Graph Optimal Transport for Cross-Domain Alignment

Liqun Chen, Zhe Gan, Yu Cheng, Linjie Li, Lawrence Carin, Jingjing Liu

Presenter: Ngoc Bui - VIDY Reading Group - Feb 2024

Cross-Domain Alignment

- Cross-domain alignment aims to align related entities in different domains.
- In tasks involving different types of data (e.g., images and text), aligning correlated features across modalities helps enhance the learning process with more interpretable alignment.

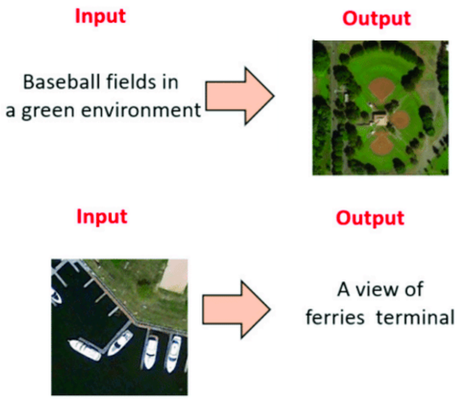
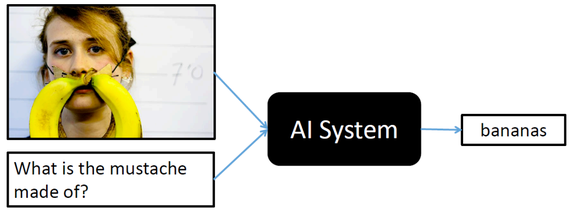
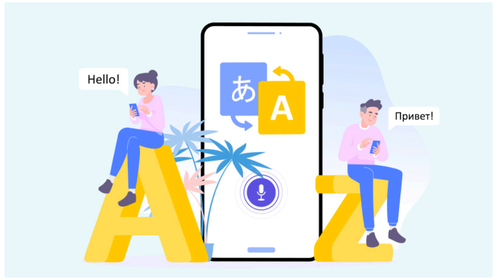


Image-text retrieval



Visual Question Answering (VQA)

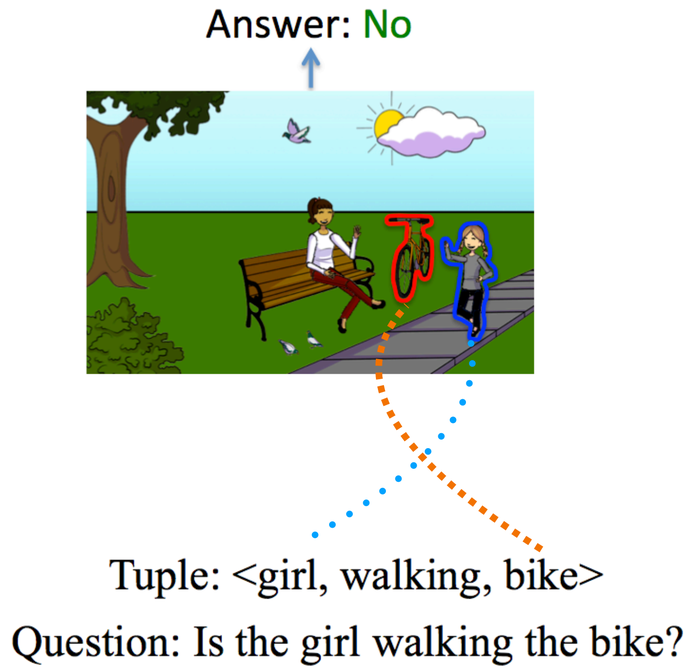


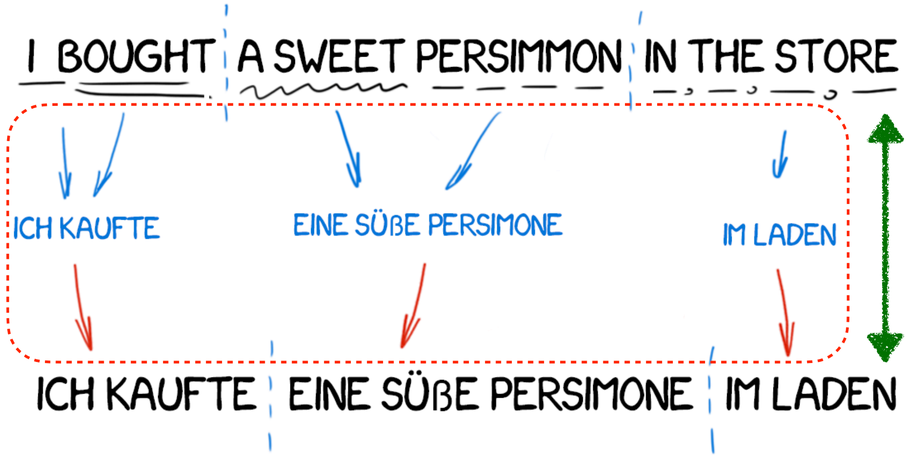
Machine Translation

Speech-To-Text

Motivation

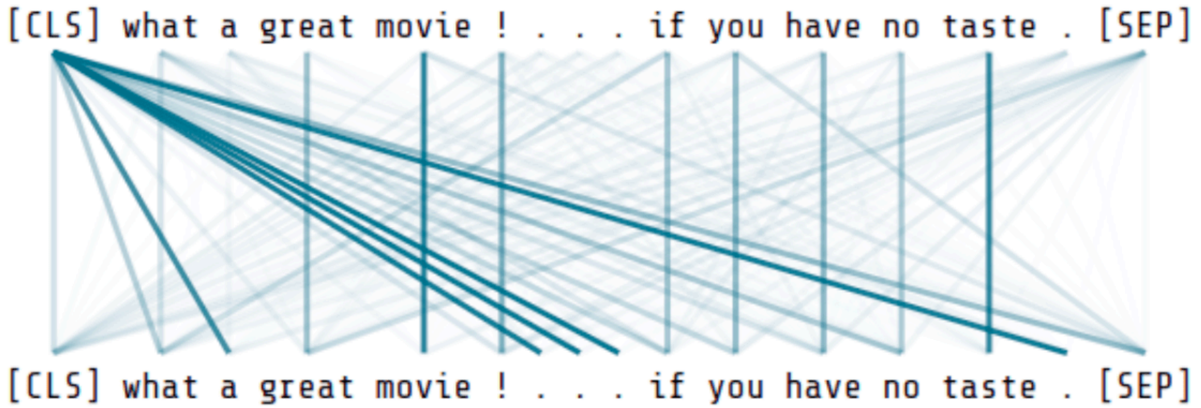
- **Challenge:** No ground truth alignments. Only paired spaces are given (e.g., an image paired with a question)
  - No supervision signal for a “girl” region in an image aligning with the word “girl” in the question





Motivation

- **Challenge:** No ground truth alignments. Only paired spaces are given (e.g., an image paired with a question)
  - No supervision signal for a “girl” region in an image aligning with the word “girl” in the question
- **Attention mechanisms** are guided by task-specific losses; no specific training signal guides/or encourages alignment *explicitly*.
  - Attention matrices are usually dense and less interpretable.



Contributions

- Propose to add an additional term to the training loss using **Graph Optimal Transport (GOT) as a regularizer** to encourage *sparse* alignment for entities across domains.
- Experiments are conducted in several domains/tasks
  - Image-text retrieval
  - Visual question answering
  - Image captioning
  - Machine translation
  - Text summarization

Outline

- Preliminaries: Optimal Transport
- Methodology
- Experiments
- Conclusion

Outline

- Preliminaries: Optimal Transport
- Methodology
- Experiments
- Conclusion

Optimal Transport

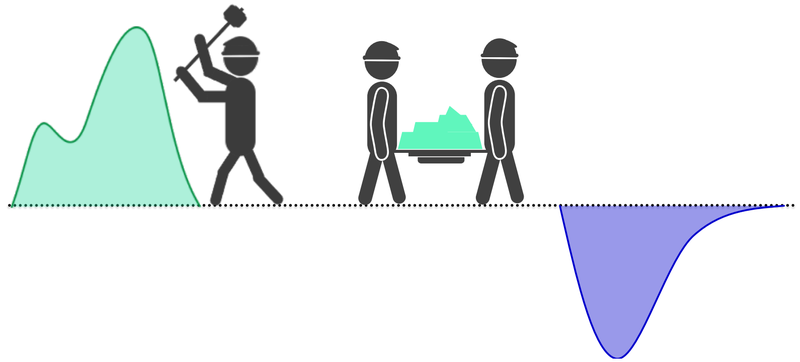
- **Optimal transport problem:**
  - Given a large pile of sand lying on a construction site. A worker aims to move all the sand to a target pile with a prescribed shape.
  - How can the worker minimize her/his total effort?

1781

MÉMOIRE
SUR LA
THÉORIE DES DÉBLAIS
ET DES REMBLAIS.
Par M. MONGE.

