Abstract
Cross-domain alignment between two sets of entities (e.g., objects in an image, words in a sentence) is fundamental to both computer vision and natural language processing. Existing methods mainly focus on designing advanced attention mechanisms to simulate soft alignment, with no training signals to explicitly encourage alignment. The learned attention matrices are also dense and lacks interpretability. We propose Graph Optimal Transport (GOT), a principled framework that germinates from recent advances in Optimal Transport (OT). In GOT, cross-domain alignment is formulated as a graph matching problem, by representing entities into a dynamically-constructed graph. Two types of OT distances are considered: (i) Wasserstein distance (WD) for node (entity) matching; and (ii) Gromov-Wasserstein distance (GWD) for edge (structure) matching. Both WD and GWD can be incorporated into existing neural network models, effectively acting as a drop-in regularizer. The inferred transport plan also yields sparse and self-normalized alignment, enhancing the interpretability of the learned model. Experiments show consistent outperformance of GOT over baselines across a wide range of tasks, including image-text retrieval, visual question answering, image captioning, machine translation, and text summarization.

1. Introduction
Cross-domain Alignment (CDA), which aims to associate related entities across different domains, plays a central role in a wide range of deep learning tasks, such as image-text retrieval (Karpathy & Fei-Fei, 2015; Lee et al., 2018), visual question answering (VQA) (Malinowski & Fritz, 2014; Antol et al., 2015), and machine translation (Bahdanau et al., 2015; Vaswani et al., 2017). Considering VQA as an example, in order to understand the contexts in the image and the question, a model needs to interpret the latent alignment between regions in the input image and words in the question. Specifically, a good model should: (i) identify entities of interest in both the image (e.g., objects/regions) and the question (e.g., words/phrases), (ii) quantify both intra-domain (within the image or sentence) and cross-domain relations between these entities, and then (iii) design good metrics for measuring the quality of cross-domain alignment drawn from these relations, in order to optimize towards better results.

CDA is particularly challenging as it constitutes a weakly supervised learning task. That is, only paired spaces of entity are given (e.g., an image paired with a question), while the ground-truth relations between these entities are not provided (e.g., no supervision signal for a “dog” region in an image aligning with the word “dog” in the question). State-of-the-art methods principally focus on designing advanced attention mechanisms to simulate soft alignment (Bahdanau et al., 2015; Xu et al., 2015; Yang et al., 2016b,a; Vaswani et al., 2017). For example, Lee et al. (2018); Kim et al. (2018); Yu et al. (2019) have shown that learned co-attention can model dense interactions between entities and infer cross-domain latent alignments for vision-and-language tasks. Graph attention has also been applied to relational reasoning for image captioning (Yao et al., 2018) and VQA (Li et al., 2019a), such as graph attention network (GAT) (Veličković et al., 2018) for capturing relations between entities in a graph via masked attention, and graph matching network (GMN) (Li et al., 2019b) for graph alignment via cross-graph soft attention. However, conventional attention mechanisms are guided by task-specific losses, with no training signal to explicitly encourage alignment. And the learned attention matrices are often dense and uninterpretable, thus inducing less effective relational inference.

We address whether there is a more principled approach to scalable discovery of cross-domain relations. To explore this, we present Graph Optimal Transport (GOT), a new

---

Most of this work was done when the first author was an intern at Microsoft. 1Duke University 2Microsoft Dynamics 365 AI Research. Correspondence to: Liqun Chen <liqun.chen@duke.edu>, Zhe Gan <zhe.gan@microsoft.com>.

1Another GOT framework was proposed in Maretic et al. (2019) for graph comparison. We use the same acronym for the proposed algorithm; however, our method is very different from theirs.
framework for cross-domain alignment that leverages recent advances in Optimal Transport (OT). OT-based learning aims to optimize for distribution matching via minimizing the cost of transporting one distribution to another. We extend this to CDA (here a domain can be language, images, videos, etc.). The transport plan is thus redefined as transporting the distribution of embeddings from one domain (e.g., language) to another (e.g., images). By minimizing the cost of the learned transport plan, we explicitly minimize the embedding distance between the domains, i.e., optimizing towards better cross-domain alignment.

Specifically, we convert entities (e.g., objects, words) in each domain (e.g., image, sentence) into a graph, where each entity is represented by a feature vector, and the graph representations are recurrently updated via graph propagation. Cross-domain alignment can then be formulated into a graph matching problem, and be addressed by calculating matching scores based on graph distance. In our GOT framework, we utilize two types of OT distance: (i) Wasserstein distance (WD) (Peyré et al., 2019) is applied to node (entity) matching, and (ii) Gromov-Wasserstein distance (GWD) (Peyré et al., 2016) is adopted for edge (structure) matching. WD only measures the distance between node embeddings across domains, without considering topological information encoded in the graphs. GWD, on the other hand, compares graph structures by measuring the distance between a pair of nodes within each graph. When fused together, the two distances allow the proposed GOT framework to effectively take into account both node and edge information for better graph matching.

The main contributions of this work are summarized as follows. (i) We propose Graph Optimal Transport (GOT), a new framework that tackles cross-domain alignment by adopting Optimal Transport for graph matching. (ii) GOT is compatible with existing neural network models, acting as an effective drop-in regularizer to the original objective. (iii) To demonstrate the versatile generalization ability of the proposed approach, we conduct experiments on five diverse tasks: image-text retrieval, visual question answering, image captioning, machine translation, and text summarization. Results show that GOT provides consistent performance enhancement over strong baselines across all the tasks.

