# 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.

*Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, pages 910–921
Melbourne, Australia, July 15 - 20, 2018. ©2018 Association for Computational Linguistics

the “visual view”). 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 nonlinear 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 nonlinear 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 minibatches 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 $\boldsymbol{x}$ and $\boldsymbol{y}$ in two languages and an image $\boldsymbol{z}$ 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 $\boldsymbol{z}$ 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 $\boldsymbol{X} \in \mathbb{R}^{D_x \times N}$ and $\boldsymbol{Y} \in \mathbb{R}^{D_y \times N}$, where $D_x$ and $D_y$ are input vector dimensionalities, 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 $\boldsymbol{f} : \mathbb{R}^{D_x \times N} \to \mathbb{R}^{D'_x \times N}$ and $\boldsymbol{g} : \mathbb{R}^{D_y \times N} \to \mathbb{R}^{D'_y \times N}$ for $\boldsymbol{X}$ and $\boldsymbol{Y}$. $D'_x$ 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 $\boldsymbol{X}$ and $\boldsymbol{Y}$ into a shared $L$-dimensional ($1 \leq L \leq min(D'_x, D'_y)$) space such that the canonical correlation of the final outputs $\boldsymbol{F}(\boldsymbol{X}) = \boldsymbol{W^T}\boldsymbol{f}(\boldsymbol{X})$ and $\boldsymbol{G}(\boldsymbol{Y}) = \boldsymbol{V^T}\boldsymbol{g}(\boldsymbol{Y})$ is maximized. $\boldsymbol{W} \in \mathbb{R}^{D'_x \times L}$ and $\boldsymbol{V} \in \mathbb{R}^{D'_y \times L}$ are projection matrices: they project the final outputs of the DNNs to the shared space. $\boldsymbol{W_f}$ and $\boldsymbol{V_g}$ (the parameters of $\boldsymbol{f}$ and $\boldsymbol{g}$) and the projection matrices are the model parameters: $\boldsymbol{W_F} = \{\boldsymbol{W_f}, \boldsymbol{W}\}$; $\boldsymbol{V_G} = \{\boldsymbol{V_g}, \boldsymbol{V}\}$.[2] Formally, the DCCA objective can be written as:

$$\begin{aligned} &\max_{\boldsymbol{W_F}, \boldsymbol{V_G}} Tr(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{FG}}) \\ &\text{so that } \hat{\boldsymbol{\Sigma}}_{\boldsymbol{FF}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{GG}} = \boldsymbol{I}. \end{aligned} \tag{1}$$

$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{FG}} \equiv \frac{1}{N-1}\boldsymbol{F}(\boldsymbol{X})\boldsymbol{G}(\boldsymbol{Y})^{\boldsymbol{T}}$ is the estimation of the cross-covariance matrix of the outputs, and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{FF}} \equiv \frac{1}{N-1}\boldsymbol{F}(\boldsymbol{X})\boldsymbol{F}(\boldsymbol{X})^{\boldsymbol{T}}$, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{GG}} \equiv \frac{1}{N-1}\boldsymbol{G}(\boldsymbol{Y})\boldsymbol{G}(\boldsymbol{Y})^{\boldsymbol{T}}$ are the estimations of the autocovariance 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:

$$\begin{aligned} &\min_{\boldsymbol{W_F}, \boldsymbol{V_G}} \frac{1}{N-1}\|\boldsymbol{F}(\boldsymbol{X}) - \boldsymbol{G}(\boldsymbol{Y})\|_F^2 \\ &s.t.\ \hat{\boldsymbol{\Sigma}}_{\boldsymbol{FF}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{GG}} = \boldsymbol{I}. \end{aligned} \tag{2}$$

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

$$\boldsymbol{F}(\boldsymbol{X}) = \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{F}(\boldsymbol{X}|\boldsymbol{Z}), \tag{3}$$

$$\boldsymbol{G}(\boldsymbol{Y}) = \boldsymbol{B}\boldsymbol{Z} + \boldsymbol{G}(\boldsymbol{Y}|\boldsymbol{Z}). \tag{4}$$

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

[3] The 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 $\boldsymbol{f}(\boldsymbol{X}) = id_X$, $\boldsymbol{g}(\boldsymbol{Y}) = id_Y$; $id$ is the identity mapping.