LORSQU'ON doit transporter des terres d'un lieu dans un autre, on a coutume de donner le nom de *Déblai* au volume des terres que l'on doit transporter, & le nom de *Remblai* à l'espace qu'elles doivent occuper après le transport.

Le prix du transport d'une molécule étant, toutes choses d'ailleurs égales, proportionnel à son poids & à l'espace qu'on lui fait parcourir, & par conséquent le prix du transport total devant être proportionnel à la somme des produits des molécules multipliées chacune par l'espace parcouru, il s'ensuit que le déblai & le remblai étant donnés de figure & de position, il n'est pas indifférent que telle molécule du déblai soit transportée dans tel ou tel autre endroit du remblai, mais qu'il y a une certaine distribution à faire des molécules du premier dans le second, d'après laquelle la somme de ces produits sera la moindre possible, & le prix du transport total sera un *minimum*.



Optimal Transport

- **Optimal transport problem:**
  - Given a large pile of sand lying on a construction site. A worker aims to move all the sand to a target pile with a prescribed shape.
  - How can the worker minimize her/his total effort?
- **Monge formulation:**

$$\inf \mathbb{E}_{\mu}[c(x, T(x)]$$
$$\text{s.t. } T_{\#}\mu = \nu$$

for continuous distribution

μ, ν : probability distributions

$c(x, y)$: the cost function to move from x to y

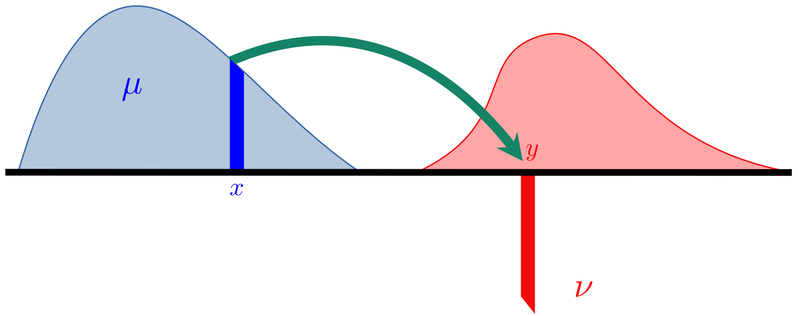
$y = T(x)$: transport map from x to y

1781

MÉMOIRE
SUR LA
THÉORIE DES DÉBLAIS
ET DES REMBLAIS.
Par M. MONGE.

LORSQU'ON doit transporter des terres d'un lieu dans un autre, on a coutume de donner le nom de *Déblai* au volume des terres que l'on doit transporter, & le nom de *Remblai* à l'espace qu'elles doivent occuper après le transport.

Le prix du transport d'une molécule étant, toutes choses d'ailleurs égales, proportionnel à son poids & à l'espace qu'on lui fait parcourir, & par conséquent le prix du transport total devant être proportionnel à la somme des produits des molécules multipliées chacune par l'espace parcouru, il s'ensuit que le déblai & le remblai étant donnés de figure & de position, il n'est pas indifférent que telle molécule du déblai soit transportée dans tel ou tel autre endroit du remblai, mais qu'il y a une certaine distribution à faire des molécules du premier dans le second, d'après laquelle la somme de ces produits sera la moindre possible, & le prix du transport total sera un *minimum*.



Optimal Transport

- **Optimal transport problem:**
  - Given a large pile of sand lying on a construction site. A worker aims to move all the sand to a target pile with a prescribed shape.
  - How can the worker minimize her/his total effort?
- **Monge formulation:**

$$\inf \mathbb{E}_{\mu}[c(x, T(x)] \quad = \min_{T} \frac{1}{n} \sum_{i=1}^{n} c(x_i, T(x_i))$$
$$\text{s.t. } T_{\#}\mu = \nu$$

$$\text{s.t. } v_j = \sum_{i:T(x_i)=y_j} u_i$$

for discrete distributions

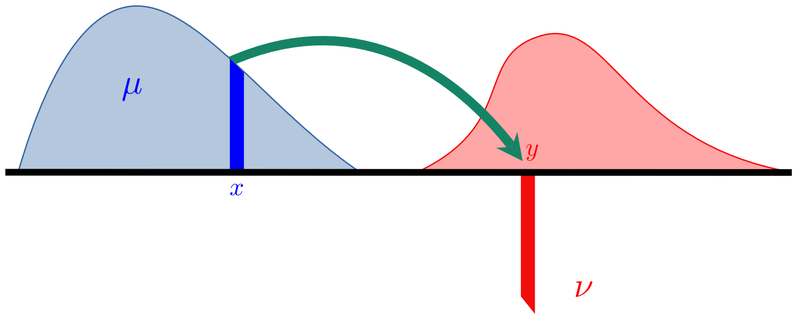
$$\mu = \sum_{i=1}^{n} u_i \delta_{x_i}, \nu = \sum_{i=1}^{m} v_i \delta_{y_i}$$

1781

MÉMOIRE
SUR LA
THÉORIE DES DÉBLAIS
ET DES REMBLAIS.
Par M. MONGE.

LORSQU'ON doit transporter des terres d'un lieu dans un autre, on a coutume de donner le nom de *Déblai* au volume des terres que l'on doit transporter, & le nom de *Remblai* à l'espace qu'elles doivent occuper après le transport.

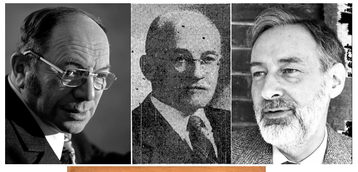
Le prix du transport d'une molécule étant, toutes choses d'ailleurs égales, proportionnel à son poids & à l'espace qu'on lui fait parcourir, & par conséquent le prix du transport total devant être proportionnel à la somme des produits des molécules multipliées chacune par l'espace parcouru, il s'ensuit que le déblai & le remblai étant donnés de figure & de position, il n'est pas indifférent que telle molécule du déblai soit transportée dans tel ou tel autre endroit du remblai, mais qu'il y a une certaine distribution à faire des molécules du premier dans le second, d'après laquelle la somme de ces produits sera la moindre possible, & le prix du transport total sera un *minimum*.

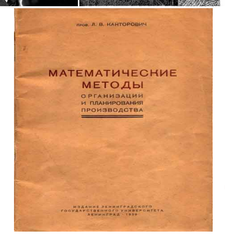


Non-Convexity & Infeasiblity

Optimal Transport

- **Kantorovich relaxation:**
  - Allowing the mass from the distribution μ to be splittable and move to different points in the support of the distribution ν.
  - The optimization problem is more tractable





1941

THE DISTRIBUTION OF A PRODUCT FROM SEVERAL SOURCES TO NUMEROUS LOCALITIES

By Frank L. Hitchcock

1. Statement of the problem. When several factories supply a product to a number of cities we desire the least costly manner of distribution. Due to freight rates and other matters the cost of a ton of product to a particular city will vary according to which factory supplies it, and will also vary from city to city.

OPTIMUM UTILIZATION OF THE TRANSPORTATION SYSTEM*

1949

by Tjalling C. Koopmans

Professor of Economics, The University of Chicago, and Research Associate, Cowles Commission for Research in Economics

The purpose of this paper is to give an application of the theory of optimum allocation of resources to one particular industry. I shall, therefore, not speak on that theory in general. I shall use one of its basic propositions, which was very admirably put forth in the paper presented by M. A l l a i s. This proposition says that a system of prices

- **Monge formulation** <-> **Kantorovich relaxation:**

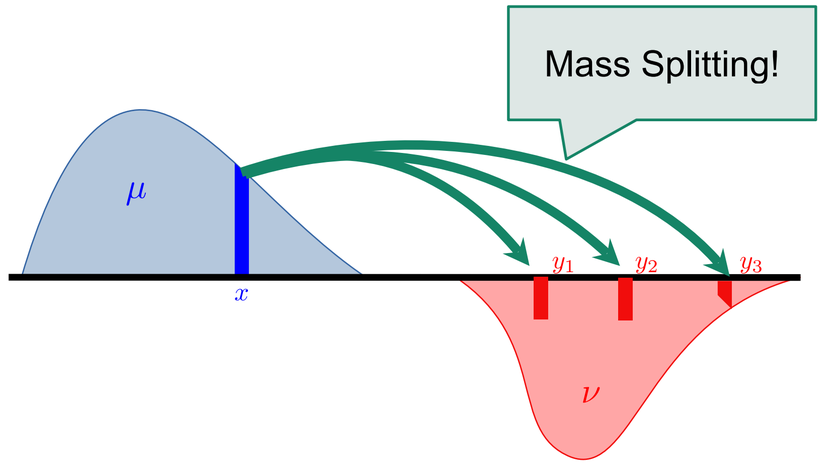
$$\min_{T} \frac{1}{n} \sum_{i=1}^{n} c(x_i, T(x_i))$$

$$\text{s.t. } v_j = \sum_{i:T(x_i)=y_j} u_i$$

$$\min_{\mathbf{T} \in \mathbb{R}^{n \times m}} \sum_{i=1}^{n} \sum_{j=1}^{m} c(x_i, y_i) \mathbf{T}_{ij}$$

