FILTERED INNER PRODUCT PROJECTION FOR MULTILINGUAL EMBEDDING ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Due to widespread interest in machine translation and transfer learning, there are numerous algorithms for mapping multiple embeddings to a shared representation space. Recently, these algorithms have been studied in the setting of bilingual lexicon induction where one seeks to align the embeddings of a source and a target language such that translated word pairs lie close to one another in a common representation space. In this paper, we propose a method, Filtered Inner Product Projection (FIPP), for mapping embeddings to a common representation space. As semantic shifts are pervasive across languages and domains, FIPP first identifies the common geometric structure in both embeddings and then, only on the common structure, aligns the Gram matrices of these embeddings. FIPP is applicable even when the source and target embeddings are of differing dimensionalities. Additionally, FIPP provides computational benefits in ease of implementation and is faster to compute than current approaches. Following the baselines in Glavaš et al. (2019), we evaluate FIPP in the context of bilingual lexicon induction and downstream language tasks. We show that FIPP outperforms existing methods on the XLING (5K) BLI dataset for most language pairs while also providing robust performance across downstream tasks.

1 INTRODUCTION

The problem of aligning sets of embeddings, or high dimensional real valued vectors, is of great interest in natural language processing, with applications in machine translation and transfer learning, and shares connections to graph matching and assignment problems. In machine translation, aligning embeddings trained on corpora from different languages has led to improved performance of supervised and unsupervised word and sentence translation (Zou et al., 2013). Additionally, aligned source and target input embeddings have been shown to improve the transfer of models learned on a source domain to a target domain Artetxe et al. (2018a); Wang et al. (2018); Mogadala & Rettinger (2016).

In the bilingual lexicon induction task, one seeks to learn a transformation on the embeddings of a source and a target language so that translated word pairs lie close to one another in the shared representation space. Specifically, one is given embeddings for a source and a target language, $X_s \in \mathbb{R}^{n \times d}$ and $X_t \in \mathbb{R}^{n \times d}$, and a small seed dictionary D containing pairs of translated words. Using this seed dictionary, a transformation is learned on X_s and X_t with the objective that unseen translation pairs can be induced, often through nearest neighbors search.

Previous literature on this topic has focused on aligning embeddings by minimizing matrix or distributional distances Grave et al. (2019); Jawanpuria et al. (2019); Joulin et al. (2018a). For instance, Mikolov et al. (2013a) proposed using the Orthogonal Procrustes solution restricted to pairs of words in the seed dictionary, $\Omega^* = \arg\min_{\Omega \in O(d)} \|X_s^D \Omega - X_t^D\|_F$, which achieves high word translation accuracy for similar languages. However, these methods usually require the dimensions of the source and target language embeddings to be the same, which often may not hold. Furthermore, due to semantic shifts across languages, it's often the case that a word and its translation may not co-occur with the same sets of words Gulordava & Baroni (2011). Therefore, seeking an alignment which minimizes all pairwise distances among translated pairs results in using information not common to both the source and target embeddings.

To address these problems, we propose Filtered Inner Product Projection (FIPP) for mapping embeddings from different languages to a shared representation space. FIPP aligns a source embedding

$X_s \in \mathbb{R}^{n \times d_1}$ to a target embedding $X_t \in \mathbb{R}^{m \times d_2}$ and maps vectors in X_s to the $\mathbb{R}^{d_2}$ space of X_t. Instead of word-level information, FIPP focuses on pairwise distance information, specified by the Gram matrices $X_s X_s^T$ and $X_t X_t^T$. During alignment, FIPP tries to achieve the following two goals. First, it is desired that the aligned source embedding FIPP$(X_s) = \tilde{X}_s \in \mathbb{R}^{n \times d_2}$ be structurally close to the original source embedding to ensure that semantic information is retained and prevent against overfitting on the seed dictionary. This goal is reflected in the minimization of the *reconstruction loss*: $\|\tilde{X}_s \tilde{X}_s^T - X_s X_s^T\|_F^2$.

Second, as the usage of words and their translations vary across languages, instead of requiring $\tilde{X}_s$ to use all of the distance information from X_t, FIPP selects a filtered set K of word pairs that have similar distances in both the source and target languages: $K = \{(i,j) \in D : \{|X_s X_s^T - X_t X_t^T|_{ij} \leq \epsilon\}$. FIPP then minimizes a *transfer loss* on this set K, the squared difference in distances between the aligned source embeddings and the target embeddings: $\sum_{(i,j) \in K} (\tilde{X}_s[i] \tilde{X}_s[j]^T - X_t[i] X_t[j]^T)^2$.

We show FIPP can be efficiently solved using either low-rank semidefinite approximations or stochastic gradient descent. Also, we formulate a least squares projection to infer aligned representations for words outside the seed dictionary and present a weighted Procrustes objective which recovers an orthogonal operator which takes into consideration the degree of structural similarity among translation pairs. The method is illustrated in Figure 1.