2. Graph Optimal Transport Framework

We first introduce the problem formulation of Cross-domain Alignment in Sec. 2.1, then present the proposed Graph Optimal Transport (GOT) framework in Secs. 2.2-2.4.

2.1. Problem Formulation

Assume we have two sets of entities from two different domains (denoted as $\mathcal{D}_x$ and $\mathcal{D}_y$). For each set, every entity is represented by a feature vector, i.e., $\mathbf{x} = \{\mathbf{x}_i\}_{i=1}^n$ and $\mathbf{y} = \{\mathbf{y}_j\}_{j=1}^m$, where $n$ and $m$ are the number of entities in each domain, respectively. The scope of this paper mainly focuses on tasks involving images and text, thus entities here correspond to objects in an image or words in a sentence. An image can be represented as a set of detected objects, each associated with a feature vector (e.g., from a pre-trained Faster RCNN (Anderson et al., 2018)). With a word embedding layer, a sentence can be represented as a sequence of word feature vectors.

A deep neural network $f_\theta(\cdot)$ can be designed to take both $\mathbf{X}$ and $\mathbf{Y}$ as initial inputs, and generate contextualized representations:

$$\mathbf{X}, \mathbf{Y} = f_\theta(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}),$$

where $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$, $\mathbf{Y} = \{\mathbf{y}_j\}_{j=1}^m$, and advanced attention mechanisms (Bahdanau et al., 2015; Vaswani et al., 2017) can be applied to $f_\theta(\cdot)$ to simulate soft alignment. The final supervision signal $l$ is then used to learn $\theta$, i.e., the training objective is defined as:

$$\mathcal{L}(\theta) = \mathcal{L}_{sup}(\mathbf{X}, \mathbf{Y}, l).$$

Several instantiations for different tasks are summarized as follows. (i) Image-text Retrieval. $\mathbf{X}$ and $\mathbf{Y}$ are image and text features, respectively. $l$ is the binary label, indicating whether the input image and sentence are paired or not. Here $f_\theta(\cdot)$ can be the SCAN model (Lee et al., 2018), and $\mathcal{L}_{sup}(\cdot)$ corresponds to ranking loss (Faghri et al., 2018; Chechik et al., 2010). (ii) VQA. Here $l$ denotes the ground-truth answer, $f_\theta(\cdot)$ can be BUTD or BAN model (Anderson et al., 2018; Kim et al., 2018), $\mathcal{L}_{sup}(\cdot)$ is cross-entropy loss. (iii) Machine Translation. $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ are textual features from the source and target sentences, respectively. Here $f_\theta(\cdot)$ can be an encoder-decoder Transformer model (Vaswani et al., 2017), and $\mathcal{L}_{sup}(\cdot)$ corresponds to cross-entropy loss that models the conditional distribution of $p(Y|X)$, and here $l$ is not needed. To simplify subsequent discussions, all the tasks are abstracted into $f_\theta(\cdot)$ and $\mathcal{L}_{sup}(\cdot)$.

In most previous work, the learned attention can be interpreted as a soft alignment between $\mathbf{X}$ and $\mathbf{Y}$. However, only the final supervision signal $\mathcal{L}_{sup}(\cdot)$ is used for model training, thus lacking an objective explicitly encouraging cross-domain alignment. To enforce alignment and cast a regularizing effect on model training, we propose a new objective for Cross-domain Alignment:

$$\mathcal{L}(\theta) = \mathcal{L}_{sup}(\mathbf{X}, \mathbf{Y}, l) + \alpha \cdot \mathcal{L}_{\text{CDA}}(\mathbf{X}, \mathbf{Y}),$$

where $\mathcal{L}_{\text{CDA}}(\cdot)$ is a regularization term that encourages alignments explicitly, and $\alpha$ is a hyper-parameter that balances the two terms. Through gradient back-propagation, the learned $\theta$ supports more effective relational inference. In Section 2.4 we describe $\mathcal{L}_{\text{CDA}}(\cdot)$ in detail.
2.2. Dynamic Graph Construction

Image and text data inherently contain rich sequential/spatial structures. By representing them as graphs and performing graph alignment, not only cross-domain relations can be modeled, but also intra-domain relations are exploited (e.g., semantic/spatial relations among detected objects in an image (Li et al., 2019a)).

Given $X$, we aim to construct a graph $G_x(V_x, E_x)$, where each node $i \in V_x$ is represented by a feature vector $x_i$. To add edges $E_x$, we first calculate the similarity between a pair of entities inside a graph: $C_x = \{\cos(x_i, x_j)\}_{i,j} \in \mathbb{R}^{n \times n}$. Further, we define $C_x = \max(C_x - \tau, 0)$, where $\tau$ is a threshold hyper-parameter for the graph cost matrix. Empirically, $\tau$ is set to 0.1. If $[C_x]_{ij} > 0$, an edge is added between node $i$ and $j$. Given $Y$, another graph $G_y(V_y, E_y)$ can be similarly constructed. Since both $X$ and $Y$ are evolving through the update of parameters $\theta$ during training, this graph construction process is considered “dynamic”.