$$\text{s.t. } \mathbf{T}\ \mathbf{1}_m = \mathbf{u}$$

$$\mathbf{T}^\top \mathbf{1}_n = \mathbf{v}$$

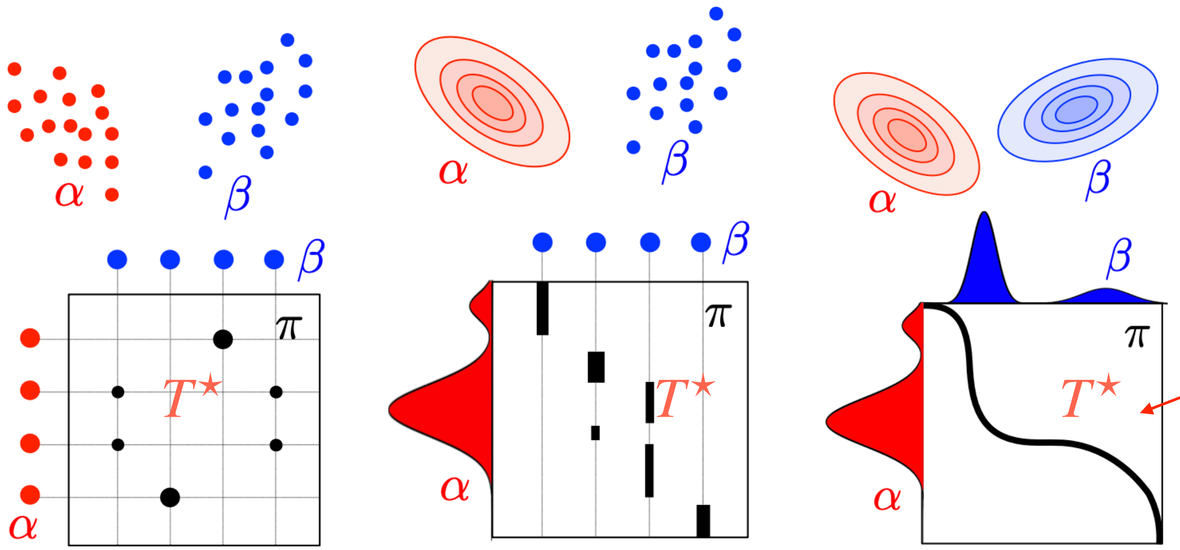


$$\mu = \sum_{i=1}^{n} u_i \delta_{x_i}, \nu = \sum_{i=1}^{m} v_i \delta_{y_i}$$

Linear Program

Optimal Transport

- **Some cases of optimal transport**



the optimal transport plan

usually easier to solve

need numerical or approximations

We will focus on discrete distribution only

Graph matching

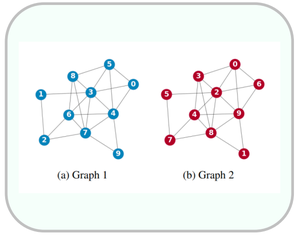
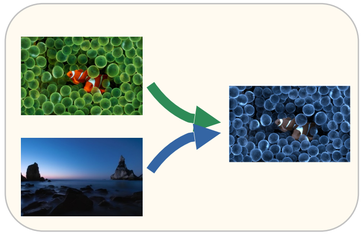
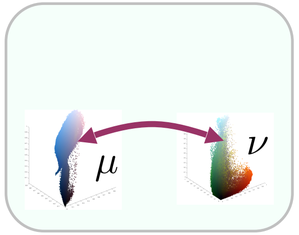


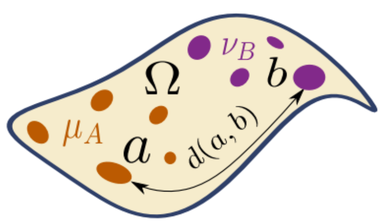
Image processing

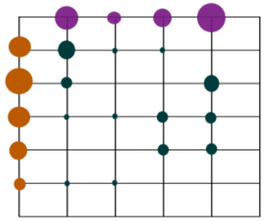


Generative models



Wasserstein Distance





- **One feature of OT** is that it defines the distance between histograms or distributions

Definition (p-Wasserstein distance): Let $\mu = \sum_i^n u_i \delta_{\mathbf{x}_i}, \nu = \sum_j^m v_j \delta_{\mathbf{y}_j}$ be two discrete distributions, the p-Wasserstein distance is defined as

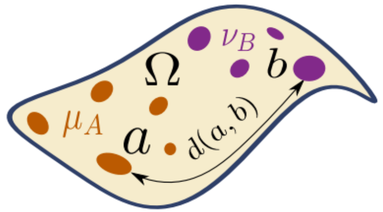
$$W_p(\mu, \nu) = \left(\min_{\mathbf{T} \in \Pi(\mu,\nu)} \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbf{T}_{ij} c(\mathbf{x}_i, \mathbf{y}_j)^p \right)^{1/p}$$

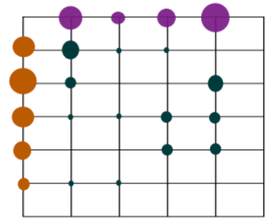
If $p = 1$, we have Kantorovich relaxation <-> Earth Mover distance

where $\Pi(\mu, \nu) = \{\mathbf{T} \in R_+^{n \times m} | \mathbf{T}\mathbf{1}_m = \mathbf{u}, \mathbf{T}^\top \mathbf{1}_n = \mathbf{v}\}$, denotes joint distribution with marginals $\mu(\mathbf{x}), \nu(\mathbf{y})$.

- Wasserstein satisfies the triangle inequality $W_p(\mu, \nu) \leq W_p(\mu, \gamma) + W_p(\gamma, \nu)$
- $W_p(\mu, \nu) = 0 \Leftrightarrow \mu = \nu$

Computation - W

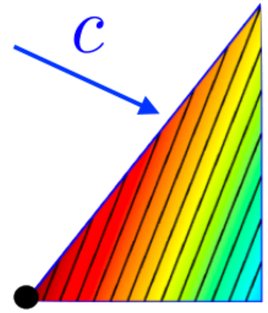




1-Wasserstein dist. (p = 1).

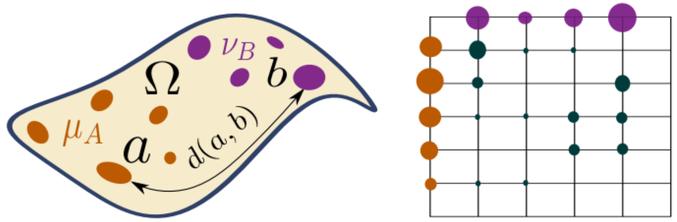
$$W_1(\mu,\nu) = \min_{\mathbf{T}\in\Pi(\mu,\nu)} \sum_{i,j} \mathbf{T}_{ij} c(\mathbf{x}_i, \mathbf{y}_j)$$

- **Linear programming**
  - **Complexity:** $\mathcal{O}(n^{2+\tau})$: $\tau \approx 0.5$, n is the number of variables.
  - **Cons**:
    - Often yields really ***sparse*** solutions as linear programs with a non-empty and bounded feasible set attain their minimum at an extremal point of the feasible set.
    - Sparse is good, but too sparse is not good because the actual transport plans in practice are usually more diffuse than the ones predicted by OT.



linear objective function on a simplex space

Computation - W



1-Wasserstein dist. (p = 1).

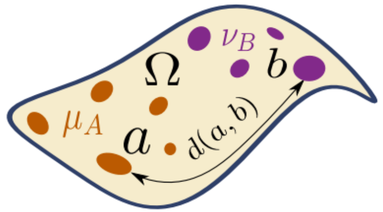
$$W_1^\epsilon(\mu,\nu) = \min_{\mathbf{T}\in\Pi(\mu,\nu)} \sum_{i,j} \mathbf{T}_{ij} c(\mathbf{x}_i, \mathbf{y}_j) + \epsilon \sum_{ij} \log \mathbf{T}_{ij} \qquad (*)$$

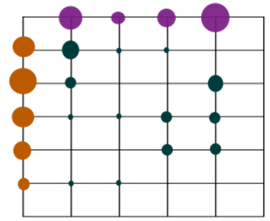
$-\epsilon H(\mathbf{T})$

- **Entropic Regularization**
  - Penalize the negative entropy of $\mathbf{T}$ to control the sparsity of the transport plan (alignment).

    Lemma. As $\epsilon \to 0$, we have $W_1^\epsilon(\mu,\nu) \to W_1(\mu,\nu)$
  - For small enough ϵ, W_1^ϵ approximates the solution of the Kantorovich problem.
  - **Pros**:
    - The ***approximate distance is smooth*** w.r.t. input histogram weights and position of the Diracs. This is ***important in DL applications*** where we use OT as the objective and want to compute the gradient w.r.t. the input positions.
    - ***Easier to solve*** with a simpler alternate minimization scheme (Sinkhorn algo.)