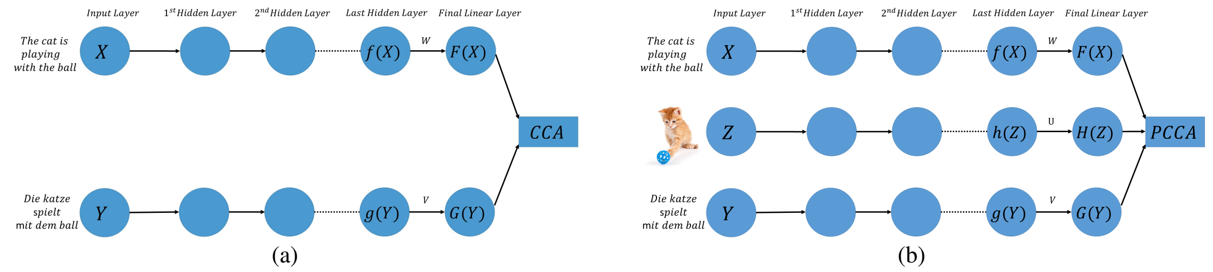

Figure 1: DCCA and DPCCA architectures. **(a)**: DCCA. $\boldsymbol{X}$ and $\boldsymbol{Y}$ (English and German image descriptions) are fed through two identical deep feed-forward neural networks followed by a final linear layer. The final nodes of the networks $\boldsymbol{F}(\boldsymbol{X})$ and $\boldsymbol{G}(\boldsymbol{Y})$ are then maximally correlated via the CCA objective. **(b)**: DPCCA. In addition, a third (shared) variable $\boldsymbol{Z}$ (an image) is either optimized via an identical architecture of the two main views (DPCCA Variant B, illustrated here) or kept fixed (DPCCA Variant A). The final nodes of the networks $\boldsymbol{F}(\boldsymbol{X})$ and $\boldsymbol{G}(\boldsymbol{Y})$ are maximally correlated conditioned on the final node in the middle network $\boldsymbol{H}(\boldsymbol{Z})$ (or directly on the input node $\boldsymbol{Z}$ in DPCCA Variant A).

$\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{L \times D_z}$ are matrices of coefficients, and $\boldsymbol{F}(\boldsymbol{X}|\boldsymbol{Z}), \boldsymbol{G}(\boldsymbol{Y}|\boldsymbol{Z}) \in \mathbb{R}^{L \times N}$ are normal random error matrices: residuals. We then minimize the mean-squared error regression criterion:

$$\min_{\boldsymbol{A}} \frac{1}{N-1} \|\boldsymbol{F}(\boldsymbol{X}) - \boldsymbol{A}\boldsymbol{Z}\|_F^2, \tag{5}$$

$$\min_{\boldsymbol{B}} \frac{1}{N-1} \|\boldsymbol{G}(\boldsymbol{Y}) - \boldsymbol{B}\boldsymbol{Z}\|_F^2. \tag{6}$$

After obtaining the optimal solutions for the coefficients, $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$, the residuals are as follows:

$$\begin{aligned} \boldsymbol{F}(\boldsymbol{X}|\boldsymbol{Z}) &= \boldsymbol{F}(\boldsymbol{X}) - \hat{\boldsymbol{A}}\boldsymbol{Z} \\ &= \boldsymbol{F}(\boldsymbol{X}) - \hat{\boldsymbol{\Sigma}}_{\boldsymbol{FZ}} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{ZZ}}^{-1} \boldsymbol{Z}. \end{aligned} \tag{7}$$

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

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

$$\begin{aligned} &\min_{\boldsymbol{W_F}, \boldsymbol{V_G}} \frac{1}{N-1} \|\boldsymbol{F}(\boldsymbol{X}|\boldsymbol{Z}) - \boldsymbol{G}(\boldsymbol{Y}|\boldsymbol{Z})\|_F^2 \\ &\quad s.t.\ \hat{\boldsymbol{\Sigma}}_{\boldsymbol{FF}|\boldsymbol{Z}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{GG}|\boldsymbol{Z}} = \boldsymbol{I}. \end{aligned} \tag{8}$$

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

$$\begin{aligned} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{FF}|\boldsymbol{Z}} &\equiv \frac{1}{N-1} \boldsymbol{F}(\boldsymbol{X}|\boldsymbol{Z}) \boldsymbol{F}(\boldsymbol{X}|\boldsymbol{Z})^{\boldsymbol{T}} \\ &= \hat{\boldsymbol{\Sigma}}_{\boldsymbol{FF}} - \hat{\boldsymbol{\Sigma}}_{\boldsymbol{FZ}} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{ZZ}}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{FZ}}^{\boldsymbol{T}}. \end{aligned} \tag{9}$$

The other conditional covariance matrix $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{GG}|\boldsymbol{Z}}$ is again computed in the analogous manner, replacing $\boldsymbol{F}$ with $\boldsymbol{G}$ and $\boldsymbol{X}$ with $\boldsymbol{Y}$.[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 $\boldsymbol{F}(\boldsymbol{X})$ and $\boldsymbol{G}(\boldsymbol{Y})$ 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 $\boldsymbol{Z}$ is kept fixed; **2)** DPCCA Variant B also optimizes over $\boldsymbol{Z}$, 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, $\boldsymbol{H}(\boldsymbol{Z}) = \boldsymbol{U}^{\boldsymbol{T}}\boldsymbol{h}(\boldsymbol{Z})$, during training, where $\boldsymbol{h} : \mathbb{R}^{D_z \times N} \to \mathbb{R}^{D_{z'} \times N}$ is a DNN having the same architecture as $\boldsymbol{f}$ and $\boldsymbol{g}$. $\boldsymbol{U} \in \mathbb{R}^{D_{z'} \times L}$ is the final linear layer of $\boldsymbol{H}$, such that overall, the additional parameters of the model are $\boldsymbol{U_H} = \{\boldsymbol{U_h}, \boldsymbol{U}\}$. Instead of assuming a linear connection between $\boldsymbol{F}(\boldsymbol{X})$ and $\boldsymbol{G}(\boldsymbol{Y})$ to $\boldsymbol{Z}$, as in Variant A, we now assume that the linear connection takes place with $\boldsymbol{H}(\boldsymbol{Z})$. 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 $D'_x = D_x$, $D'_y = D_y$ and $\boldsymbol{f}(\boldsymbol{X}) = id_X, \boldsymbol{g}(\boldsymbol{Y}) = id_Y$.