By representing the entities in both domains as graphs, cross-domain alignment is naturally formulated into a graph matching problem.

In our proposed framework, we use Optimal Transport (OT) for graph matching, where a transport plan $T \in \mathbb{R}^{n \times m}$ is learned to optimize the alignment between $X$ and $Y$. OT possesses several idiosyncratic characteristics that make it a good choice for solving CDA problem. (i) Self-normalization: all the elements of $T^*$ sum to 1 (Peyré et al., 2019). (ii) Sparsity: when solved exactly, OT yields a sparse solution $T^*$ containing $(2r - 1)$ non-zero elements at most, where $r = \max(n, m)$, leading to a more interpretable and robust alignment (De Goes et al., 2011). (iii) Efficiency: compared with conventional linear programming solvers, our solution can be readily obtained using iterative procedures that only require matrix-vector products (Xie et al., 2018), hence readily applicable to large deep neural networks.

![Figure 1. Illustration of the Wasserstein Distance (WD) and the Gromov-Wasserstein Distance (GWD) used for node and structure matching, respectively.](image)

**Algorithm 1** Computing Wasserstein Distance.

1. **Input:** $\{x_i\}_{i=1}^n, \{y_j\}_{j=1}^m$, $\beta$
2. $\sigma = \frac{1}{\beta} \mathbf{1}_n$, $T^{(1)} = \mathbf{1}^T$
3. $C_{ij} = c(x_i, y_j)$, $A_{ij} = e^{-\frac{C_{ij}}{\beta}}$
4. for $t = 1, 2, 3 \ldots$ do
5. $Q = A \odot T^{(t)}$ // $\odot$ is Hadamard product
6. for $k = 1, 2, 3 \ldots K$ do
7. $\delta = \frac{1}{\beta} Q \sigma$
8. end for
9. $T^{(t+1)} = \text{diag}(\delta) Q \text{diag}(\sigma)$
10. end for
11. $D_{wd} = \langle C^T, T \rangle$
12. Return $T$, $D_{wd}$ // $\langle \cdot, \cdot \rangle$ is the Frobenius dot-product

**Algorithm 2** Computing Gromov-Wasserstein Distance.

1. **Input:** $\{x_i\}_{i=1}^n, \{y_j\}_{j=1}^m$, probability vectors $p, q$
2. Compute intra-domain similarities:
3. $[C_x]_{ij} = \cos(x_i, x_j)$, $[C_y]_{ij} = \cos(y_i, y_j)$,
4. Compute cross-domain similarities:
5. $C_{xy} = C_x p_{1m} + C_y q_{Km}$
6. for $t = 1, 2, 3 \ldots$ do
7. $L = \|C_{xy} - 2C_x C_y\|$
8. $\delta = L \odot \text{diag}(\delta)$
9. Apply Algorithm 1 to solve transport plan $T$
10. end for
11. $D_{gw} = \langle L^T, T \rangle$
12. Return $T$, $D_{gw}$

2.3. Optimal Transport Distances

As illustrated in Figure 1, two types of OT distance are adopted for our graph matching: Wasserstein distance for node matching, and Gromov-Wasserstein distance for edge matching.

**Wasserstein Distance** Wasserstein distance (WD) is commonly used for matching two distributions (e.g., two sets of node embeddings). In our setting, discrete WD can be used as a solver for network flow and bipartite matching (Luise et al., 2018). The definition of WD is described as follows.

**Definition 2.1.** Let $\mu \in P(X), \nu \in P(Y)$ denote two discrete distributions, formulated as $\mu = \sum_{i=1}^n u_i \delta_{x_i}$ and $\nu = \sum_{j=1}^m v_j \delta_{y_j}$, with $\delta_x$ as the Dirac function centered on $x$. $\Pi(\mu, \nu)$ denotes all the joint distributions $\gamma(x, y)$, with marginals $\mu(x)$ and $\nu(y)$. The weight vectors $u = \{u_i\}_{i=1}^n \in \Delta_n$ and $v = \{v_j\}_{j=1}^m \in \Delta_m$ belong to the $n$- and $m$-dimensional simplex, respectively (i.e., $\sum_{i=1}^n u_i = \sum_{j=1}^m v_j = 1$), where both $\mu$ and $\nu$ are probability distributions. The Wasserstein distance between the two discrete distributions $\mu, \nu$ is defined as:

$$D_{wd}(\mu, \nu) = \inf_{\gamma \in \Pi(\mu, \nu)} \mathbb{E}(x, y) \sim \gamma [c(x, y)] = \min_{T \in \Pi(u, v)} \sum_{i=1}^n \sum_{j=1}^m T_{ij} \cdot c(x_i, y_j)$$
where $\Pi(u, v) = \{T \in \mathbb{R}^{m \times n} | T 1_n = u, T^T 1_n = v \}$, $1_n$ denotes an $n$-dimensional all-one vector, and $c(x, y)$ is the cost function evaluating the distance between $x_i$ and $y_j$. For example, the cosine distance $c(x, y) = 1 - \frac{|x_i \cdot y_j|}{\|x_i\|_2 \cdot \|y_j\|_2}$ is a popular choice. The matrix $T$ is denoted as the transport plan, where $T_{i,j}$ represents the amount of mass shifted from $u_i$ to $v_j$.