Computation - W



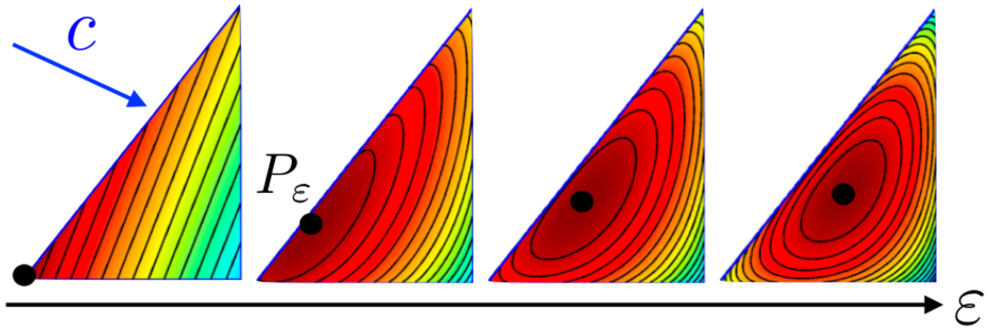


1-Wasserstein dist. (p = 1).

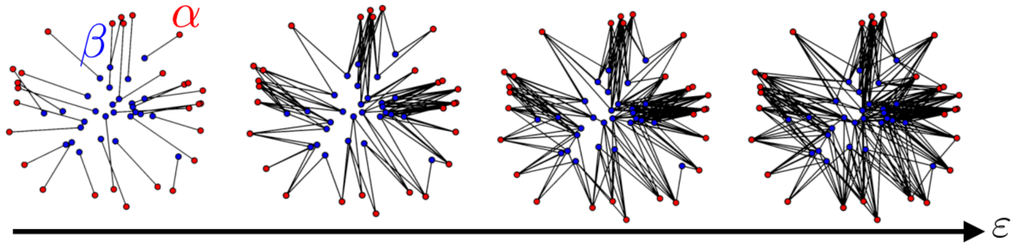
$$W_1^{\epsilon}(\mu,\nu) = \min_{\mathbf{T}\in\Pi(\mu,\nu)} \sum_{i,j} \mathbf{T}_{ij} c(\mathbf{x}_i, \mathbf{y}_j) + \epsilon \sum_{ij} \log \mathbf{T}_{ij} \qquad (*)$$

$-\epsilon H(\mathbf{T})$

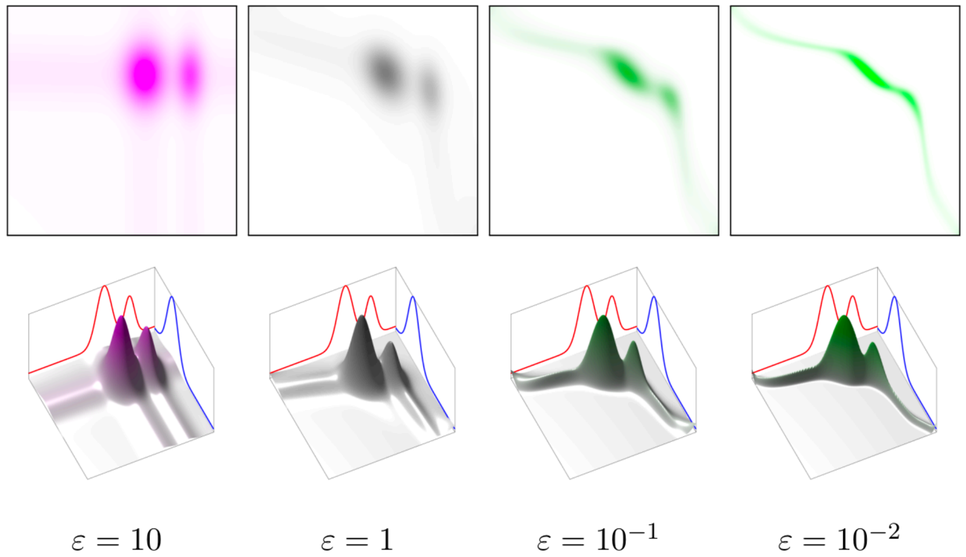
- **Entropic Regularization**



Optimization landscape

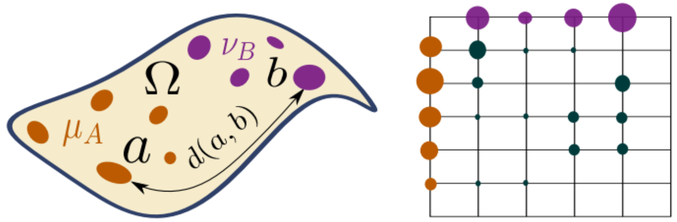


Transport plan in a discrete OT



Transport plan in a continuous OT

Computation - W



1-Wasserstein dist. (p = 1).

$$W_1^{\epsilon}(\mu,\nu) = \min_{\mathbf{T}\in\Pi(\mu,\nu)} \sum_{i,j} \mathbf{T}_{ij} c(\mathbf{x}_i, \mathbf{y}_j) + \epsilon \sum_{ij} \log \mathbf{T}_{ij} \qquad (*)$$

$-\epsilon H(\mathbf{T})$

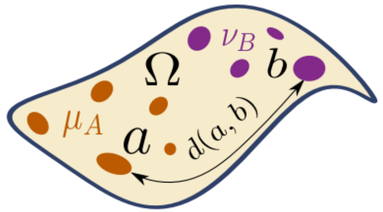
- **Sinkhorn Algorithm**

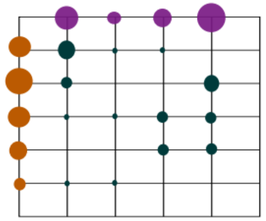
Theorem. (*) has a unique solution in the form $\mathbf{T} = \mathrm{diag}(\mathbf{a})\mathbf{K}\,\mathrm{diag}(\mathbf{b})$ where $\mathbf{K}$ is a matrix with elements defined as $\mathbf{K}_{ij} = \exp(-\frac{c(\mathbf{x}_i,\mathbf{y}_j)}{\epsilon})$ and $\mathbf{a}$ and $\mathbf{b}$ are scaling vectors satisfying the following system

$$\mathrm{diag}(\mathbf{a})\mathbf{K}\mathbf{b} = \mathbf{u} \qquad (1)$$

$$\mathrm{diag}(\mathbf{b})\mathbf{K}^{\top}\mathbf{a} = \mathbf{v} \qquad (2)$$

Computation - W





1-Wasserstein dist. (p = 1).

$$W_1^{\epsilon}(\mu,\nu) = \min_{\mathbf{T}\in\Pi(\mu,\nu)} \sum_{i,j} \mathbf{T}_{ij} c(\mathbf{x}_i, \mathbf{y}_j) + \epsilon \sum_{ij} \log \mathbf{T}_{ij} \qquad (*)$$

$-\epsilon H(\mathbf{T})$

- **Sinkhorn Algorithm**

Theorem. (*) has a unique solution in the form $\mathbf{T} = \mathrm{diag}(\mathbf{a})\mathbf{K}\,\mathrm{diag}(\mathbf{b})$ where $\mathbf{K}$ is a matrix with elements defined as $\mathbf{K}_{ij} = \exp(-\frac{c(\mathbf{x}_i,\mathbf{y}_j)}{\epsilon})$ and $\mathbf{a}$ and $\mathbf{b}$ are scaling vectors satisfying the following system

$$\mathrm{diag}(\mathbf{a})\mathbf{K}\mathbf{b} = \mathbf{u} \qquad (1)$$

$$\mathrm{diag}(\mathbf{b})\mathbf{K}^{\top}\mathbf{a} = \mathbf{v} \qquad (2)$$

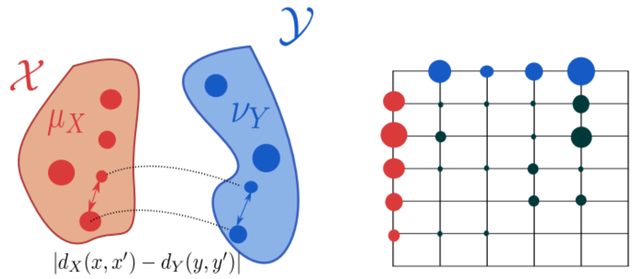
- **Procedure**: Initialize $\mathbf{b}^{(0)} = \mathbf{1}$
  Repeat
  1. Solve (1) by $\mathbf{a}^{(l+1)} = \mathbf{u}/(\mathbf{K}\mathbf{b}^{(l)})$
  2. Solve (2) by $\mathbf{b}^{(l+1)} = \mathbf{v}/(\mathbf{K}^{\top}\mathbf{a}^{(l+1)})$

- After L iterations

$$\mathbf{T} = \mathrm{diag}(\mathbf{a}^{(L)})\mathbf{K}\,\mathrm{diag}(\mathbf{b}^{(L)})$$

- **Complexity**: $\mathcal{O}(Ln^2)$

Gromov-Wasserstein Distance



- **Wasserstein dist.** requires a cost function between two points $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.
- **Gromov-Wasserstein dist.** is designed to measure the distance of two distributions from different domains where a cost function between two spaces is not trivial.