[6] For Variant A, in order for $\boldsymbol{Z}$ to be on the same range of values as in $\boldsymbol{F}$ and $\boldsymbol{G}$, we pass it through the activation function of the network, $\boldsymbol{Z} = \sigma(\boldsymbol{Z})$. Due to space constraints we discuss DPCCA Variant A in the supplementary material only.

changes Eq. (3) and Eq. (4) to:[7]

$$\boldsymbol{F}(\boldsymbol{X}) = \boldsymbol{A'} \cdot \boldsymbol{H}(\boldsymbol{Z}) + \boldsymbol{F}(\boldsymbol{X}|\boldsymbol{H}(\boldsymbol{Z})), \tag{10}$$

$$\boldsymbol{G}(\boldsymbol{Y}) = \boldsymbol{B'} \cdot \boldsymbol{H}(\boldsymbol{Z}) + \boldsymbol{G}(\boldsymbol{Y}|\boldsymbol{H}(\boldsymbol{Z})). \tag{11}$$

## 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 $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{FF_t}}$ at time $t$ is calculated by a moving average over the minibatches:

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{FF_t}} \leftarrow \rho \hat{\boldsymbol{\Sigma}}_{\boldsymbol{FF_{t-1}}} + (1-\rho)\Big(\frac{|b_t|}{N-1}\Big)^{-1} \boldsymbol{F}(\boldsymbol{X_{b_t}})\boldsymbol{F}(\boldsymbol{X_{b_t}})^{\boldsymbol{T}}. \tag{12}$$

$b_t$ is the minibatch at time $t$, $\boldsymbol{X_{b_t}}$ 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:

$$\widetilde{\boldsymbol{F}(\boldsymbol{X_{b_t}})} = \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}_{\boldsymbol{FF_t}} \boldsymbol{F}(\boldsymbol{X_{b_t}}). \tag{13}$$

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

$$\Big(\frac{|b_t|}{N-1}\Big)^{-1} \widetilde{\boldsymbol{F}(\boldsymbol{X_{b_t}})}\,\widetilde{\boldsymbol{F}(\boldsymbol{X_{b_t}})}^{\boldsymbol{T}} = \boldsymbol{I}. \tag{14}$$

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

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

initialization: Initialize weights ($\boldsymbol{W_F}, \boldsymbol{V_G}$).
Randomly choose a minibatch ($\boldsymbol{X_{b_0}}, \boldsymbol{Y_{b_0}}$).
Initialize covariances:
$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{FF}} \leftarrow \frac{N-1}{|b_0|}\boldsymbol{F}(\boldsymbol{X_{b_0}})\boldsymbol{F}(\boldsymbol{X_{b_0}})^{\boldsymbol{T}}$
$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{GG}} \leftarrow \frac{N-1}{|b_0|}\boldsymbol{G}(\boldsymbol{Y_{b_0}})\boldsymbol{G}(\boldsymbol{Y_{b_0}})^{\boldsymbol{T}}$

  for $t = 1, 2, \ldots, n$ do
  Randomly choose a minibatch ($\boldsymbol{X_{b_t}}, \boldsymbol{Y_{b_t}}$).

  Update covariances:
  $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{FF}} \leftarrow \rho\hat{\boldsymbol{\Sigma}}_{\boldsymbol{FF}} + (1-\rho)\frac{N-1}{|b_t|}\boldsymbol{F}(\boldsymbol{X_{b_t}})\boldsymbol{F}(\boldsymbol{X_{b_t}})^{\boldsymbol{T}}$
  $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{GG}} \leftarrow \rho\hat{\boldsymbol{\Sigma}}_{\boldsymbol{GG}} + (1-\rho)\frac{N-1}{|b_t|}\boldsymbol{G}(\boldsymbol{Y_{b_t}})\boldsymbol{G}(\boldsymbol{Y_{b_t}})^{\boldsymbol{T}}$
  Fix $\widetilde{\boldsymbol{G}(\boldsymbol{Y_{b_t}})} = \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}_{\boldsymbol{GG}}\boldsymbol{G}(\boldsymbol{Y_{b_t}})$, and compute $\nabla\boldsymbol{W_F}$ with respect to:
  $\min_{\boldsymbol{W_F}} \frac{1}{|b_t|}\|\boldsymbol{F}(\boldsymbol{X_{b_t}}) - \widetilde{\boldsymbol{G}(\boldsymbol{Y_{b_t}})}\|^2_F$

  Update parameters:
  $\boldsymbol{W_F} \leftarrow \boldsymbol{W_F} - \eta\nabla\boldsymbol{W_F}$
  Fix $\widetilde{\boldsymbol{F}(\boldsymbol{X_{b_t}})} = \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}_{\boldsymbol{FF}}\boldsymbol{F}(\boldsymbol{X_{b_t}})$, and compute $\nabla\boldsymbol{V_G}$ with respect to:
  $\min_{\boldsymbol{V_G}} \frac{1}{|b_t|}\|\boldsymbol{G}(\boldsymbol{Y_{b_t}}) - \widetilde{\boldsymbol{F}(\boldsymbol{X_{b_t}})}\|^2_F$