$D_{w} (\mu, \nu)$ defines an optimal transport distance that measures the discrepancy between each pair of samples across the two domains. In our graph matching, this is a natural choice for node (entity) matching.

**Gromov-Wasserstein Distance** Instead of directly calculating distances between two sets of nodes as in WD, Gromov-Wasserstein distance (GWD) (Peyré et al., 2016; Chowdhury & Mémoli, 2019) can be used to calculate distances between two sets of nodes as in WD, but fails to capture the similarity between edges. If there are duplicated entities represented by different nodes in the same graph, WD will treat them as identical and ignore their neighboring relations. For example, given a sentence "there is a red book on the blue desk" paired with an image containing several desks and books in different colors, it is difficult to correctly identify which book in the image the sentence is referring to, without understanding the relations among the objects in the image.

On the other hand, WD can match nodes in different graphs, but fails to capture the similarity between edges. If there are duplicated entities represented by different nodes in the same graph, WD will treat them as identical and ignore their neighboring relations. For example, given a sentence "there is a red book on the blue desk" paired with an image containing several desks and books in different colors, it is difficult to correctly identify which book in the image the sentence is referring to, without understanding the relations among the objects in the image.

**2.4. Graph Matching via OT Distances**

Though GWD is capable of capturing edge similarity between graphs, it cannot be directly applied to graph alignment, since only the similarity between $c_1(x_i, x'_i)$ and $c_2(y_j, y'_j)$ is considered, without taking into account node representations. For example, the word pair ("boy", "girl") has similar cosine similarity as the pair ("football", "basketball"), but the semantic meanings of the two pairs are completely different, and should not be matched.

On the other hand, WD can match nodes in different graphs, but fails to capture the similarity between edges. If there are duplicated entities represented by different nodes in the same graph, WD will treat them as identical and ignore their neighboring relations. For example, given a sentence "there is a red book on the blue desk" paired with an image containing several desks and books in different colors, it is difficult to correctly identify which book in the image the sentence is referring to, without understanding the relations among the objects in the image.

To best couple WD and GWD and unify these two distances in a mutually-beneficial way, we propose a transport plan $T$ shared by both WD and GWD. Compared with naively employing two different transport plans, we observe that this joint plan works better (see Table 8), and faster, since we only need to solve $T$ once (instead of twice). Intuitively,
with a shared transport plan, WD and GWD can enhance each other effectively, as $T$ utilizes both node and edge information simultaneously. Formally, the proposed GOT distance is defined as:

$$D_{got}({\mu}, {\nu}) = \min_{\Pi \in \Pi(u,v)} \sum_{i,j} T_{ij} \left( \lambda c(x_i, y_j) + (1 - \lambda) L(x_i, y_j, x'_i, y'_j) \right).$$

We apply the Sinkhorn algorithm (Cuturi, 2013; Cuturi & Peyré, 2017) to solve WD (4) with an entropic regularizer (Benamou et al., 2015):

$$\min_{T \in \Pi(u,v)} \sum_{i=1}^{n} \sum_{j=1}^{m} T_{ij} c(x_i, y_j) + \beta H(T),$$

where $H(T) = \sum_{i,j} T_{ij} \log T_{ij}$, and $\beta$ is the hyper-parameter controlling the importance of the entropy term. Details are provided in Algorithm 1. The solver for GWD can enhance WD and GWD can enhance WD with a shared transport plan, WD and GWD can enhance each other effectively, as $T$ utilizes both node and edge information simultaneously. Formally, the proposed GOT distance is defined as:

$$\min_{\Pi \in \Pi(u,v)} \sum_{i,j} T_{ij} \left( \lambda c(x_i, y_j) + (1 - \lambda) L(x_i, y_j, x'_i, y'_j) \right).$$

We apply the Sinkhorn algorithm (Cuturi, 2013; Cuturi & Peyré, 2017) to solve WD (4) with an entropic regularizer (Benamou et al., 2015):

$$\min_{T \in \Pi(u,v)} \sum_{i,j} T_{ij} c(x_i, y_j) + \beta H(T),$$

where $H(T) = \sum_{i,j} T_{ij} \log T_{ij}$, and $\beta$ is the hyper-parameter controlling the importance of the entropy term. Details are provided in Algorithm 1. The solver for GWD can enhance WD and GWD can enhance WD with a shared transport plan, WD and GWD can enhance each other effectively, as $T$ utilizes both node and edge information simultaneously. Formally, the proposed GOT distance is defined as:

$$D_{got}({\mu}, {\nu}) = \min_{\Pi \in \Pi(u,v)} \sum_{i,j} T_{ij} \left( \lambda c(x_i, y_j) + (1 - \lambda) L(x_i, y_j, x'_i, y'_j) \right).$$

There are different ways to solve the OT distance, such as linear programming. However, this solver is not differentiable, thus it cannot be applied in deep learning frameworks. Recently, WGAN (Arjovsky et al., 2017) proposes to approximate the dual form of WD by imposing a 1-Lipschitz constraint on the discriminator. Note that the duality used for WGAN is restricted to the W-1 distance, i.e., $\| \cdot \|_1$. The Sinkhorn algorithm was first proposed in Cuturi (2013) as a solver for calculating an entropic regularized OT distance. Thanks to the Envelop Theorem (Cuturi & Peyré, 2017), the Sinkhorn algorithm can be efficiently calculated and readily applied to neural networks. More recently, Vayer et al. (2018) proposed the fused GWD for graph matching. Our proposed GOT framework enjoys the benefits of both Sinkhorn algorithm and fused GWD: it is (i) capable of capturing more structured information via marrying both WD and GWD; and (ii) scalable to large datasets and trainable with deep neural networks.
Graph Optimal Transport for Cross-Domain Alignment