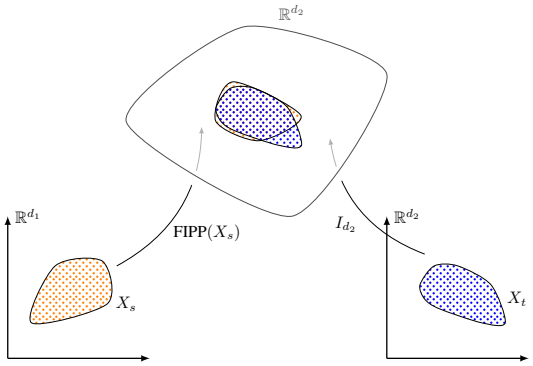

Figure 1: FIPP alignment of source and target embeddings, X_s and X_t, to a common representation space. Note X_s is modified using information from X_t and mapped to $\mathbb{R}^{d_2}$ while X_t is unchanged.

Compared to previous approaches, FIPP has improved generality, stability, and efficiency. First, since FIPP's alignment between the source and target embeddings is performed on Gram matrices, i.e. $X_s X_s^T$ and $X_t X_t^T \in \mathbb{R}^{|D| \times |D|}$, embeddings are not required to be of the same dimensions. This is particularly helpful for aligning embeddings trained on smaller corpuses, such as in low resource domains, or compute-intensive settings where embeddings may have been compressed to lower dimensions. Secondly, alignment modifications made on filtered Gram matrices can incorporate varying constraints on alignment at the most granular level, pairwise distances. Lastly, FIPP is easy to implement as it involves only matrix operations, is deterministic, and takes an order of magnitude less time to compute than either the current best supervised Joulin et al. (2018b) or unsupervised approach Artetxe et al. (2018c).

We conduct a thorough evaluation of FIPP using baselines outlined in Glavaš et al. (2019) including bilingual lexicon induction with 5K and 1K supervision sets and downstream evaluation on MNLI Natural Language Inference and Ted CLDC Document Classification tasks. The rest of this paper is organized as follows. We discuss related work in Section 2. We introduce our FIPP model in Section 3 and usage for inference in Section 4. We present experimental results in Section 5 and further discuss findings in Section 6. We conclude the paper in Section 7.

2 Related Work

2.1 Distributional Methods for Quantifying Semantic Shifts

Prior work has shown monolingual text corpora from different communities or time periods exhibit variations in semantics and syntax Hamilton et al. (2016a;b). Word embeddings Mikolov et al. (2013b); Pennington et al. (2014); Bojanowski et al. (2017) map words to representations in a continuous space with the objective that the inner product between any two words representations is approximately proportional to their co-occurrence frequency. By comparing pairwise distances in monolingual embeddings trained on separate corpora, one can quantify semantic shifts associated with biases, cultural norms, and temporal differences Gulordava & Baroni (2011); Sagi et al. (2011);

Kim et al. (2014). Recently proposed metrics on embeddings compare all pairwise inner products of two embeddings, E and F, of the form $\|EE^T - FF^T\|_F$ Yin et al. (2018). While these metrics have been applied in quantifying monolingual semantic variation, they have not been explored in context of mapping embeddings to a common representation space or in multilingual settings.

2.2 Multilingual Embedding Alignment

The first work on this topic is by Mikolov et al. (2013a) who proposed the problem of bilingual embedding alignment with the presence of an incomplete matching between words of two languages or a bilingual seed dictionary. The authors suggested using the operator recovered by the Orthogonal Procrustes problem, $\Omega^* = \arg\min_{\Omega \in O(d)} \|X_s^D \Omega - X_t^D\|_F$, and found that the approach was able to induce a matching between similar languages with $> 75\%$ accuracy. Dinu & Baroni (2015) worked on corrections to the "hubness" problem in embedding alignment, where certain word vectors may be close to many other word vectors, arising due to nonuniform density of vectors in the R^d space. Smith et al. (2017) proposed the inverted softmax metric for inducing matchings between words in embeddings of different languages. Artetxe et al. (2016) studied the impact of normalization, centering and orthogonality constraints in linear alignment functions. Jawanpuria et al. (2019) presented a composition of orthogonal operators and a Mahalanobis metric of the form $UBV^T, U, V^T \in O(d), B \succ 0$ to account for observed correlations and moment differences between embedding dimensions Søgaard et al. (2018). Joulin et al. (2018a) proposed an alignment based on neighborhood information to account for differences in density and shape of embeddings in their respective $\mathbb{R}^d$ spaces. Artetxe et al. (2018c) outlined a framework which unifies many existing alignment approaches as compositions of matrix operations such as Orthogonal mappings, Whitening, and Dimensionality Reduction. In the unsupervised setting, where a bilingual seed dictionary is not provided, approaches using adversarial learning, distributional matching, and noisy self-supervision have been used to concurrently learn a matching and an alignment between embeddings Cao et al. (2016); Zhang et al. (2017); Hoshen & Wolf (2018); Grave et al. (2019); Artetxe et al. (2017; 2018b); Alvarez-Melis & Jaakkola (2018). Discussion on unsupervised approaches has been deferred to the Appendix. Recently, progress in the bilingual lexicon induction task has been spurred by the release of various large-scale datasets and measurement tools Joulin et al. (2018a); Glavaš et al. (2019).

3 Filtered Inner Product Projection (FIPP)

In this section, we introduce Filtered Inner Product Projection (FIPP), a method for aligning embeddings in a shared representation space. The rest of this section details the method definition and is split into defining the optimization problem and approaches for finding approximate solutions.