  Update parameters:
  $\boldsymbol{V_G} \leftarrow \boldsymbol{V_G} - \eta\nabla\boldsymbol{V_G}$
  end for
Output: ($\boldsymbol{W_F}, \boldsymbol{V_G}$)
```

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 $\boldsymbol{F}(\boldsymbol{X_{b_t}})$ and $\boldsymbol{G}(\boldsymbol{Y_{b_t}})$ 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:

$$\min_{\boldsymbol{W_F}} \frac{1}{|b_t|}\|\boldsymbol{F}(\boldsymbol{X_{b_t}}) - \widetilde{\boldsymbol{G}(\boldsymbol{Y_{b_t}})}\|^2_F, \tag{15}$$

$$\min_{\boldsymbol{V_G}} \frac{1}{|b_t|}\|\boldsymbol{G}(\boldsymbol{Y_{b_t}}) - \widetilde{\boldsymbol{F}(\boldsymbol{X_{b_t}})}\|^2_F. \tag{16}$$

Wang et al. (2015b) show that the projection matrices $\boldsymbol{W}$ and $\boldsymbol{V}$ converge to the exact solutions of CCA as t→ $\infty$ 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 $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{FF}}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{GG}}$, when calculating the conditional variables $\boldsymbol{F}(\boldsymbol{X}|\boldsymbol{Z})$, $\boldsymbol{G}(\boldsymbol{Y}|\boldsymbol{Z})$, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{FF|Z}}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{GG|Z}}$, we additionally have to obtain the stochastic estimators $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{FZ}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{GZ}}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{ZZ}}$. To this end, we use the moving average estimation from Eq. (12). Next, we define the whitening transformation on the residuals:

[7] Note that the matrices of coefficients $\boldsymbol{A'}$ , $\boldsymbol{B'} \in \mathbb{R}^{L \times L}$.

[8] Setting $\rho$ to a high value indicates slow updates of the estimator; setting it low mostly erases the overall estimation and relies more on the current minibatch estimation.

$$\boldsymbol{F}(\widetilde{\boldsymbol{X}_{b_t}|\boldsymbol{Z}_{b_t}}) = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{FF_t|Z}}^{-\frac{1}{2}} \boldsymbol{F}(\boldsymbol{X}_{b_t}|\boldsymbol{Z}_{b_t}), \tag{17}$$

$$\boldsymbol{G}(\widetilde{\boldsymbol{Y}_{b_t}|\boldsymbol{Z}_{b_t}}) = \boldsymbol{\Sigma}_{\boldsymbol{GG_t|Z}}^{-\frac{1}{2}} \boldsymbol{G}(\boldsymbol{Y}_{b_t}|\boldsymbol{Z}_{b_t}). \tag{18}$$

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

$$\min_{\boldsymbol{W_F}} \frac{1}{|b_t|} \|\boldsymbol{F}(\boldsymbol{X}_{b_t}|\boldsymbol{Z}_{b_t}) - \boldsymbol{G}(\widetilde{\boldsymbol{Y}_{b_t}|\boldsymbol{Z}_{b_t}})\|_F^2, \tag{19}$$

$$\min_{\boldsymbol{V_G}} \frac{1}{|b_t|} \|\boldsymbol{G}(\boldsymbol{Y}_{b_t}|\boldsymbol{Z}_{b_t}) - \boldsymbol{F}(\widetilde{\boldsymbol{X}_{b_t}|\boldsymbol{Z}_{b_t}})\|_F^2. \tag{20}$$

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 $\boldsymbol{Z}$ with $\boldsymbol{H}(\boldsymbol{Z})$ in all equations where it appears, and we optimize over $\boldsymbol{U_H}$ along with $\boldsymbol{W_F}$ and $\boldsymbol{V_G}$ in Eq. (19) and Eq. (20), respectively.

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

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

initialization: Initialize weights $(\boldsymbol{W_F}, \boldsymbol{V_G}, \boldsymbol{U_H})$.
Randomly choose a minibatch $(\boldsymbol{X}_{b_0}, \boldsymbol{Y}_{b_0}, \boldsymbol{Z}_{b_0})$.
Initialize covariances:
$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{FF}} \leftarrow \frac{N-1}{|b_0|} \boldsymbol{F}(\boldsymbol{X}_{b_0}) \boldsymbol{F}(\boldsymbol{X}_{b_0})^{\boldsymbol{T}}$
$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{GG}} \leftarrow \frac{N-1}{|b_0|} \boldsymbol{G}(\boldsymbol{Y}_{b_0}) \boldsymbol{G}(\boldsymbol{Y}_{b_0})^{\boldsymbol{T}}$
$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{HH}} \leftarrow \frac{N-1}{|b_0|} \boldsymbol{H}(\boldsymbol{Z}_{b_0}) \boldsymbol{H}(\boldsymbol{Z}_{b_0})^{\boldsymbol{T}}$