Definition (Gromov-Wasserstein distance): Let $\mu = \sum_i^n u_i \delta_{\mathbf{x}_i}, \nu = \sum_j^m v_j \delta_{\mathbf{y}_j}$ be two discrete distributions, the Gromov-Wasserstein distance is defined as

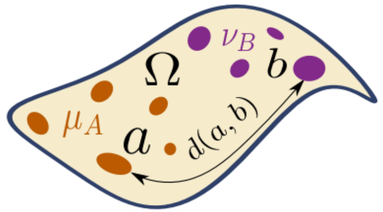
$$GW_p(\mu, \nu) = \left(\min_{\mathbf{T} \in \Pi(\mu,\nu)} \sum_{i,i',j,j'} \mathbf{T}_{ij} \mathbf{T}_{i'j'} L(\mathbf{x}_i, \mathbf{y}_j, \mathbf{x}_{i'}, \mathbf{y}_{j'})^p \right)^{1/p}$$

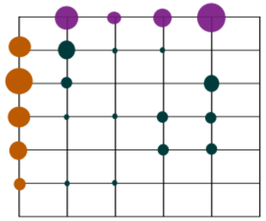
where $\Pi(\mu, \nu) = \{\mathbf{T} \in R_+^{n \times m} | \mathbf{T}\mathbf{1}_m = \mathbf{u}, \mathbf{T}^\top \mathbf{1}_n = \mathbf{v}\}$, denotes joint distribution with marginals $\mu(\mathbf{x}), \nu(\mathbf{y})$ and $L(\mathbf{x}_i, \mathbf{y}_i, \mathbf{x}_{i'}, \mathbf{y}_{j'}) = |d_{\mathcal{X}}(\mathbf{x}_i, \mathbf{x}_{i'}) - d_{\mathcal{Y}}(\mathbf{y}_i, \mathbf{y}_{i'})|$.

- GW tends to associate pairs of points with similar distances together.

distance metrics in $\mathcal{X}, \mathcal{Y}$ respectively.

Computation - GW



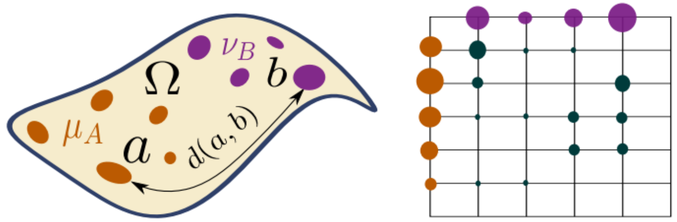


1-Gromove-Wasserstein dist.

$$GW_1(\mu,\nu) = \min_{\mathbf{T}\in\Pi(\mu,\nu)} \sum_{i,i',j,j'} \mathbf{T}_{ij}\mathbf{T}_{i'j'}L(\mathbf{x}_i,\mathbf{y}_j,\mathbf{x}_{i'},\mathbf{y}_{j'})$$

- **Quadratic programming —> Difficult to solve**
  - **Complexity:** The objective is non-convex -> NP-hard.

Computation - GW



1-Gromove-Wasserstein dist.

$$GW_1(\mu,\nu) = \min_{\mathbf{T}\in\Pi(\mu,\nu)} \sum_{i,i',j,j'} \mathbf{T}_{ij}\mathbf{T}_{i'j'}L(\mathbf{x}_i,\mathbf{y}_j,\mathbf{x}_{i'},\mathbf{y}_{j'}) - \epsilon H(\mathbf{T})$$

- **Quadratic programming —> Difficult to solve**
  - **Complexity:** The objective is non-convex -> NP-hard.
- **Pseudo cost matrix + Iterative Sinkhorn algorithm**
  - Using projected gradient descent to solve the entropic GW and choosing the proper parameters, it is proved that each iteration of this algorithm reads

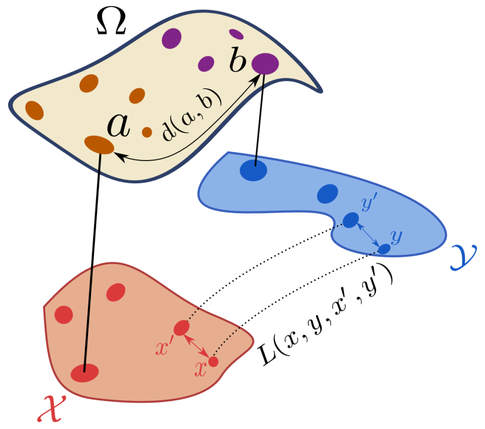
Can be solved by Sinkhorn algorithm

$$\mathbf{T}^{(k+1)} = \arg\min_{\mathbf{T}\in\Pi(\mu,\nu)} \langle \mathbf{L}^{(k)}, \mathbf{T}\rangle$$

where $\mathbf{L}^{(k)}$ is a *pseudo cost matrix* defined by $\mathbf{L}^{(k)}_{ij} = \sum_{i'j'} L(\mathbf{x}_i,\mathbf{y}_j,\mathbf{x}_{i'},\mathbf{y}_{j'})\mathbf{T}^{(k)}_{i'j'}$

  - No known convergence guarantee, but it performs well in practice.

Fused Gromov-Wasserstein Distance



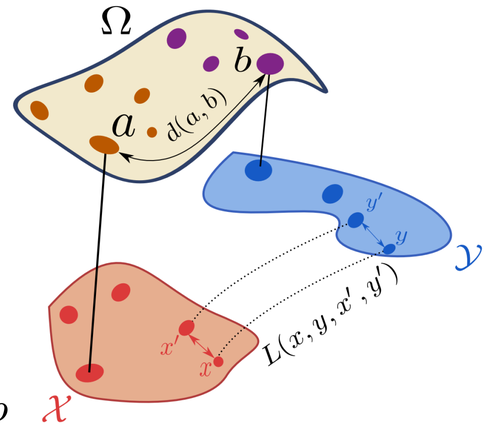
- **Wasserstein dist.** focuses on comparing node's similarity.
- **Gromov-Wasserstein dist.** focuses on comparing edge's similarity.
- **Fused Gromov-Wasserstein dist.** combines them to account for both nodes and edges.

Definition (Fused Gromov-Wasserstein): Let $\mu = \sum_i^n u_i \delta_{\mathbf{x}_i}, \nu = \sum_j^m v_j \delta_{\mathbf{y}_j}$ be two discrete distributions, the Fused Gromov-Wasserstein distance is defined as

$$FGW_{p,\lambda}(\mu,\nu) = \left(\min_{\mathbf{T}\in\Pi(\mu,\nu)} \sum_{i,i',j,j'} \mathbf{T}_{ij}\lambda c_\Omega(\mathbf{x}_i,\mathbf{y}_j)^p + (1-\lambda)\mathbf{T}_{ij}\mathbf{T}_{i'j'}L(\mathbf{x}_i,\mathbf{y}_i,\mathbf{x}_{i'},\mathbf{y}_{j'})^p \right)^{1/p}$$

where $\Pi(\mu,\nu) = \{\mathbf{T} \in R_+^{n\times m} | \mathbf{T}\mathbf{1}_m = \mathbf{u}, \mathbf{T}^\top \mathbf{1}_n = \mathbf{v}\}$, denotes joint distribution with marginals $\mu(\mathbf{x}), \nu(\mathbf{y})$, $L(\mathbf{x}_i,\mathbf{y}_i,\mathbf{x}_{i'},\mathbf{y}_{j'}) = |d_\mathcal{X}(\mathbf{x}_i,\mathbf{x}_{i'}) - d_\mathcal{Y}(\mathbf{y}_i,\mathbf{y}_{i'})|$, and $c_\Omega(\mathbf{x}_i,\mathbf{y}_j)$ is the cost function to transport $\mathbf{x}_i$ to $\mathbf{y}_j$ in a common space Ω.