3.1 Filtered Inner Product Projection Objective

FIPP aligns a source embedding $\mathfrak{X}_s \in \mathbb{R}^{n \times d_1}$ to a target embedding $\mathfrak{X}_t \in \mathbb{R}^{m \times d_2}$ and projects vectors in $\mathfrak{X}_s$ to $\tilde{\mathfrak{X}}_s \in \mathbb{R}^{n \times d_2}$. Let $X_s \in \mathbb{R}^{c \times d_1}$ and $X_t \in \mathbb{R}^{c \times d_2}$ be the source and target embeddings for pairs in the seed dictionary D, $|D| = c \ll min(n, m)$. FIPP's objective is to minimize a linear combination of a reconstruction loss, which regularizes changes in the pairwise inner products of the source embedding, and a transfer loss, which aligns the source and target embeddings on common portions of their geometries.

$$\min_{\tilde{X}_s \in \mathbb{R}^{c \times d_2}} \overbrace{\|\tilde{X}_s \tilde{X}_s^T - X_s X_s^T\|_F^2}^{\text{Reconstruction Loss}} + \lambda \underbrace{\|\Delta^\epsilon \circ (\tilde{X}_s \tilde{X}_s^T - X_t X_t^T)\|_F^2}_{\text{Transfer Loss}} \tag{1}$$

where $\lambda, \epsilon \in \mathbb{R}^+$ are tunable scalar hyperparameters whose effects are discussed in Section B, $\circ$ is the Hadamard product, and Δ^ϵ is a binary matrix discussed in 3.1.2.

3.1.1 Reconstruction Loss

Due to the limited, noisy supervision in our problem setting, an alignment should be regularized against overfitting. Specifically, the aligned space needs to retain a similar geometric structure to the original source embeddings; this has been enforced in previous works by ensuring that alignments

are close to orthogonal mappings Mikolov et al. (2013a); Joulin et al. (2018a); Jawanpuria et al. (2019). As $\tilde{X}_s$ and X_s can be of differing dimensionality, we check structural similarity by comparing pairwise inner products, captured by a reconstruction loss known as the PIP distance or Global Anchor Metric: $\|\tilde{X}_s\tilde{X}_s^T - X_sX_s^T\|_F^2$ Yin & Shen (2018); Yin et al. (2018).

Theorem 1. *Suppose $E \in \mathbb{R}^{n\times d}$, $F \in \mathbb{R}^{n\times d}$ are two matrices with orthonormal columns and $\Omega^* = \arg\min_{\Omega\in O(d)} \|E\Omega - F\|_F$. It follows that Yin et al. (2018):*

$$\|E\Omega^* - F\| \leq \|EE^T - FF^T\| \leq \sqrt{2}\|E\Omega^* - F\|. \tag{2}$$

This metric has been used in quantifying semantic shifts and has been proven Yin et al. (2018) to be equivalent to the residual of the Orthogonal Procrustes problem up to a small constant factor, as shown in Theorem 1. Note that the PIP distance is invariant to orthogonal operations such as rotations which are known to be present in unaligned embeddings.

3.1.2 Transfer Loss

In aligning X_s to X_t, we should seek to only utilize common geometric information between the two embedding spaces. We propose a simple approach, although FIPP can admit other forms of filtering mechanisms, denoted as inner product filtering where we find pairwise distances similar in both embedding spaces defined by a threshold ϵ. Specifically, compute a matrix $\Delta^\epsilon \in \{0,1\}^{c\times c}$ where Δ^ϵ_{ij} is an indicator on whether $|X_{s,i}X_{s,j}^T - X_{t,i}X_{t,j}^T| < \epsilon$. In this form, ϵ is a hyperparameter which determines how close pairwise distances must be in the source and target embeddings to be deemed similar. We then define a transfer loss as being the squared difference between the converted source embedding $\tilde{X}_s$ and target embedding X_t, but only on pairs of words in K: $\|\Delta^\epsilon \circ (\tilde{X}_s\tilde{X}_s^T - X_tX_t^T)\|_F^2$, where $\circ$ is the Hadamard product. The FIPP objective is a linear combination of the reconstruction and transfer losses.

3.2 Approximate Solutions to the FIPP Objective

We provide two approaches for obtaining solutions to the FIPP objective using Low-rank Semidefinite Approximations and Stochastic Gradient Descent.

3.2.1 Solutions using Low-rank Semidefinite Approximations

Denote the Gram matrices $G^s \triangleq X_sX_s^T$, $G^t \triangleq X_tX_t^T$ and $\tilde{G}^s \triangleq \tilde{X}_s\tilde{X}_s^T$.