$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{FH}} \leftarrow \frac{N-1}{|b_0|} \boldsymbol{F}(\boldsymbol{X}_{b_0}) \boldsymbol{H}(\boldsymbol{Z}_{b_0})^{\boldsymbol{T}}$
$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{GH}} \leftarrow \frac{N-1}{|b_0|} \boldsymbol{G}(\boldsymbol{Y}_{b_0}) \boldsymbol{H}(\boldsymbol{Z}_{b_0})^{\boldsymbol{T}}$

  for $t = 1, 2, \ldots, n$ do
  Randomly choose a minibatch $(\boldsymbol{X}_{b_t}, \boldsymbol{Y}_{b_t}, \boldsymbol{Z}_{b_t})$.
  Update covariances:
  $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{FF}} \leftarrow \rho\hat{\boldsymbol{\Sigma}}_{\boldsymbol{FF}} + (1-\rho)\frac{N-1}{|b_t|} \boldsymbol{F}(\boldsymbol{X}_{b_t}) \boldsymbol{F}(\boldsymbol{X}_{b_t})^{\boldsymbol{T}}$
  $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{GG}} \leftarrow \rho\hat{\boldsymbol{\Sigma}}_{\boldsymbol{GG}} + (1-\rho)\frac{N-1}{|b_t|} \boldsymbol{G}(\boldsymbol{Y}_{b_t}) \boldsymbol{G}(\boldsymbol{Y}_{b_t})^{\boldsymbol{T}}$
  $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{HH}} \leftarrow \rho\hat{\boldsymbol{\Sigma}}_{\boldsymbol{HH}} + (1-\rho)\frac{N-1}{|b_t|} \boldsymbol{H}(\boldsymbol{Z}_{b_t}) \boldsymbol{H}(\boldsymbol{Z}_{b_t})^{\boldsymbol{T}}$
  $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{FH}} \leftarrow \rho\hat{\boldsymbol{\Sigma}}_{\boldsymbol{FH}} + (1-\rho)\frac{N-1}{|b_t|} \boldsymbol{F}(\boldsymbol{X}_{b_t}) \boldsymbol{H}(\boldsymbol{Z}_{b_t})^{\boldsymbol{T}}$
  $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{GH}} \leftarrow \rho\hat{\boldsymbol{\Sigma}}_{\boldsymbol{GH}} + (1-\rho)\frac{N-1}{|b_t|} \boldsymbol{G}(\boldsymbol{Y}_{b_t}) \boldsymbol{H}(\boldsymbol{Z}_{b_t})^{\boldsymbol{T}}$

  Update conditional variables:
  $\boldsymbol{F|H} \leftarrow \boldsymbol{F}(\boldsymbol{X}_{b_t}) - \hat{\boldsymbol{\Sigma}}_{\boldsymbol{FH}} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{HH}}^{-1} \boldsymbol{H}(\boldsymbol{Z}_{b_t})$
  $\boldsymbol{G|H} \leftarrow \boldsymbol{G}(\boldsymbol{Y}_{b_t}) - \hat{\boldsymbol{\Sigma}}_{\boldsymbol{GH}} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{HH}}^{-1} \boldsymbol{H}(\boldsymbol{Z}_{b_t})$
  $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{FF|H}} \leftarrow \hat{\boldsymbol{\Sigma}}_{\boldsymbol{FF}} - \hat{\boldsymbol{\Sigma}}_{\boldsymbol{FH}} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{HH}}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{FH}}^{\boldsymbol{T}}$
  $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{GG|H}} \leftarrow \hat{\boldsymbol{\Sigma}}_{\boldsymbol{GG}} - \hat{\boldsymbol{\Sigma}}_{\boldsymbol{GH}} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{HH}}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{GH}}^{\boldsymbol{T}}$

  Fix $\widetilde{\boldsymbol{G|H}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{GG|H}}^{-\frac{1}{2}} \boldsymbol{G|H}$, and compute $\nabla \boldsymbol{W_F}$, $\nabla \boldsymbol{U_H}$
  with respect to:
  $\min_{\boldsymbol{W_F}, \boldsymbol{U_H}} \frac{1}{|b_t|} \|\boldsymbol{F|H} - \widetilde{\boldsymbol{G|H}}\|_F^2$
  Update parameters:
  $\boldsymbol{W_F} \leftarrow \boldsymbol{W_F} - \eta\nabla\boldsymbol{W_F}, \boldsymbol{U_H} \leftarrow \boldsymbol{U_H} - \eta\nabla\boldsymbol{U_H}$
  Fix $\widetilde{\boldsymbol{F|H}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{FF|H}}^{-\frac{1}{2}} \boldsymbol{F|H}$, and compute $\nabla \boldsymbol{V_G}$, $\nabla \boldsymbol{U_H}$
  with respect to:
  $\min_{\boldsymbol{V_G}, \boldsymbol{U_H}} \frac{1}{|b_t|} \|\boldsymbol{G|H} - \widetilde{\boldsymbol{F|H}}\|_F^2$
  Update parameters:
  $\boldsymbol{V_G} \leftarrow \boldsymbol{V_G} - \eta\nabla\boldsymbol{V_G}, \boldsymbol{U_H} \leftarrow \boldsymbol{U_H} - \eta\nabla\boldsymbol{U_H}$
  end for
Output: $(\boldsymbol{W_F}, \boldsymbol{V_G}, \boldsymbol{U_H})$
```