<table>
<thead>
<tr>
<th>Method</th>
<th>Sentence Retrieval</th>
<th>Image Retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R@1</td>
<td>R@5</td>
</tr>
<tr>
<td>VSE++ (ResNet) (Faghri et al., 2018)</td>
<td>52.9</td>
<td>–</td>
</tr>
<tr>
<td>DPC (ResNet) (Zheng et al., 2020)</td>
<td>55.6</td>
<td>81.9</td>
</tr>
<tr>
<td>DAN (ResNet) (Nam et al., 2017)</td>
<td>55.0</td>
<td>81.8</td>
</tr>
<tr>
<td>SCO (ResNet) (Huang et al., 2018)</td>
<td>55.5</td>
<td>82.0</td>
</tr>
<tr>
<td>SCAN (Faster R-CNN, ResNet) (Lee et al., 2018)</td>
<td>67.7</td>
<td>88.9</td>
</tr>
<tr>
<td><strong>Ours (Faster R-CNN, ResNet):</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SCAN + WD</strong></td>
<td>70.9</td>
<td>92.3</td>
</tr>
<tr>
<td><strong>SCAN + GWD</strong></td>
<td>69.5</td>
<td>91.2</td>
</tr>
<tr>
<td><strong>SCAN + GOT</strong></td>
<td>70.9</td>
<td>92.8</td>
</tr>
<tr>
<td>VSE++ (ResNet) (Faghri et al., 2018)</td>
<td>41.3</td>
<td>–</td>
</tr>
<tr>
<td>DPC (ResNet) (Zheng et al., 2020)</td>
<td>41.2</td>
<td>70.5</td>
</tr>
<tr>
<td>GXN (ResNet) (Gu et al., 2018)</td>
<td>42.0</td>
<td>–</td>
</tr>
<tr>
<td>SCO (ResNet) (Huang et al., 2018)</td>
<td>42.8</td>
<td>72.3</td>
</tr>
<tr>
<td>SCAN (Faster R-CNN, ResNet) (Lee et al., 2018)</td>
<td>46.4</td>
<td>77.4</td>
</tr>
<tr>
<td><strong>Ours (Faster R-CNN, ResNet):</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SCAN + WD</strong></td>
<td>50.2</td>
<td>80.1</td>
</tr>
<tr>
<td><strong>SCAN + GWD</strong></td>
<td>47.2</td>
<td>78.3</td>
</tr>
<tr>
<td><strong>SCAN + GOT</strong></td>
<td>50.5</td>
<td>80.2</td>
</tr>
</tbody>
</table>

Table 1. Results on image-text retrieval evaluated on Recall@K (R@K). Upper panel: Flickr30K; lower panel: COCO.

of handling varying node degrees through an appropriate normalization of the adjacency matrix of the data. To further improve the classification accuracy, the graph attention network (GAT) (Veličković et al., 2018) was proposed by using a learned weight matrix instead of the adjacency matrix, with masked attention to aggregate node neighborhood information.

Recently, the graph neural network has been extended to other tasks beyond classification. Li et al. (2019b) proposed graph matching network (GMN) for learning similarities between graphs. Similar to GAT, masked attention is applied to aggregate information from each node within a graph, and cross-graph information is further exploited via soft attention. Task-specific losses are then used to guide model training. In this setting, an adjacency matrix can be directly obtained from the data and soft attention is used to induce alignment. In contrast, our GOT framework does not rely on explicit graph structures in the data, and uses OT for graph alignment.

4. Experiments

To validate the effectiveness of the proposed GOT framework, we evaluate performance on a selection of diverse tasks. We first consider vision-and-language understanding, including: (i) image-text retrieval, and (ii) visual question answering. We further consider text generation tasks, including: (iii) image captioning, (iv) machine translation, and (v) abstractive text summarization. Code is available at https://github.com/LiqunChen0606/Graph-Optimal-Transport.

4.1. Vision-and-Language Tasks

Image-Text Retrieval For image-text retrieval task, we use pre-trained Faster R-CNN (Ren et al., 2015) to extract bottom-up-attention features (Anderson et al., 2018) as the image representation. A set of 36 features is created for each image, each feature represented by a 2048-dimensional vector. For captions, a bi-directional GRU (Schuster & Paliwal, 1997; Bahdanau et al., 2015) is used to obtain textual features.

We evaluate our model on the Flickr30K (Plummer et al., 2015) and COCO (Lin et al., 2014) datasets. Flickr30K contains 31,000 images, with five human-annotated captions per image. We follow previous work (Karpathy & Fei-Fei, 2015; Faghri et al., 2018) for the data split: 29,000, 1,000 and 1,000 images are used for training, validation and test, respectively. COCO contains 123,287 images, each image also accompanied with five captions. We follow the data split in Faghri et al. (2018), where 113,287, 5,000 and 5,000 images are used for training, validation and test, respectively.