Lemma 2. *The matrix G^* which minimizes the FIPP objective for a fixed λ and ϵ has entries:*

$$G^*_{ij} = \begin{cases} \frac{(X_sX_s^T)_{ij} + \lambda(X_tX_t^T)_{ij}}{1+\lambda}, & \text{if } (i,j) \in K \\ (X_sX_s^T)_{ij}, & \text{otherwise} \end{cases} \tag{3}$$

Proof. For a fixed λ and ϵ, $\mathcal{L}_{FIPP,\lambda,\epsilon}(\tilde{X}_s\tilde{X}_s^T)$ can be decomposed as follows:

$$\begin{aligned} \mathcal{L}_{FIPP,\lambda,\epsilon}(\tilde{X}_s\tilde{X}_s^T) =& \|\tilde{X}_s\tilde{X}_s^T - X_sX_s^T\|_F^2 + \lambda\|\Delta^\epsilon \circ (\tilde{X}_s\tilde{X}_s^T - X_tX_t^T)\|_F^2 \\ =& \sum_{i,j\in K} ((\tilde{G}^s_{ij} - G^s_{ij})^2 + \lambda(\tilde{G}^s_{ij} - G^t_{ij})^2) + \sum_{i,j\notin K} (\tilde{G}^s_{ij} - G^s_{ij})^2 \end{aligned} \tag{4}$$

By taking derivatives with respect to $\tilde{G}^s_{ij}$, the matrix G^* which minimizes $\mathcal{L}_{FIPP,\lambda,\epsilon}(\cdot)$ is:

$$G^* = \underset{\tilde{X}_s\tilde{X}_s^T \in \mathbb{R}^{c\times c}}{\arg\min}\ \mathcal{L}_{FIPP,\lambda,\epsilon}(\tilde{X}_s\tilde{X}_s^T), \quad G^*_{ij} = \begin{cases} \frac{(X_sX_s^T)_{ij} + \lambda(X_tX_t^T)_{ij}}{1+\lambda}, & \text{if } (i,j) \in K \\ (X_sX_s^T)_{ij}, & \text{otherwise} \end{cases} \tag{5}$$

□

We now have the matrix $G^* \in \mathbb{R}^{c\times c}$ which minimizes the FIPP objective. However, for G^* to be a Gram matrix, it is required that $G^* \in S_+^{c\times c}$, the set of symmetric Positive Semidefinite matrices. Additionally, to recover an $\tilde{X}_s \in \mathbb{R}^{c\times d_2}$ such that $\tilde{X}_s\tilde{X}_s^T = G^*$, we must have $Rank(G^*) \leq d_2$.

Note that G^* is symmetric by construction since the set K is commutative and G^s, G^t are symmetric. However, G^* is not necessarily positive semidefinite nor is it necessarily true that $Rank(G^*) \leq d_2$. Therefore, to recover an aligned embedding $\tilde{X}_s \in \mathbb{R}^{c \times d_2}$, we instead perform a rank-constrained semidefinite approximation to find $\min_{\tilde{X}_s \in \mathbb{R}^{c \times d_2}} \|\tilde{X}_s \tilde{X}_s^T - G^*\|_F$.

Theorem 3. *Let $G^* = Q\Lambda Q^T$ be the Eigendecomposition of G^*. A matrix $\tilde{X}_s \in \mathbb{R}^{m \times d_2}$ which minimizes $\|\tilde{X}_s \tilde{X}_s^T - G^*\|_F$ is given by $\sum_{i=1, \lambda_i \geq 0}^{d_2} \lambda_i^{\frac{1}{2}} q_i$, where λ_i and q_i are the ith largest eigenvalue and corresponding eigenvector.*

Proof. Since $G^* \in S^{c \times c}$, its Eigendecomposition is $G^* = Q\Lambda Q^T$ where Q is orthonormal. Let $\tilde{\lambda}, \tilde{q}$ be the d_2 largest nonnegative eigenvalues in Λ and their corresponding eigenvectors; additionally, denote the complementary eigenvalues and associated eigenvectors as $\tilde{\lambda}^{\perp} = \Lambda \setminus \tilde{\lambda}, \tilde{q}^{\perp} = Q \setminus \tilde{q}$. Using the Eckart–Young–Mirsky Theorem for the Frobenius norm Kishore Kumar & Schneider (2017), note that for $G \in S_+^{c \times c}, Rank(G) \leq d_2$; $\|G^* - G\|_F \geq \|\tilde{q}^{\perp} \tilde{\lambda}^{\perp} \tilde{q}^{\perp T}\|_F = \sum_{\lambda_i \in \tilde{\lambda}^{\perp}} \lambda_i^{\frac{1}{2}}$ and that $\|G^* - G\|_F$ is minimized for $G = \tilde{q}\tilde{\lambda}\tilde{q}^T$. Using this result, we can recover $\tilde{X}_s$:

$$\underset{\substack{G \in S_+^{c \times c}, \\ Rank(G) \leq d_2}}{\arg\min} \|G^* - G\|_F = \sum_{\lambda_i \in \tilde{\lambda}} (\lambda_i^{\frac{1}{2}} q_i)(\lambda_i^{\frac{1}{2}} q_i)^T = \tilde{X}_s \tilde{X}_s^T \tag{6}$$

Using the above matrix approximation, we find our aligned embedding $\tilde{X}_s = \sum_{\lambda_i \in \tilde{\lambda}} \lambda_i^{\frac{1}{2}} q_i$, a minimizer of $\|\tilde{X}_s \tilde{X}_s^T - G^*\|_F$. □

Due to the rank constraint on G, we are only interested in the d_2 largest eigenvalues and corresponding eigenvectors which incurs a complexity of $\mathcal{O}(d_2 c^2)$ using power iteration Panju (2011).