Fused Gromov-Wasserstein Distance



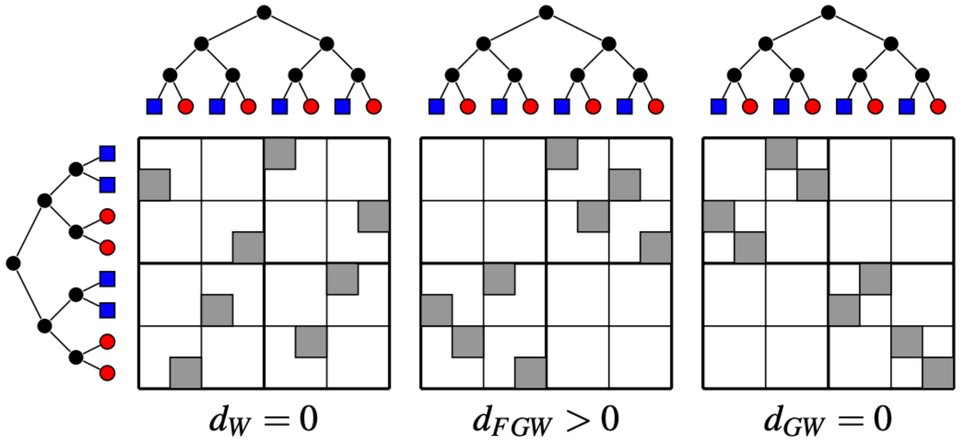
- **Properties of FGW:**
  - $FGW_{p,\lambda}(\mu,\nu) \geq \lambda W_p(\mu_\Omega,\nu_\Omega)^p \quad \text{and} \quad FGW_{p,\lambda}(\mu,\nu) \geq (1-\lambda) GW_p(\mu,\nu)^p$
  - If the FGW distance is zero, then GW and W distances vanish.
    The converse is not necessarily true

- **Toy Example:**
  - Distance between two nodes used for Wasserstein: Euclidean distance
  - Distance between two nodes used for Gromov-Wasserstein: The shortest path.

FGW > 0: the bottom and first level structure is preserved, as well as the feature matching (red on red and blue on blue), and the cost to transport the topology is not zero -> FGW discriminates two structures

W = 0: red nodes are transported on red ones and the blue nodes on the blue ones.

But tree structures are completely discarded



GW = 0: all couples of points are transported to another couple of points

It matches the tree structure without taking into account the features

Fused Gromov-Wasserstein Distance

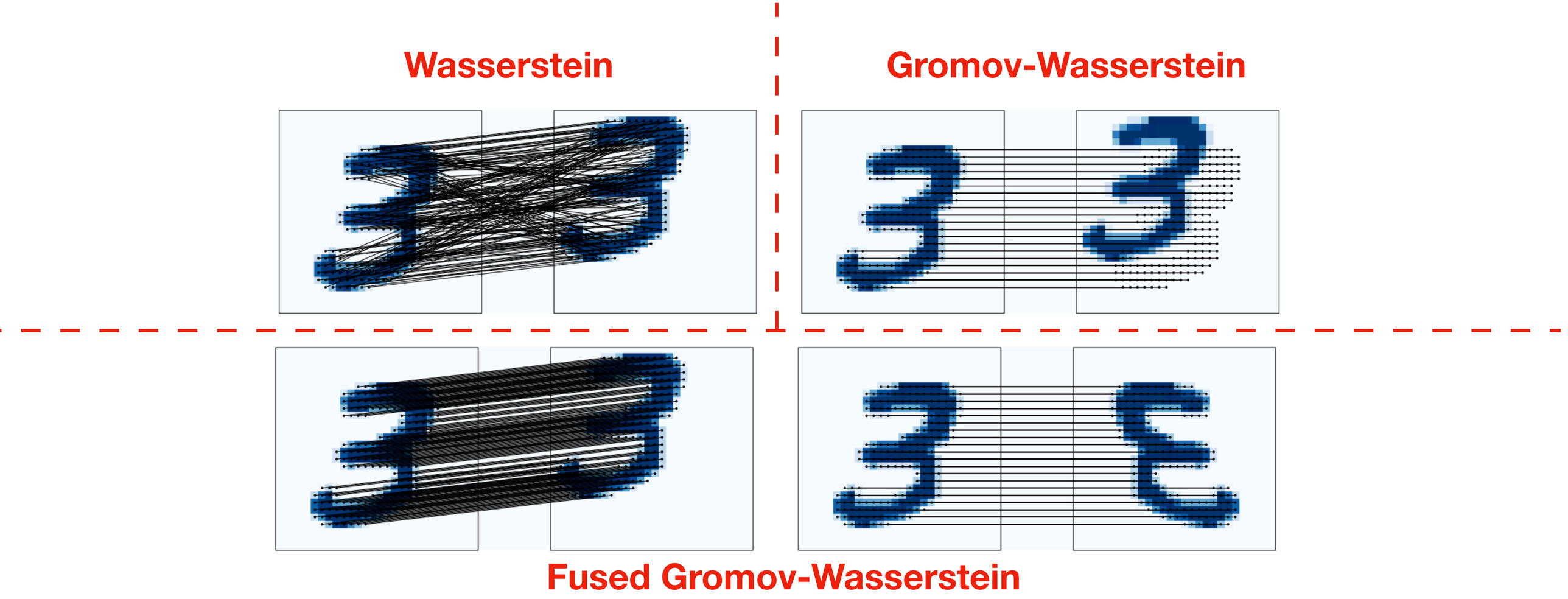


FIG. 10. Couplings obtained when considering (Top left) the features only, where we have $d^{\Omega}_{W,1} = 0$ (Top right) the structure only, with $d_{GW,1} = 0$ (Bottom left and right) both the features and the structure, with $d^{\Omega}_{FGW,0.1,1,2}$. For readibility issues, only the couplings starting from non white pixels on the left picture are depicted.

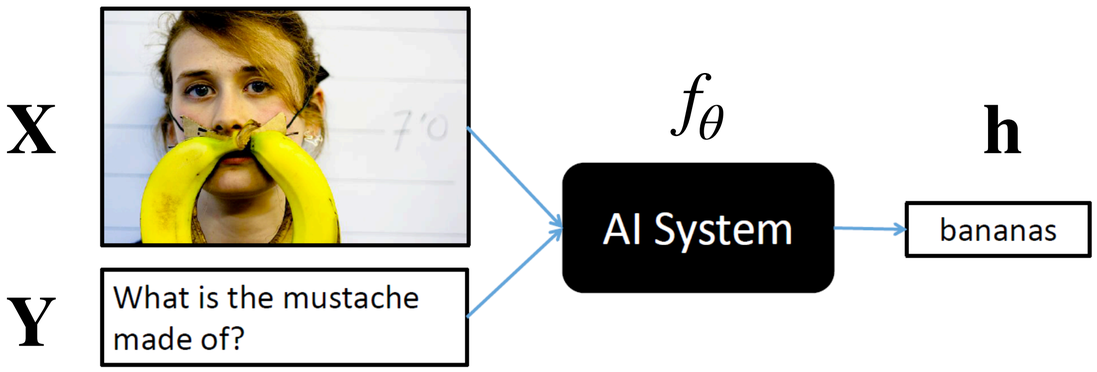
Outline

- Preliminaries: Optimal Transport
- Methodology
- Experiments
- Conclusion

Settings

- **Input**: Two sets of entities $\mathbf{X} = \{\mathbf{x}_i\}_n, \mathbf{Y} = \{\mathbf{y}_j\}_m$ in different domains ($\mathbf{x}_i \in \mathbb{D}_1, \mathbf{y}_i \in \mathbb{D}_2$)
- **Model**: $f_\theta(\mathbf{X}, \mathbf{Y}) \rightarrow \mathbf{h}$, where $\mathbf{h}$ is the output signal.
- **Supervision loss**:

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



Visual Question Answering (VQA)

$\mathcal{L}_{\text{sup}}(\mathbf{X}, \mathbf{Y}, \mathbf{h})$ is entropy-loss

Graph Optimal Transport Framework

- **Idea**: add an additional term to the training loss using **Graph Optimal Transport (GOT) as a regularizer** to encourage *sparse* alignment for entities across domains.

- **Supervision loss** + Cross-domain alignment (CDA) loss:

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

where

$$\mathcal{L}_{\text{CDA}}(\mathbf{X}, \mathbf{Y}) = D_{\text{FGW}}(\mu, \nu) = \inf_{\mathbf{T} \in \Pi(\mu,\nu)} \sum_{i,i',j,j'} \mathbf{T}_{ij} \left(\lambda c(\mathbf{x}_i, \mathbf{y}_j + \lambda \mathbf{T}_{i'j'} L(\mathbf{x}_i, \mathbf{y}_j, \mathbf{x}_{i'}, \mathbf{y}_{j'})\right)$$

Construct graph/distribution from sets of entities????

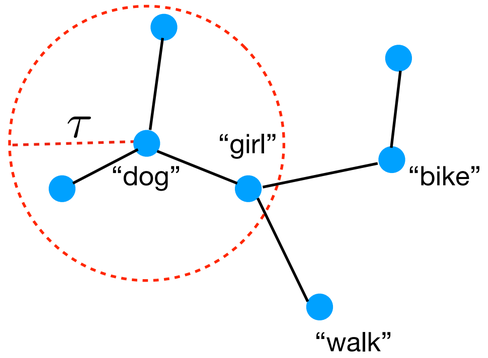
the same transport plan

Graph/Distribution Construction

- **Dynamic graph construction**: Given a set of entities $\mathbf{X} = \{\mathbf{x}_i\}_n \rightarrow \mathcal{G}_x(V_x, E_x)$ we use a unit (τ) disk graph to construct the graph representation
  - Define $\mathbf{C}^x = \{\cos(\mathbf{x}_i, \mathbf{x}_j) - \tau\}_{ij} \in \mathbb{R}^{n \times n}$
  - The set of edges: $E_x = \{ij | \mathbf{C}^x_{ij} > 0\}$
  - The term “Dynamic” is because feature vectors $\mathbf{X}$ will be changed during learning.