We measure the performance of image retrieval and sentence retrieval on Recall at $K$ (R@K) (Karpathy & Fei-Fei, 2015), defined as the percentage of queries retrieving the correct images/sentences within the top $K$ highest-ranked results. In our experiment, $K = \{1, 5, 10\}$, and Rsum (Huang et al., 2017) (summation over all R@K) is used to evaluate the overall performance. Results are summarized in Table 1. Both WD and GWD can boost the performance of the SCAN model, while WD achieves a larger margin than GWD. This indicates that when used alone, GWD may not be a good metric for graph alignment. When combining the
Graph Optimal Transport for Cross-Domain Alignment

Figure 3. (a) A comparison of the inferred transport plan from GOT (top chart) and the learned attention matrix from SCAN (bottom chart). Both serve as a lens to visualize cross-domain alignment. The horizontal axis represents image regions, and the vertical axis represents word tokens. (b) The original image.

<table>
<thead>
<tr>
<th>Model</th>
<th>BAN</th>
<th>BAN+GWD</th>
<th>BAN+WD</th>
<th>BAN+GOT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score</td>
<td>66.00</td>
<td>66.21</td>
<td>66.26</td>
<td>66.44</td>
</tr>
</tbody>
</table>

Table 2. Results (accuracy) on VQA 2.0 validation set, using BAN (Kim et al., 2018) as baseline.

<table>
<thead>
<tr>
<th>Model</th>
<th>BUTD</th>
<th>BAN-1</th>
<th>BAN-2</th>
<th>BAN-4</th>
<th>BAN-8</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o GOT</td>
<td>65.37</td>
<td>65.37</td>
<td>65.61</td>
<td>65.81</td>
<td>66.00</td>
</tr>
<tr>
<td>w/ GOT</td>
<td>65.01</td>
<td>65.68</td>
<td>65.88</td>
<td>66.10</td>
<td>66.44</td>
</tr>
</tbody>
</table>

Table 3. Results (accuracy) of applying GOT to BUTD (Anderson et al., 2018) and BAN-\(m\) (Kim et al., 2018) on VQA 2.0. \(m\) denotes the number of glimpses.

Two distances together, GOT achieves the best performance.

Figure 3 provides visualization on the learned transport plan in GOT and the learned attention matrix in SCAN. Both serve as a proxy to lend insights into the learned alignment. As shown, the attention matrix from SCAN is much denser and noisier than the transport plan inferred by GOT. This shows our model can better discover cross-domain relations between image-text pairs, since the inferred transport plan is more interpretable and has less ambiguity. For example, both the words “sidewalk” and “skateboard” match the corresponding image regions very well.

Because of the Envelope Theorem (Cuturi & Peyré, 2017), GOT needs to be calculated only during the forward phase of model training. Therefore, it does not introduce much extra computation time. For example, when using the same machine for image-text retrieval experiments, SCAN required 6hr 34min for training and SCAN+GOT 6hr 57min.

Visual Question Answering We also consider the VQA 2.0 dataset (Goyal et al., 2017), which contains human-annotated QA pairs on COCO images (Lin et al., 2014). For each image, an average of 3 questions are collected, with 10 candidate answers per question. The most frequent answer from the annotators is selected as the correct answer. Following previous work (Kim et al., 2018), we take the answers that appear more than 9 times in the training set as candidate answers, which results in 3129 candidates. Classification accuracy is used as the evaluation metric, defined as \(\min(1, \frac{\# \text{ humans provided ans.}}{3})\).

The BAN model (Kim et al., 2018) is used as baseline, with the original codebase used for fair comparison. Results are summarized in Table 2. Both WD and GWD improve the BAN model on the validation set, and GOT achieves further performance lift.

We also investigate whether different architecture designs affect the performance gain. We consider BUTD (Anderson et al., 2018) as an additional baseline, and apply different number of glimpses \(m\) to the BAN model, denoted as BAN-\(m\). Results are summarized in Table 3, with the following observations: (i) When the number of parameters in the tested model is small, such as BUTD, the improvement brought by GOT is more significant. (ii) BAN-4, a simpler model than BAN-8, when combined with GOT, can outperform BAN-8 without using GOT (66.10 v.s. 66.00). (iii) For complex models such as BAN-8 that might have limited space for improvement, GOT is still able to achieve performance gain.

4.2. Text Generation Tasks

Image Captioning We conduct experiments on image captioning using the same COCO dataset. The same bottom-up-attention features (Anderson et al., 2018) used in image-
Graph Optimal Transport for Cross-Domain Alignment