3.2.2 Solutions using Stochastic Gradient Descent

Alternatively, solutions to the FIPP objective can be obtained using Stochastic Gradient Descent (SGD). This requires defining a single variable $\tilde{X}_s \in \mathbb{R}^{c \times d_2}$ over which to optimize. We find that the solutions obtained after convergence of SGD are close, with respect to the Frobenius norm, to those obtained with low rank PSD approximations up to a rotation. However, the complexity of solving FIPP using SGD is $\mathcal{O}(Tc^2)$, where T is the number of training epochs. Empirically we find $T \gg d_2$ for SGD convergence and, as a result, this approach incurs a complexity much greater than that of low-rank semidefinite approximations.

4 Inference and Alignment

4.1 Inference with Least Squares Projection

To infer aligned source embeddings for words outside of the supervision dictionary, we make the assumption that source words not used for alignment should preserve their distances to those in the seed dictionary in their respective spaces, i.e., $X_s \mathfrak{X}_s^T \approx \tilde{X}_s \tilde{\mathfrak{X}}_s^T$. Using this assumption, we formulate a least squares projection Boyd & Vandenberghe (2004) on an overdetermined system of equations to recover $\tilde{\mathfrak{X}}_s^T$:

$$\tilde{\mathfrak{X}}_s^T = (\tilde{X}_s^T \tilde{X}_s)^{-1} \tilde{X}_s^T X_s \mathfrak{X}_s^T \tag{7}$$

4.2 Weighted Orthogonal Procrustes

As $\tilde{X}_s \in \mathbb{R}^{c \times d_2}$ has been optimized only with concern for its inner products, $\tilde{X}_s$ must be rotated so that it's basis matches that of X_t. We propose a weighted variant of the Orthogonal Procrustes solution to account for differing levels of translation uncertainty among pairs in the seed dictionary, which may arise due to polysemy, semantic shifts and translation errors. In Weighted Least Squares problems, an inverse variance-covariance weighting, W, is used Strutz (2010); Brignell et al. (2015)

to account for differing levels of measurement uncertainty among samples. We solve a weighted Procrustes objective, where measurement error is assumed to be uncorrelated across samples and approximated as the transfer loss for each translation pair, $W_{ii}^{-1} = \|\tilde{X}_s[i]X_s^T - X_t[i]X_t^T\|^2$:

$$\text{SVD}((WX_t)^T W\tilde{X}_s) = U\Sigma V^T, \Omega^W = \underset{\Omega \in O(d_2)}{\arg\min} \|W(\tilde{X}_s\Omega - X_t)\|_F^2 = UV^T, \tag{8}$$

where $O(d_2)$ is the group of $d_2 \times d_2$ unitary matrices. After obtaining Ω^W, we apply the rotation on $\tilde{X}_s$.

5 Experimentation

In this section, we report experimental bilingual lexicon induction performed on the XLING (5K) dataset and downstream experiments performed on MNLI Natural Language Inference and Ted CLDC Document Classification tasks.

5.1 Bilingual Lexicon Induction - XLING (5K)

The XLing BLI task dictionaries constructed by Glavaš et al. (2019) include all 28 pairs between 8 languages in different language families, Croatian (HR), English (EN), Finnish (FI), French (FR), German (DE), Italian (IT), Russian (RU), and Turkish (TR).

Table 1: Mean Average Precision (MAP) of alignment methods on XLING with 5K supervision dictionaries, retrieval method is nearest neighbors.

Method	VecMap	MUSE	ICP	GWA	CCA	PROC	PROC-B*	DLV	RCSLS	FIPP
DE-FI	0.302	0.000	0.251	0.216	0.353	0.359	0.362	0.357	**0.395**	0.389
DE-FR	0.505	0.005	0.454	0.433	0.509	0.511	0.514	0.506	0.536	**0.543**
DE-HR	0.300	0.245	0.240	0.015	0.318	0.329	0.324	0.328	0.359	**0.360**
DE-IT	0.493	0.496	0.447	0.440	0.506	0.510	0.508	0.510	0.529	**0.533**
DE-RU	0.322	0.272	0.245	0.222	0.411	0.425	0.413	0.423	**0.458**	0.449
DE-TR	0.253	0.237	0.215	0.101	0.280	0.284	0.278	0.284	**0.324**	0.321
EN-DE	0.521	0.520	0.486	0.029	0.542	0.544	0.532	0.545	0.580	**0.590**
EN-FI	0.292	0.000	0.262	0.022	0.383	0.396	0.380	0.396	0.438	**0.439**
EN-FR	0.626	0.632	0.613	0.462	0.652	0.654	0.642	0.649	0.675	**0.679**
EN-HR	0.268	0.000	0.000	0.003	0.325	0.336	0.336	0.334	0.375	**0.382**
EN-IT	0.600	0.608	0.577	0.017	0.624	0.625	0.612	0.625	**0.652**	0.649
EN-RU	0.323	0.000	0.259	0.002	0.454	0.464	0.449	0.467	**0.510**	0.502
EN-TR	0.288	0.294	0.000	0.003	0.327	0.335	0.328	0.335	0.386	**0.407**
FI-FR	0.368	0.348	0.000	0.121	0.362	0.362	0.350	0.351	0.395	**0.407**
FI-HR	0.280	0.228	0.208	0.009	0.288	0.294	0.293	0.294	0.321	**0.335**
FI-IT	0.355	0.000	0.263	0.173	0.353	0.355	0.348	0.356	0.388	**0.407**
FI-RU	0.312	0.001	0.231	0.086	0.340	0.342	0.327	0.342	0.376	**0.379**
HR-FR	0.402	0.000	0.282	0.018	0.372	0.374	0.365	0.364	0.412	**0.426**
HR-IT	0.389	0.000	0.045	0.021	0.366	0.364	0.368	0.366	0.399	**0.415**
HR-RU	0.376	0.000	0.309	0.001	0.367	0.372	0.365	0.374	0.404	**0.408**
IT-FR	0.667	0.662	0.629	0.655	0.668	0.669	0.664	0.665	0.682	**0.684**
RU-FR	0.463	0.005	0.000	0.188	0.469	0.470	0.478	0.466	0.494	**0.497**
RU-IT	0.463	0.450	0.394	0.190	0.474	0.474	0.476	0.475	0.491	**0.503**
TR-FI	0.246	0.000	0.173	0.102	0.260	0.269	0.270	0.268	0.300	**0.306**
TR-FR	0.341	0.000	0.000	0.106	0.337	0.338	0.333	0.333	0.375	**0.380**
TR-HR	0.223	0.133	0.138	0.016	0.250	0.259	0.244	0.255	0.285	**0.288**
TR-IT	0.332	0.000	0.243	0.142	0.331	0.335	0.330	0.336	0.368	**0.372**
TR-RU	0.200	0.000	0.119	0.040	0.285	0.290	0.262	0.289	**0.324**	0.319
Avg.	**0.375**	0.183	0.253	0.137	0.400	0.405	0.398	0.403	0.437	**0.442**