## 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 $\boldsymbol{F}(\boldsymbol{X})$ and $\boldsymbol{G}(\boldsymbol{Y})$, rather than on $\boldsymbol{F}(\boldsymbol{X}|\boldsymbol{Z})$ and $\boldsymbol{G}(\boldsymbol{Y}|\boldsymbol{Z})$ (see §3.2).

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

**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)

|  | EN-DE | EN-IT | EN-RU |
|---|---|---|---|
| Nouns | 4606 | 4735 | 4106 |
| Adjectives | 405 | 416 | 348 |
| Verbs | 392 | 400 | 227 |
| Adverbs | 167 | 161 | 142 |
| Prepositions | 12 | 12 | 9 |
| **Total** | **5598** | **5740** | **4838** |

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.

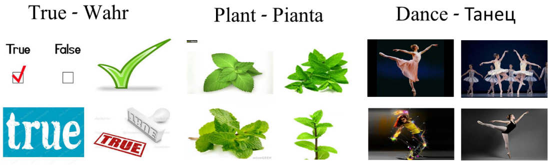

Figure 2: WIW examples from each of the three bilingual lexicons. Note that the designated words can be either abstract (*true*), express an action (*dance*) or be more concrete (*plant*).

word pairs from SimLex-999 (Hill et al., 2015) were translated to German (DE), Italian (IT), and Russian (RU), and similarity scores were crowd-sourced from native speakers.

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 lowercased 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 Nonparametric 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.

| | **R@1** | | **BLEU+1** | |
|---|---|---|---|---|
| **Model** | EN→DE | DE→EN | EN→DE | DE→EN |
| DPCCA (Variant A) | 0.795 | 0.779 | 0.836 | 0.827 |
| DPCCA (Variant B) | 0.809 | **0.794** | 0.848 | **0.839** |
| DPCCA(B)+DCCA_NOI (concat) | **0.826** | 0.791 | **0.863** | 0.837 |
| DCCA_NOI (Wang et al., 2015b) | 0.812 | 0.788 | 0.849 | 0.830 |
| DCCA_SDL (Chang et al., 2017) | 0.507 | 0.487 | 0.552 | 0.533 |
| DCCA (Wang et al., 2015a) | 0.619 | 0.621 | 0.664 | 0.673 |
| DCCAE (Wang et al., 2015a) | 0.564 | 0.542 | 0.607 | 0.598 |
| IMG_PIVOT (Gella et al., 2017) | 0.772 | 0.763 | 0.789 | 0.781 |
| BCN (Rajendran et al., 2016) | 0.579 | 0.570 | 0.628 | 0.629 |
| PCCA (Rao, 1969) | 0.785 | 0.737 | 0.825 | 0.787 |
| CCA (Hotelling, 1936) | 0.764 | 0.704 | 0.803 | 0.754 |
| GCCA (Funaki and Nakayama, 2015) | 0.699 | 0.690 | 0.742 | 0.743 |
| NCCA (Michaeli et al., 2016) | 0.157 | 0.165 | 0.205 | 0.213 |
| PPCCA (Mukuta and Harada, 2014) | 0.035 | 0.050 | 0.063 | 0.086 |

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 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.

| | English-German | | | | | |
|---|---|---|---|---|---|---|
| **Model** | EN-Adj | EN-Verbs | EN-Nouns | DE-Adj | DE-Verbs | DE-Nouns |
| DPCCA (Variant A) | **0.640** | 0.311 | 0.369 | 0.430 | **0.321** | **0.404** |
| DPCCA (Variant B) | 0.626 | **0.316** | **0.382** | **0.462** | 0.319 | 0.399 |
| DCCA_NOI (Wang et al., 2015b) | 0.611 | 0.308 | 0.361 | 0.441 | 0.297 | 0.398 |
| DCCA (Wang et al., 2015a) | 0.618 | 0.261 | 0.327 | 0.404 | 0.290 | 0.362 |
| PCCA (Rao, 1969) | 0.614 | 0.296 | 0.340 | 0.305 | 0.143 | 0.340 |
| CCA (Hotelling, 1936) | 0.557 | 0.297 | 0.321 | 0.284 | 0.157 | 0.346 |
| GCCA (Funaki and Nakayama, 2015) | 0.636 | 0.280 | 0.378 | 0.446 | 0.277 | 0.398 |
| INIT_EMB | 0.582 | 0.160 | 0.306 | 0.407 | 0.164 | 0.285 |

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).

| | EN-DE WIW | | EN-IT WIW | | EN-RU WIW | |
|---|---|---|---|---|---|---|
| **Model** | EN | DE | EN | IT | EN | RU |
| DPCCA (A) | 0.398 | **0.400** | 0.412 | **0.429** | 0.404 | **0.407** |
| DPCCA (B) | **0.405** | **0.400** | 0.413 | 0.427 | **0.413** | 0.402 |
| PCCA | 0.374 | 0.301 | 0.370 | 0.386 | 0.374 | 0.374 |
| DCCA_NOI | 0.390 | 0.398 | 0.413 | 0.422 | 0.407 | 0.398 |
| GCCA | 0.395 | 0.386 | **0.414** | 0.407 | 0.412 | 0.396 |
| INIT_EMB | 0.321 | 0.278 | 0.321 | 0.361 | 0.321 | 0.385 |