???

$\mathbf{C}^x = \max(\mathbf{C}^x, 0)$ ⟶ cost matrix



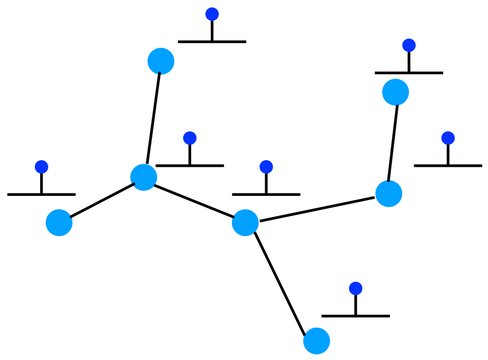
Graph/Distribution Construction

- **Dynamic graph construction**: Given a set of entities $\mathbf{X} = \{\mathbf{x}_i\}_n \rightarrow \mathcal{G}_x(V_x, E_x)$ we use a unit (τ) disk graph to construct the graph representation
  - Define $\mathbf{C}^x = \{\cos(\mathbf{x}_i, \mathbf{x}_j) - \tau\}_{ij} \in \mathbb{R}^{n \times n}$
  - The set of edges: $E_x = \{ij | \mathbf{C}^x_{ij} > 0\}$
  - The term “Dynamic” is because feature vectors $\mathbf{X}$ will be changed during learning.
- **Distribution construction**: construct a discrete distribution over the set of entities

$$\mathbf{X} = \{\mathbf{x}_i\}_n \rightarrow \mu = \sum_{i}^{n} u_i \delta_{\mathbf{x}_i}$$

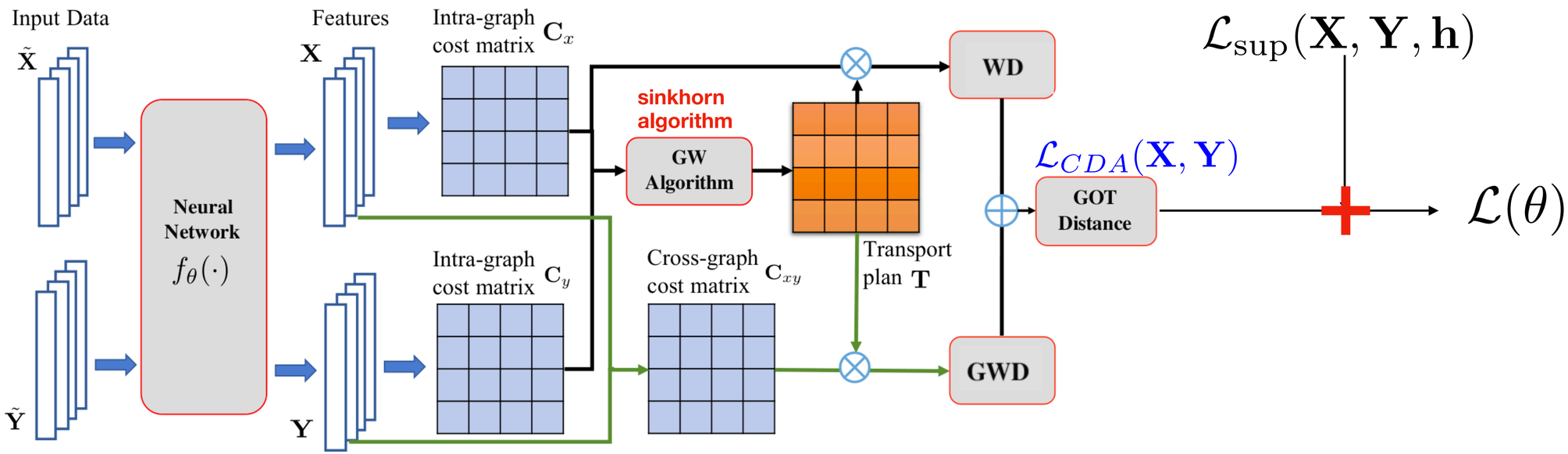
Dirac function

$u_i = \frac{1}{n}$: **uniform distribution**



Computation Graph

- Forward-pass



- Backward-pass

$$\nabla_\theta \mathcal{L}(\theta) = \nabla_\theta \mathcal{L}_{\text{sup}}(\mathbf{X}, \mathbf{Y}, \mathbf{h}) + \nabla_\theta \mathcal{L}_{\text{CDA}}(\mathbf{X}, \mathbf{Y}; \mathbf{T}^\star)$$

the optimal transport plan

Outline

- Preliminaries: Optimal Transport
- Methodology
- Experiments
- Conclusion

Image Text Retrieval

- **For image**: use Faster-RNN to extract bottom-up features (36 features $\in \mathbb{R}^{2048}$).
- **For caption**: use bi-directional GRU to extract text features.
- **Evaluation datasets**: Flickr30K and COCO.
- **Metric**: Recall@K : the percentage of queries retrieving the correct images/sentences within the top K highest-ranked results.

Image Text Retrieval

- **Both** SCAN + WD and +GWD improve the performance but WD has a larger margin than GWD
- **GOT (WD + GWD)** improves SCAN's performance significantly.
- The improvement of GOT compared to WD is **not** significant.

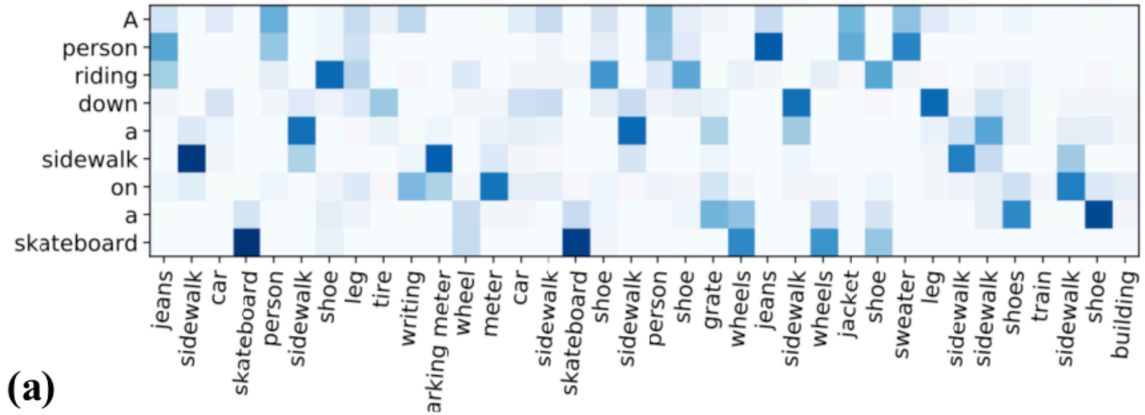
Method	Sentence Retrieval R@1	R@5	R@10	Image Retrieval R@1	R@5	R@10	Rsum
VSE++ (ResNet) (Faghri et al., 2018)	52.9	–	87.2	39.6	–	79.5	–
DPC (ResNet) (Zheng et al., 2020)	55.6	81.9	89.5	39.1	69.2	80.9	416.2
DAN (ResNet) (Nam et al., 2017)	55.0	81.8	89.0	39.4	69.2	79.1	413.5
SCO (ResNet) (Huang et al., 2018)	55.5	82.0	89.3	41.1	70.5	80.1	418.5
SCAN (Faster R-CNN, ResNet) (Lee et al., 2018)	67.7	88.9	94.0	44.0	74.2	82.6	452.2
Ours (Faster R-CNN, ResNet):							
SCAN + WD	70.9	92.3	95.2	49.7	78.2	86.0	472.3
SCAN + GWD	69.5	91.2	95.2	48.8	78.1	85.8	468.6
SCAN + GOT	**70.9**	**92.8**	**95.5**	**50.7**	**78.7**	**86.2**	**474.8**
VSE++ (ResNet) (Faghri et al., 2018)	41.3	–	81.2	30.3	–	72.4	–
DPC (ResNet) (Zheng et al., 2020)	41.2	70.5	81.1	25.3	53.4	66.4	337.9
GXN (ResNet) (Gu et al., 2018)	42.0	–	84.7	31.7	–	74.6	–
SCO (ResNet) (Huang et al., 2018)	42.8	72.3	83.0	33.1	62.9	75.5	369.6
SCAN (Faster R-CNN, ResNet)(Lee et al., 2018)	46.4	77.4	87.2	34.4	63.7	75.7	384.8
Ours (Faster R-CNN, ResNet):							
SCAN + WD	50.2	80.1	89.5	37.9	66.8	78.1	402.6
SCAN + GWD	47.2	78.3	87.5	34.9	64.4	76.3	388.6
SCAN + GOT	**50.5**	**80.2**	**89.8**	**38.1**	**66.8**	**78.5**	**403.9**