The dictionaries use the same vocabulary across languages and are constructed based on word frequency, to reduce biases known to exist in other datasets Kementchedjhieva et al. (2019). We evaluate FIPP across the all language pairs using a supervision dictionary of size 5K; FIPP outperforms other approaches on 22 of 28 language pairs.

5.2 Downstream Evaluations

We perform downstream evaluation on Document Classification and Natural Language Inference tasks. Different methods each achieve strong downstream performance on a subset of tasks and we find that FIPP performs competitively on both tasks.

5.2.1 TED CLDC - Document Classification

The TED-CLDC corpus Hermann & Blunsom (2014) contains cross-lingual documents across 15 topics and 12 language pairs. Following the evaluation of Glavaš et al. (2019), a simple CNN based classifier is trained and evaluated over each topic for language pairs included in our BLI evaluations (EN-DE, EN-FR, EN-IT, EN-RU, and EN-TR). The micro-averaged F_1 scores for each language pair are shown in Table 2. RCSLS outperforms other methods overall and on DE, FR, and RU while FIPP performs best on IT and TR.

Method	VecMap	MUSE	ICP	GWA	PROC	PROC-B*	DLV	RCSLS	FIPP
DE	0.433	0.288	0.492	0.180*	0.345	0.352	0.299	**0.588**	0.520
FR	0.316	0.223	0.254	0.209*	0.239	0.210	0.175	**0.540**	0.433
IT	0.333	0.198*	0.457*	0.206*	0.310	0.218	0.234	0.451	**0.459**
RU	0.504	0.226*	0.362	0.151*	0.251	0.186	0.375	**0.527**	0.481
TR	0.439	0.264*	0.175	0.173*	0.190	0.310	0.208	0.447	**0.491**
Avg.	0.405	0.240	0.348	0.184	0.267	0.255	0.258	**0.510**	0.477

Table 2: TED-CLDC micro-averaged F_1 scores using a CNN model with embeddings from different alignment methods, evaluation follows Glavaš et al. (2019).

5.2.2 MNLI - Natural Language Inference

The multilingual XNLI corpus Williams et al. (2018) includes 5 of the 8 languages used in BLI: EN, DE, FR, RU, and TR. We perform the same evaluation as Glavaš et al. (2019) by training an ESIM model Chen et al. (2017) using EN word embeddings from a shared EN-L2 embedding space for $L2 \in \{\text{DE, FR, RU, TR}\}$. The trained model is then evaluated without further training on the L2 XNLI test set using L2 embeddings from the shared space. The test set accuracy for different alignment methods is reported below in Table 3. The bootstrap Procrustes approach Glavaš et al. (2019) outperforms other methods narrowly while RCSLS performs worst despite having high BLI accuracy.

Method	VecMap	MUSE	ICP	GWA	PROC	PROC-B*	DLV	RCSLS	FIPP
EN-DE	0.604	0.611	0.580	0.427*	0.607	**0.615**	0.614	0.390	0.603
EN-FR	**0.613**	0.536	0.510	0.383*	0.534	0.532	0.556	0.363	0.509
EN-RU	0.574	0.363*	0.572	0.376*	0.585	**0.599**	0.571	0.399	0.577
EN-TR	0.534	0.359*	0.400*	0.359*	0.568	0.573	**0.579**	0.387	0.547
Avg.	**0.581**	0.467	0.516	0.386	0.574	**0.580**	0.571	0.385	0.559

Table 3: MNLI test accuracy using an ESIM model with embeddings from different alignment methods, evaluation follows Glavaš et al. (2019).