<table>
<thead>
<tr>
<th>Method</th>
<th>CIDEr</th>
<th>BLEU-4</th>
<th>BLEU-3</th>
<th>BLEU-2</th>
<th>BLEU-1</th>
<th>ROUGE</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soft Attention (Xu et al., 2015)</td>
<td>-</td>
<td>24.3</td>
<td>34.4</td>
<td>49.2</td>
<td>70.7</td>
<td>-</td>
<td>23.9</td>
</tr>
<tr>
<td>Hard Attention (Xu et al., 2015)</td>
<td>-</td>
<td>25.0</td>
<td>35.7</td>
<td>50.4</td>
<td>71.8</td>
<td>-</td>
<td>23.0</td>
</tr>
<tr>
<td>Show &amp; Tell (Vinyals et al., 2015)</td>
<td>85.5</td>
<td>27.7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>23.7</td>
</tr>
<tr>
<td>ATT-FCN (You et al., 2016)</td>
<td>-</td>
<td>30.4</td>
<td>40.2</td>
<td>53.7</td>
<td>70.9</td>
<td>-</td>
<td>24.3</td>
</tr>
<tr>
<td>SCN-LSTM (Gan et al., 2017)</td>
<td>101.2</td>
<td>33.0</td>
<td>43.3</td>
<td>56.6</td>
<td>72.8</td>
<td>-</td>
<td>25.7</td>
</tr>
<tr>
<td>Adaptive Attention (Lu et al., 2017)</td>
<td>108.5</td>
<td>33.2</td>
<td>43.9</td>
<td>58.0</td>
<td>74.2</td>
<td>-</td>
<td>26.6</td>
</tr>
<tr>
<td>MLE</td>
<td>106.3</td>
<td>34.3</td>
<td>45.3</td>
<td>59.3</td>
<td>75.6</td>
<td>55.2</td>
<td>26.2</td>
</tr>
<tr>
<td>MLE + WD</td>
<td>107.9</td>
<td>34.8</td>
<td>46.1</td>
<td>60.1</td>
<td>76.2</td>
<td>55.6</td>
<td>26.5</td>
</tr>
<tr>
<td>MLE + GWD</td>
<td>106.6</td>
<td>33.3</td>
<td>45.2</td>
<td>59.1</td>
<td>75.7</td>
<td>55.0</td>
<td>25.9</td>
</tr>
<tr>
<td>MLE + GOT</td>
<td><strong>109.2</strong></td>
<td><strong>35.1</strong></td>
<td><strong>46.5</strong></td>
<td><strong>60.3</strong></td>
<td><strong>77.0</strong></td>
<td><strong>56.2</strong></td>
<td><strong>26.7</strong></td>
</tr>
</tbody>
</table>

Table 4. Results of image captioning on the COCO dataset.

<table>
<thead>
<tr>
<th>Model</th>
<th>EN-VI uncased</th>
<th>EN-VI cased</th>
<th>EN-DE uncased</th>
<th>EN-DE cased</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transformer (Vaswani et al., 2017)</td>
<td>29.25 ± 0.18</td>
<td>28.46 ± 0.17</td>
<td>25.60 ± 0.07</td>
<td>25.12 ± 0.12</td>
</tr>
<tr>
<td>Transformer + WD</td>
<td>29.49 ± 0.10</td>
<td>28.68 ± 0.14</td>
<td>25.83 ± 0.12</td>
<td>25.30 ± 0.11</td>
</tr>
<tr>
<td>Transformer + GWD</td>
<td>28.65 ± 0.14</td>
<td>28.34 ± 0.16</td>
<td>25.42 ± 0.17</td>
<td>24.82 ± 0.15</td>
</tr>
<tr>
<td>Transformer + GOT</td>
<td><strong>29.92 ± 0.11</strong></td>
<td><strong>29.09 ± 0.18</strong></td>
<td><strong>26.05 ± 0.17</strong></td>
<td><strong>25.54 ± 0.15</strong></td>
</tr>
</tbody>
</table>

Table 5. Results of neural machine translation on EN-DE and EN-VI.

<table>
<thead>
<tr>
<th>Method</th>
<th>ROUGE-1</th>
<th>ROUGE-2</th>
<th>ROUGE-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABS+ (Rush et al., 2015)</td>
<td>31.00</td>
<td>12.65</td>
<td>28.34</td>
</tr>
<tr>
<td>LSTM (Hu et al., 2018)</td>
<td>36.11</td>
<td>16.39</td>
<td>32.32</td>
</tr>
<tr>
<td>LSTM + GWD</td>
<td>36.31</td>
<td>17.32</td>
<td>33.15</td>
</tr>
<tr>
<td>LSTM + WD</td>
<td>36.81</td>
<td>17.34</td>
<td>33.34</td>
</tr>
<tr>
<td>LSTM + GOT</td>
<td>37.10</td>
<td>17.61</td>
<td>35.70</td>
</tr>
</tbody>
</table>

Table 6. Results of abstractive text summarization on the English Gigawords dataset.

Machine Translation In machine translation (and abstractive summarization), the word embedding spaces of the source and target sentences are different, which can be considered as different domains. Therefore, GOT can be used to align those words with similar semantic meanings between the source and target sentences for better translation/summarization. We choose two machine translation benchmarks for experiments: (i) English-Vietnamese TED-talks corpus, which contains 133K pairs of sentences from the IWSLT Evaluation Campaign (Cettolo et al., 2015); and (ii) a large-scale English-German parallel corpus with 4.5M pairs of sentences, from the WMT Evaluation Campaign (Vaswani et al., 2017). The Texar codebase (Hu et al., 2018) is used in our experiments.

We apply GOT to the Transformer model (Vaswani et al., 2017) and use BLEU score (Papineni et al., 2002) as the evaluation metric. Results are summarized in Table 5. As also observed in Chen et al. (2019a), using WD can improve the performance of the Transformer for sequence-to-sequence learning. However, if only GWD is used, the test BLEU score drops. Since GWD can only match the edges, it ignores supervision signals from node representations. This serves as empirical evidence to support our hypothesis that using GWD alone may not be enough to improve performance. However, GWD serves as a complementary method for capturing graph information that might be missed by WD. Therefore, when combining the two together, GOT achieves the best performance. Example translations are provided in Table 7.