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

a selection of strongest baselines. Further, Table 4 presents results on all SimLex word pairs. The POS class result patterns for EN-IT and EN-RU are very similar to the patterns in Table 3 and are provided in the supplementary material. First, the results over the initial monolingual embeddings before training (INIT_EMB) clearly indicate that multilingual information is beneficial for the word similarity task. We observe improvements with all models (the only exception being extremely low-scoring PPCCA and NCCA, not shown). Moreover, by additionally grounding concepts from two languages in the visual modality it is possible to further boost word similarity scores. This result is in line with prior work in monolingual settings (Chrupała et al., 2015; Kiela and Bottou, 2014; Lazaridou et al., 2015), which have shown to profit from multi-modal features.

The results on the POS classes represented in SimLex-999 (nouns, verbs, adjectives, Table 3) form our main finding: conditioning the multilingual representations on a shared image leads to improvements in verb and adjective representations. While for nouns one of the DPCCA variants is the best performing model for both languages, the gaps from the best performing baselines are much smaller. This is interesting since, e.g., verbs are more abstract than nouns (Hartmann and Søgaard, 2017; Hill et al., 2014). Considering the fact that SimLex-999 consists of 666 noun pairs, 222 verb pairs and 111 adjective pairs, this is the reason that the gains of DPCCA over the strongest baselines across the entire evaluation set are more modest (Table 4). We note again that the same patterns presented in Table 3 for EN-DE – more prominent verb and adjective gains and a smaller gain on nouns – also hold for EN-IT and EN-RU (see the supplementary material).

## 8 Conclusion and Future Work

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

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. 2013. Deep canonical correlation analysis. In *Proceedings of ICML*, pages 1247–1255.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, Lawrence C. Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of ICCV*, pages 2425–2433.

Raman Arora and Karen Livescu. 2013. Multi-view CCA-based acoustic features for phonetic recognition across speakers and domains. In *Proceedings of ICASSP*, pages 7135–7139.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky Wide Web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Lawrence W. Barsalou and Katja Wiemer-Hastings. 2005. Situating abstract concepts. In D. Pecher and R. Zwaan, editors, *Grounding cognition: The role of perception and action in memory, language, and thought*, pages 129–163.

Shane Bergsma and Benjamin Van Durme. 2011. Learning bilingual lexicons using the visual similarity of labeled web images. In *Proceedings of IJCAI*, pages 1764–1769.

Elia Bruni, Nam Khanh Tram, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. 1995. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208.

Iacer Calixto and Qun Liu. 2017. Incorporating global visual features into attention-based neural machine translation. In *Proceedings of EMNLP*, pages 992–1003.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Multilingual multi-modal embeddings for natural language processing. *arXiv preprint arXiv:1702.01101*.

Sarath Chandar, Mitesh M Khapra, Hugo Larochelle, and Balaraman Ravindran. 2016. Correlational neural networks. *Neural Computation*, 28:257–285.

Xiaobin Chang, Tao Xiang, and Timothy M. Hospedales. 2017. Deep multi-view learning with stochastic decorrelation loss. *CoRR*, abs/1707.09669.

Grzegorz Chrupała, Ákos Kádár, and Afra Alishahi. 2015. Learning language through pictures. In *Proceedings of ACL*, pages 112–118.

Desmond Elliott, Stella Frank, and Eva Hasler. 2015. Multilingual image description with neural sequence models. *arXiv preprint arXiv:1510.04709*.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *Proceedings of ECCV*, pages 15–29.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of EACL*, pages 462–471.

Yansong Feng and Mirella Lapata. 2010. Visual information in semantic representation. In *Proceedings of NAACL-HLT*, pages 91–99.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of EMNLP*, pages 457–468.

Ruka Funaki and Hideki Nakayama. 2015. Image-mediated learning for zero-shot cross-lingual document retrieval. In *Proceedings of EMNLP*, pages 585–590.

Spandana Gella, Mirella Lapata, and Frank Keller. 2016. Unsupervised visual sense disambiguation for verbs using multimodal embeddings. In *Proceedings of NAACL-HLT*, pages 182–192.

Spandana Gella, Rico Sennrich, Frank Keller, and Mirella Lapata. 2017. Image pivoting for learning multilingual multimodal representations. In *Proceedings of EMNLP*, pages 2839–2845.

Goran Glavaš, Ivan Vulić, and Simone Paolo Ponzetto. 2017. If sentences could see: Investigating visual information for semantic textual similarity. In *Proceedings of IWCS*.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1–3).

Mareike Hartmann and Anders Søgaard. 2017. Limitations of cross-lingual learning from image search. *CoRR*, abs/1709.05914.

Felix Hill, Roi Reichart, and Anna Korhonen. 2014. Multi-modal models for concrete and abstract concept meaning. *Transactions of the ACL*, 2:285–296.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Julian Hitschler, Shigehiko Schamoni, and Stefan Riezler. 2016. Multimodal pivots for image caption translation. In *Proceedings of ACL*, pages 2399–2409.

