0.611</td>
<td>0.308</td>
<td>0.361</td>
</tr>
<tr>
<td>DCCA (Wang et al., 2015a)</td>
<td>0.618</td>
<td>0.261</td>
<td>0.327</td>
</tr>
<tr>
<td>PCCA (Rao, 1969)</td>
<td>0.614</td>
<td>0.296</td>
<td>0.340</td>
</tr>
<tr>
<td>CCA (Hotelling, 1936)</td>
<td>0.557</td>
<td>0.297</td>
<td>0.321</td>
</tr>
<tr>
<td>GCCA (Funaki and Nakayama, 2015)</td>
<td>0.636</td>
<td>0.280</td>
<td>0.378</td>
</tr>
<tr>
<td>INIT_EMB</td>
<td>0.582</td>
<td>0.160</td>
<td>0.306</td>
</tr>
</tbody>
</table>

Table 4: Results (Spearman rank correlation) of our models and the strongest baselines on Multilingual SimLex-999 (all data).

<table>
<thead>
<tr>
<th>Model</th>
<th>EN-DE WIW</th>
<th>EN-IT WIW</th>
<th>EN-RU WIW</th>
</tr>
</thead>
<tbody>
<tr>
<td>DPCCA (A)</td>
<td>0.390</td>
<td>0.429</td>
<td>0.404</td>
</tr>
<tr>
<td>DPCCA (B)</td>
<td>0.405</td>
<td>0.413</td>
<td>0.413</td>
</tr>
<tr>
<td>PCCA</td>
<td>0.374</td>
<td>0.370</td>
<td>0.374</td>
</tr>
<tr>
<td>DCCA_NOI</td>
<td>0.390</td>
<td>0.413</td>
<td>0.407</td>
</tr>
<tr>
<td>GCCA</td>
<td>0.395</td>
<td>0.414</td>
<td>0.412</td>
</tr>
<tr>
<td>INIT_EMB</td>
<td>0.321</td>
<td>0.321</td>
<td>0.321</td>
</tr>
</tbody>
</table>

We addressed the problem of utilizing images as a bridge between languages to learn improved bilingual text representations. Our main contribution is two-fold. First, we proposed to use the Partial CCA (PCCA) method. In addition, we proposed a stochastic optimization algorithm for the deep version of PCCA that overcomes the challenges posed by the covariance estimation required by the method. Our experiments reveal the effectiveness of these methods for both sentence-level and word-level tasks. Crucially, our proposed solution does not require access to images at inference/test time, in line with the realistic scenario where images that describe sentential queries are not readily available.

In future work we plan to improve our methods by exploiting the internal structure of images and sentences as well as by effectively integrating signals from more than two languages.

Acknowledgments

IV is supported by the ERC Consolidator Grant LEXICAL: Lexical Acquisition Across Languages (no 648909). GR and RR are supported by the Infomedia Magnet Grant and by an AOL grant on “connected experience technologies”.
References