Abstractive Summarization We evaluate abstractive summarization on the English Gigawords benchmark (Graff et al., 2003). A basic LSTM model as implemented in Texar (Hu et al., 2018) is used in our experiments. ROUGE-1, -2 and -L scores (Lin, 2004) are reported. Table 6 shows that both GWD and WD can improve the performance of the LSTM. The transport plan for source and output sentences alignment is illustrated in Figure 4. The learned alignment is sparse and interpretable. For instance, the words “largest” and “projects” in the source sentence matches the words...
Reference: India’s new prime minister, Narendra Modi, is meeting his Japanese counterpart, Shinzo Abe, in Tokyo to discuss economic and security ties, on his first major foreign visit since winning May’s election.

MLE: India’s new prime minister, Narendra Modi, meets his Japanese counterpart, Shinzo Abe, in Tokyo, during his first major foreign visit in May to discuss economic and security relations.

GOT: India’s new prime minister, Narendra Modi, is meeting his Japanese counterpart Shinzo Abe in Tokyo in his first major foreign visit since his election victory in May to discuss economic and security relations.

Reference: Chinese leaders presented the Sunday ruling as a democratic breakthrough because it gives Hong Kongers a direct vote, but the decision also makes clear that Chinese leaders would retain a firm hold on the process through a nominating committee tightly controlled by Beijing.

MLE: The Chinese leadership presented the decision of Sunday as a democratic breakthrough, because it gives Hong Kong citizens a direct right to vote, but the decision also makes it clear that the Chinese leadership maintains the expiration of a nomination committee closely controlled by Beijing.

GOT: The Chinese leadership presented the decision on Sunday as a democratic breakthrough, because Hong Kong citizens have a direct electoral right, but the decision also makes it clear that the Chinese leadership remains firmly in hand with a nominating committee controlled by Beijing.

Table 7. Comparison of German-to-English translation examples. For each example, we show the human translation (reference) and the translation from MLE and GOT. We highlight the key-phrase differences between reference and translation outputs in blue and red, and denote the error in translation in bold. In the first example, GOT correctly maintains all the information in “since winning May’s election” by translating to “since his election victory in May”, whereas MLE only generate “in May”. In the second example, GOT successfully keeps the information “Beijing”, whereas MLE generates wrong words “expiration of”.

<table>
<thead>
<tr>
<th>Model</th>
<th>EN-VI uncased</th>
<th>EN-DE uncased</th>
</tr>
</thead>
<tbody>
<tr>
<td>GOT (shared)</td>
<td>29.92 ± 0.11</td>
<td>26.05 ± 0.18</td>
</tr>
<tr>
<td>GOT (unshared)</td>
<td>29.77 ± 0.12</td>
<td>25.89 ± 0.17</td>
</tr>
</tbody>
</table>

Table 8. Ablation study on transport plan in machine translation. Both models were run 5 times with the same hyper-parameter setting.

<table>
<thead>
<tr>
<th>λ</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>28.65</td>
</tr>
<tr>
<td>0.1</td>
<td>29.31</td>
</tr>
<tr>
<td>0.3</td>
<td>29.52</td>
</tr>
<tr>
<td>0.5</td>
<td>29.65</td>
</tr>
<tr>
<td>0.8</td>
<td>29.92</td>
</tr>
<tr>
<td>1.0</td>
<td>29.49</td>
</tr>
</tbody>
</table>

Table 9. Ablation study of the hyper-parameter λ on the EN-VI machine translation dataset.

“more” and “investment” in the output summary very well.

4.3. Ablation study

We conduct additional ablation study on the EN-VI and EN-DE datasets for machine translation.

Shared Transport Plan T

As discussed in Sec. 2.4, we use a shared transport plan T to solve the GOT distance. An alternative is not to share this T matrix. The comparison results are provided in Table 8. GOT with a shared transport plan achieves better performance than the alternative. Since we only need to run the iterative Sinkhorn algorithm once, it also saves training time than the unshared case.

Hyper-parameter λ

We perform ablation study on the hyper-parameter λ in (6). We select λ from [0, 1] and report results in Table 9. When λ = 0.8, EN-VI translation performs the best, which indicates that the weight on WD needs to be larger than the weight on GWD, since intuitively node matching is more important than edge matching for machine translation. However, both WD and GWD contribute to GOT achieving the best performance.

5. Conclusions

We propose Graph Optimal Transport, a principled framework for cross-domain alignment. With the Wasserstein and Gromov-Wasserstein distances, both intra-domain and cross-domain relations are captured for better alignment. Empirically, we observe that enforcing alignment can serve as an effective regularizer for model training. Extensive experiments show that the proposed method is a generic framework that can be applied to a wide range of cross-domain tasks. For future work, we plan to apply the proposed framework to self-supervised representation learning.

Acknowledgements

The authors would like to thank the anonymous reviewers for their insightful comments. The research at Duke University was supported in part by DARPA, DOE, NIH, NSF and ONR.

References


Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A. C., Salakhutdinov, R., Zemel, R. S., and Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In ICLR, 2015.