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

Image Text Retrieval

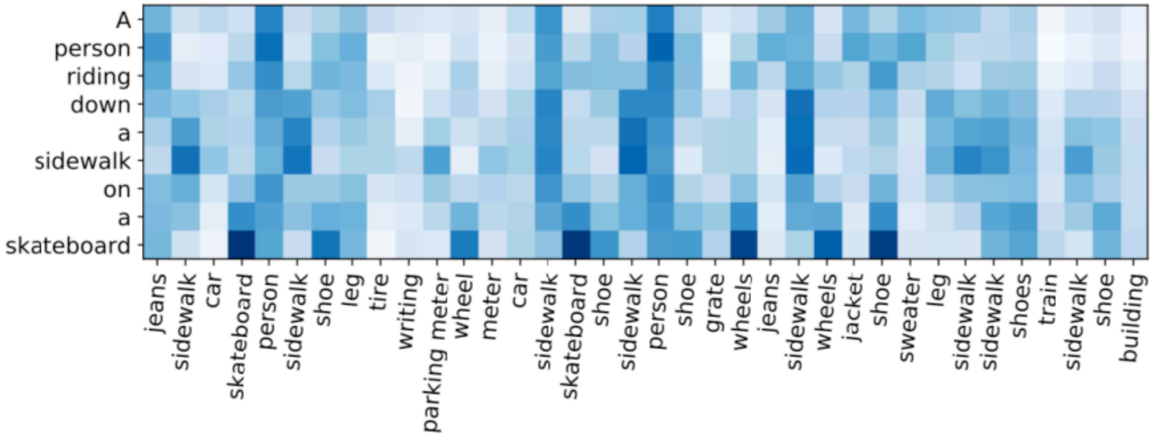
- **GOT** has a ***sparser*** transport plan compared to the attention matrix from SCAN

SCAN + GOT



(a)

SCAN



Vision Question Answering

- **Model**: BAN model and BUTD
- **Evaluation datasets**: VQA 2.0 dataset (human-annotated QA on COCO)
- **Metric**: Accuracy
- **Observation**:
  - GOT improves the performance slightly.
  - Performance gains are more significant on small models (BUTD)

Model	BAN	BAN+GWD	BAN+WD	BAN+GOT
Score	66.00	66.21	66.26	**66.44**

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

Model	BUTD	BAN-1	BAN-2	BAN-4	BAN-8
w/o GOT	63.37	65.37	65.61	65.81	66.00
w/ GOT	**65.01**	**65.68**	**65.88**	**66.10**	**66.44**

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.

Image Captioning

- **Image encoder**: use Faster-RNN to extract bottom-up features (36 features $\in \mathbb{R}^{2048}$).
- **Text encoder**: LSTM with 256 hidden units.
- **Evaluation datasets**: VQA 2.0 dataset (human-annotated QA on COCO)
- **Metrics**: CIDEr, BLUE-x, ROUGE, METEOR

Image Captioning

- **GOT** improves MLE
  - MLE: trained with maximum likelihood estimation objective.
  - MLE + GOT: The same base model as MLE but trained with an additional CDA regularizer.
- **Observation**: Again, the performance gap compared to MLE is not significant.

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

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

Machine Translation

- **Base model**: Transformer model
- **Evaluation datasets**:
  - English-Vietnamese TED-talks corpus (133K pairs of sentences from the IWSLT Evaluation Campaign)
  - English-German parallel corpus with 4.5M sentences from WMT Evaluation Campain
- **Metrics**: BLEU
- **Observation**: the same as other experiments

Model	EN-VI uncased	EN-VI cased	EN-DE uncased	EN-DE cased
Transformer (Vaswani et al., 2017)	29.25 ± 0.18	28.46 ± 0.17	25.60 ± 0.07	25.12 ± 0.12
Transformer + WD	29.49 ± 0.10	28.68 ± 0.14	25.83 ± 0.12	25.30 ± 0.11
Transformer + GWD	28.65 ± 0.14	28.34 ± 0.16	25.42 ± 0.17	24.82 ± 0.15
Transformer + GOT	$\mathbf{29.92 \pm 0.11}$	$\mathbf{29.09 \pm 0.18}$	$\mathbf{26.05 \pm 0.17}$	$\mathbf{25.54 \pm 0.15}$

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

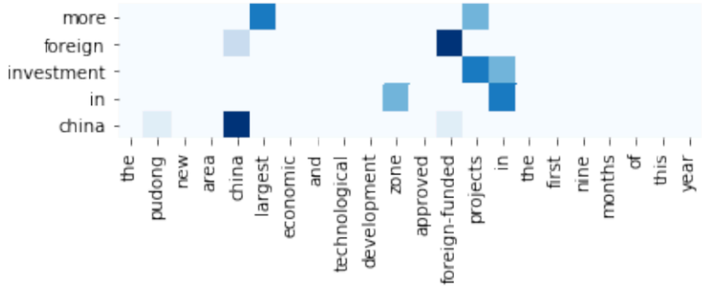


Figure 4. Inferred transport plan for aligning source and output sentences in abstractive summarization.

Machine Translation

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.

good

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.

bad

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"*.

Abstractive Summarization

- **Base model**: LSTM model
- **Evaluation datasets**: English Gigawords
- **Metrics**: ROUGE-1, -2, -L
- **Observation**: the same as other experiments

Method	ROUGE-1	ROUGE-2	ROUGE-L
ABS+ (Rush et al., 2015)	31.00	12.65	28.34
LSTM (Hu et al., 2018)	36.11	16.39	32.32
LSTM + GWD	36.31	17.32	33.15
LSTM + WD	36.81	17.34	33.34
LSTM + GOT	**37.10**	**17.61**	**33.70**

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

more <-> largest

investment <-> project

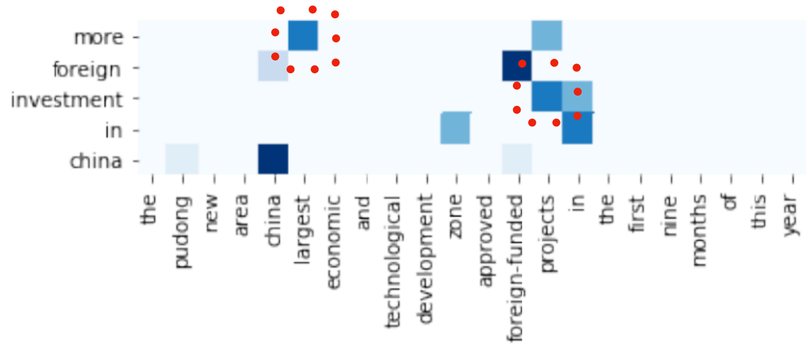


Figure 4. Inferred transport plan for aligning source and output sentences in abstractive summarization.

Ablation Study

- **Settings**: use the same settings as in the machine translation task.

- **Shared vs unshared transport plan $\mathbf{T}$**:

Model	EN-VI uncased	EN-DE uncased
GOT (shared)	29.92 ± 0.11	26.05 ± 0.18
GOT (unshared)	29.77 ± 0.12	25.89 ± 0.17

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

- **Hyperparameters λ**:

λ	0	0.1	0.3	0.5	0.8	1.0
BLEU	28.65	29.31	29.52	29.65	**29.92**	29.49

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

Outline

- Preliminaries: Optimal Transport
- Methodology
- Experiments
- Conclusion

Conclusion

- **Contribution**: a new framework ***explicitly*** promoting the ***sparse*** alignment among entities across different domains via optimal transport.
- **Experiments** are conducted across several domains/tasks.
- **Weakness**:
  - The effects of “graph” or “dynamic graph” are unclear.
  - Computationally expensive: the complexity of the Sinkhorn algorithm is $\mathcal{O}(n^2T)$, where T is the number of iterations.
  - Experiments are conducted where the number of entities is relatively small. High-dimensional data (e.g., images) need object detection. End-to-end architectures are more preferred nowadays.
  - Performance gain compared to the cost is low. The advantage of combining GWD is not really significant.


THANK YOU

Vision Question Answering

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 $\min(1, \frac{\text{\# humans provided ans.}}{3})$.

Image Captioning

- **CIDEr (Consensus-based Image Description Evaluation)**: measures the similarity between the generated sentence and the human's sentence via a weighted cosine similarity of TF-IDF representation of two sentences.

- **BLEU-x (Bilingual Evaluation Understudy)**: the precision of the x-grams in the generated caption that appears in the reference caption, and then applies a brevity penalty to discourage overly short captions.

- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation)**: mainly used for automatic summarization and machine translation. ROUGE measures the overlap between the generated caption and the reference caption. Several variants: ROUGE-N (overlap of N-grams), ROUGE-L (longest common subsequence), and ROUGE-S (skip-bigram co-occurrence statistics)

- **METEOR (Metric for Evaluation of Translation with Explicit Ordering)**: Similar to BLEU, but considering synonym matching and stemming. METEOR also accounts for the order of words and provides a penalty for serious ordering errors.