6 Discussion

6.1 Runtime Comparison

Method	FIPP	VecMap	RCSLS
CPU	23s	15,896s	387s
GPU	-	612s	-

Table 4: Average time required to run alignment approaches, supervised methods use a 5K supervision dictionary.

An advantage of FIPP compared to existing methods is it's computational efficiency. We provide a runtime comparison of FIPP and the best performing unsupervised Artetxe et al. (2018c) and supervised Joulin et al. (2018b) alignment methods. The average execution time of alignment on 3 runs of the 'EN-DE' XLING 5K dictionary is provided in Table 4. The implementation used for RCSLS is from Facebook's fastText repo [1] which provides CPU computation and the VecMap implementation is from the author's repo [2] with and without the 'cuda' flag. Runtime calculations were performed on a machine with 2 Nvidia GeForce GTX 1080 Ti GPUs, 12 Intel Core i7-6800K processors, and 112GB RAM.

6.2 Isotropic Preprocessing

Common preprocessing steps used by previous approaches, Joulin et al. (2018a); Artetxe et al. (2018a), involve normalizing the rows of $\mathfrak{X}_s, \mathfrak{X}_t$ to have an ℓ_2 norm of 1 and demeaning columns. The transfer loss of the FIPP objective makes direct comparisons on the Gram matrices, $X_s X_s^T$ and $X_t X_t^T$, of the source and target embeddings for words in the seed dictionary. To reduce the influence of dimensional biases between X_s and X_t and ensure words are weighted equally during alignment, it is desired that X_s and X_t be isotropic - i.e. $Z(c) = \sum_{i \in X_s} \exp c^T X_s[i]$ is approximately constant for any unit vector c Arora et al. (2016). Mu & Viswanath (2018) find that a first order approximation to enforce isotropic behavior is achieved by column demeaning while a second order approximation is obtained by the removal of the top PCA components. In FIPP, we apply this simple pre-processing approach by removing the top PCA component of X_s and X_t. Empirically, the distributions of inner products between a source and target embedding can differ substantially when not pre-processed, rendering a substandard alignment, which is discussed further in the Appendix.

6.3 Alignment of Embeddings of Differing Dimensionality

d_1	Linear	CCA	FIPP
300	0.544	0.542	0.590
250	0.524	0.529	0.574
200	0.486	0.501	0.543

Table 5: XLING 5K (EN-DE) MAP for Linear, RCSLS, and FIPP alignment methods on embeddings of differing dimensionality.

FIPP is able to align embeddings of different dimensionalities to a common representation space unlike competing approaches which rely on neighborhood information Joulin et al. (2018b) or utilize inner products between source and target embeddings Artetxe et al. (2018c). We evaluate BLI performance on embeddings of different dimensionalities for the (English, German) language pair. We align English embeddings with dimensions $\in \{200, 250, 300\}$ to a German embedding of dimension 300 and compare the performance of FIPP with CCA and the best linear transform with orthonormal rows, $\Omega^* = \arg\min_{\Omega \in \mathbb{R}^{d_1 \times d_2}, \Omega\Omega^T = I_{d_1}} \|X_s\Omega - X_t\|_F$, equivalent to the Orthogonal Procrustes solution when $d_1 = d_2$. While the performance of all methods decreases as $d_2 - d_1$ increases, the relative performance gap between FIPP and other approaches is maintained.

7 Conclusion

In this paper, we introduced Filtered Inner Product Projection (FIPP), a method for aligning multiple embeddings to a common representation space using pairwise inner product information. FIPP accounts for semantic shifts and aligns embeddings only on common portions of their geometries. Unlike previous approaches, FIPP can be applied to pairs of embedding spaces regardless of their dimensions. We provide two methods for finding approximate solutions to the FIPP objective and show that it can be efficiently solved even in the case of large seed dictionaries. We evaluate FIPP on the task of bilingual lexicon induction using the XLING (5K) dataset, on which it achieves state-of-the-art performance on most language pairs.

[1] https://github.com/facebookresearch/fastText/tree/master/alignment
[2] https://github.com/artetxem/vecmap

REFERENCES

Hanan Aldarmaki, Mahesh Mohan, and Mona Diab. Unsupervised word mapping using structural similarities in monolingual embeddings. *Transactions of the Association for Computational Linguistics*, 6:185–196, 2018. doi: 10.1162/tacl_a_00014. URL https://www.aclweb.org/anthology/Q18-1014.

David Alvarez-Melis and Tommi Jaakkola. Gromov-Wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1881–1890, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1214. URL https://www.aclweb.org/anthology/D18-1214.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. A latent variable model approach to PMI-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399, 2016. doi: 10.1162/tacl_a_00106. URL https://www.aclweb.org/anthology/Q16-1028.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2289–2294, 2016.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 451–462. Association for Computational Linguistics, 2017. doi: 10.18653/v1/P17-1042. URL http://aclweb.org/anthology/P17-1042.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *AAAI*, pp. 5012–5019, 2018a. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16935.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 789–798, Melbourne, Australia, July 2018b. Association for Computational Linguistics. doi: 10.18653/v1/P18-1073. URL https://www.aclweb.org/anthology/P18-1073.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 789–798, Melbourne, Australia, July 2018c. Association for Computational Linguistics. doi: 10.18653/v1/P18-1073. URL https://www.aclweb.org/anthology/P18-1073.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. ISSN 2307-387X.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004. ISBN 0521833787.