Berthold K.P. Horn, Hugh M. Hilden, and Shahriar Negahdaripour. 1988. Closed-form solution of absolute orientation using orthonormal matrices. *Journal of Optical Society of America*, 5(7):1127–1135.

Paul Horst. 1961. Generalized canonical correlations and their applications to experimental data. *Journal of Clinical Psychology*, 17(4):331–347.

Harold Hotelling. 1936. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377.

Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. 2016. Attention-based multimodal neural machine translation. In *Proceedings of WMT*, pages 639–645.

Agnan Kessy, Alex Lewin, and Korbinian Strimmer. 2017. Optimal whitening and decorrelation. *The American Statistician*.

Douwe Kiela. 2016. MMFeat: A toolkit for extracting multi-modal features. In *Proceedings of ACL System Demonstrations*, pages 55–60.

Douwe Kiela and Léon Bottou. 2014. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *Proceedings of EMNLP*, pages 36–45.

Douwe Kiela, Felix Hill, Anna Korhonen, and Stephen Clark. 2014. Improving multi-modal representations using image dispersion: Why less is sometimes more. In *Proceedings of ACL*, pages 835–841.

Douwe Kiela, Anita Lilla Verő, and Stephen Clark. 2016. Comparing data sources and architectures for deep visual representation learning in semantics. In *Proceedings of EMNLP*, pages 447–456.

Douwe Kiela, Ivan Vulić, and Stephen Clark. 2015. Visual bilingual lexicon induction with transferred ConvNet features. In *Proceedings of EMNLP*, pages 148–158.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR (Conference Track)*.

Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. 2014. Multimodal neural language models. In *Proceedings of ICML*, pages 595–603.

George Lakoff and Mark Johnson. 1999. *Philosophy in the flesh: The embodied mind and its challenge to Western thought*.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining language and vision with a multimodal skip-gram model. In *Proceedings of NAACL-HLT*, pages 153–163.

Ira Leviant and Roi Reichart. 2015. Judgment language matters: Multilingual vector space models for judgment language aware lexical semantics. *CoRR, abs/1508.00106*.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the ACL*, 3:211–225.

Chin-Yew Lin and Franz Josef Och. 2004. ORANGE: A method for evaluating automatic evaluation metrics for machine translation. In *Proceedings of COLING*, pages 501–507.

Max M. Louwerse. 2011. Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, 59(1):617–645.

Ang Lu, Weiran Wang, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Deep multilingual correlation for improved word embeddings. In *Proceedings of NAACL-HLT*, pages 250–256.

Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. 2015. Deep captioning with multimodal recurrent neural networks (m-RNN). In *Proceedings of ICLR (Conference Track)*.

Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. 2014. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*.

Tomer Michaeli, Weiran Wang, and Karen Livescu. 2016. Nonparametric canonical correlation analysis. In *Proceedings of ICML*, pages 1967–1976.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of ICLR (Conference Track)*.

Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the ACL*, 5(1):309–324.

Yusuke Mukuta and Harada. 2014. Probabilistic partial canonical correlation analysis. In *Proceedings of ICML*, pages 1449–1457.

Hideki Nakayama and Noriki Nishida. 2017. Zero-resource machine translation by multimodal encoder–decoder network with multimedia pivot. *Machine Translation*, 31(1-2):49–64.

Janarthanan Rajendran, Mitesh M. Khapra, Sarath Chandar, and Balaraman Ravindran. 2016. Bridge correlational neural networks for multilingual multimodal representation learning. In *Proceedings of NAACL-HLT*, pages 171–181.

B. Raja Rao. 1969. Partial canonical correlations. *Trabajos de estadistica y de investigación operativa*, 20(2-3):211–219.

Pushpendre Rastogi, Benjamin Van Durme, and Raman Arora. 2015. Multiview LSA: Representation learning via generalized CCA. In *Proceedings of NAACL-HLT*, pages 556–566.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *Proceedings of ICLR (Workshop Track)*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of CVPR*, pages 3156–3164.

Ivan Vulić, Douwe Kiela, Stephen Clark, and Marie-Francine Moens. 2016. Multi-modal representations for improved bilingual lexicon learning. In *Proceedings of ACL*, pages 188–194. ACL.

Ivan Vulić, Roy Schwartz, Ari Rappoport, Roi Reichart, and Anna Korhonen. 2017. Automatic selection of context configurations for improved class-specific word representations. In *Proceedings of CoNLL*, pages 112–122.

Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. 2015a. On deep multi-view representation learning. In *Proceedings of ICML*, pages 1083–1092.

Weiran Wang, Raman Arora, Karen Livescu, and Nathan Srebro. 2015b. Stochastic optimization for deep CCA via nonlinear orthogonal iterations. In *Proceedings of Communication, Control, and Computing*, pages 688–695.

Huijuan Xu and Kate Saenko. 2016. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *Proceedings of ECCV*, pages 451–466.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of ICML*, pages 2048–2057.

Fei Yan and Krystian Mikolajczyk. 2015. Deep correlation for matching images and text. In *Proceedings of CVPR*, pages 3441–3450.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the ACL*, 2:67–78.