Christopher J. Brignell, IL Dryden, and William Browne. *Covariance Weighted Procrustes Analysis*, pp. 189–209. Springer, 2015.

Hailong Cao, Tiejun Zhao, Shu Zhang, and Yao Meng. A distribution-based model to learn bilingual word embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 1818–1827, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL https://www.aclweb.org/anthology/C16-1171.

Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, Hui Jiang, and Diana Inkpen. Enhanced lstm for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, Vancouver, July 2017. ACL.

Georgiana Dinu and Marco Baroni. Improving zero-shot learning by mitigating the hubness problem. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6568.

Manaal Faruqui and Chris Dyer. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 462–471, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. doi: 10.3115/v1/E14-1049. URL https://www.aclweb.org/anthology/E14-1049.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 710–721, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1070. URL https://www.aclweb.org/anthology/P19-1070.

Edouard Grave, Armand Joulin, and Quentin Berthet. Unsupervised alignment of embeddings with wasserstein procrustes. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1880–1890. PMLR, 2019. URL http://proceedings.mlr.press/v89/grave19a.html.

Kristina Gulordava and Marco Baroni. A distributional similarity approach to the detection of semantic change in the google books ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pp. 67–71, 2011.

William L Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 595–605, 2016a.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 1, pp. 1489–1501, 2016b.

Karl Moritz Hermann and Phil Blunsom. Multilingual models for compositional distributed semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 58–68, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-1006. URL https://www.aclweb.org/anthology/P14-1006.

Yedid Hoshen and Lior Wolf. An iterative closest point method for unsupervised word translation. *CoRR*, abs/1801.06126, 2018. URL http://arxiv.org/abs/1801.06126.

Pratik Jawanpuria, Arjun Balgovind, Anoop Kunchukuttan, and Bamdev Mishra. Learning multilingual word embeddings in latent metric space: A geometric approach. *Transactions of the Association for Computational Linguistics*, 7:107–120, 2019. doi: 10.1162/tacl_a_00257. URL https://doi.org/10.1162/tacl_a_00257.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, and Edouard Grave. Improving supervised bilingual mapping of word embeddings. *CoRR*, abs/1804.07745, 2018a. URL http://arxiv.org/abs/1804.07745.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2979–2984, Brussels, Belgium, October-November 2018b. Association for Computational Linguistics. doi: 10.18653/v1/D18-1330. URL https://www.aclweb.org/anthology/D18-1330.

Yova Kementchedjhieva, Mareike Hartmann, and Anders Søgaard. Lost in evaluation: Misleading benchmarks for bilingual dictionary induction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural*

Language Processing (EMNLP-IJCNLP), pp. 3336–3341, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1328. URL https://www.aclweb.org/anthology/D19-1328.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. Temporal analysis of language through neural language models. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 61, 2014.

N Kishore Kumar and Jan Schneider. Literature survey on low rank approximation of matrices. *Linear and Multilinear Algebra*, 65(11):2212–2244, 2017.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H196sainb.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168, 2013a. URL http://arxiv.org/abs/1309.4168.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 3111–3119, 2013b.

Aditya Mogadala and Achim Rettinger. Bilingual word embeddings from parallel and non-parallel corpora for cross-language text classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 692–702, 2016.

Jiaqi Mu and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=HkuGJ3kCb.

Maysum Panju. Iterative methods for computing eigenvalues and eigenvectors. *arXiv preprint arXiv:1105.1185*, 2011.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

Sebastian Ruder, Ryan Cotterell, Yova Kementchedjhieva, and Anders Søgaard. A discriminative latent-variable model for bilingual lexicon induction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 458–468, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1042. URL https://www.aclweb.org/anthology/D18-1042.

Eyal Sagi, Stefan Kaufmann, and Brady Clark. Tracing semantic change with latent semantic analysis. *Current methods in historical semantics*, pp. 161–183, 2011.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=r1Aab85gg.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 778–788, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1072. URL https://www.aclweb.org/anthology/P18-1072.

Tilo Strutz. *Data Fitting and Uncertainty: A Practical Introduction to Weighted Least Squares and Beyond*. Vieweg and Teubner, Wiesbaden, DEU, 2010. ISBN 3834810223.

Rui Wang, Hai Zhao, Sabine Ploux, Bao-Liang Lu, Masao Utiyama, and Eiichiro Sumita. Graph-based bilingual word embedding for statistical machine translation. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 17(4), July 2018. ISSN 2375-4699. doi: 10.1145/3203078. URL https://doi.org/10.1145/3203078.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/N18-1101.

Zi Yin and Yuanyuan Shen. On the dimensionality of word embedding. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 31*, pp. 887–898. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/7368-on-the-dimensionality-of-word-embedding.pdf.

Zi Yin, Vin Sachidananda, and Balaji Prabhakar. The global anchor method for quantifying linguistic shifts and domain adaptation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 31*, pp. 9412–9423. Curran Associates, Inc., 2018.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1959–1970. Association for Computational Linguistics, 2017. doi: 10.18653/v1/P17-1179. URL http://aclweb.org/anthology/P17-1179.

Will Y Zou, Richard Socher, Daniel Cer, and Christopher D Manning. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1393–1398, 2013.