
Contrastive Learning for Supervised Graph Matching (Supplementary Material)

Gathika Ratnayaka¹

Qing Wang¹

Yang Li²

¹School of Computing, Australian National University, Canberra, Australia

²School of Information Technology, Deakin University, Melbourne, Australia

A PROOFS

Lemma 1. *Given a Sinkhorn ranking \mathbf{S} , Algorithm 1 can always return a matching \mathbf{M} that satisfies the one-to-one mapping constraint: $\sum_{j=1}^n \mathbf{M}_{i,j} = 1 \forall i \in V_S$ and $\sum_{i=1}^m \mathbf{M}_{i,j} \leq 1 \forall j \in V_T$.*

Proof. We first prove that such a matching \mathbf{M} satisfies $\sum_{j=1}^n \mathbf{M}_{i,j} = 1 \forall i \in V_S$ in two steps:

- (1) We show that each node in V_S can be matched to at most one node in V_T . In Algorithm 1, if a node $i \in V_S$ is already matched with some node in V_S , the node i is excluded for any further consideration by the condition on line 3. Further, even if a node $i \in V_S$ is not matched, the algorithm only matches the node i with at most one node in V_T during Lines 4-11. Thus, each node $i \in V_S$ can never be matched to two or more nodes in V_T .
- (2) Now we prove that each node in V_S must be matched to at least one node in V_T by contradiction. We assume that there exists a node $i \in V_S$ which is not matched with any node in V_T . Then by Algorithm 1, this means that, each time when the highest ranked node (say $j \in V_T$) is selected from i 's preference order, this the highest ranked node j does not meet the conditions specified in Line 5 and Line 9. Thus, such a node j is removed from the preference order of i (Line 12). As the node i is not matched, the algorithm must have attempted to match the node i with every node in V_T and each node in V_T fails to meet the conditions specified in Line 5 and Line 9. However, we know that $|V_T| \geq |V_S|$ and each node in V_S can be matched to at most one node in V_T (see the proof in the first step). Thus, this is impossible - contradicting with the assumption.

We then prove that such a matching \mathbf{M} satisfies $\sum_{i=1}^m \mathbf{M}_{i,j} \leq 1 \forall j \in V_T$ by showing that any node $j \in V_T$ can be matched with at most one node in V_S :

- In Algorithm 1, a node $j \in V_T$ can only be matched to a node $i \in V_S$ if node j is still not matched yet (Line 5) or if node j ranks i with a higher preference order than its currently matched node $i' \in V_S$ (Line 9). In the first case, the node j is guaranteed to be only matched to i . In the second case, the match between i' and j is broken ($\mathbf{M}_{i',j} \leftarrow 0$) before matching i with j . These ensure that any node $j \in V_T$ can only be matched with at most one node in V_S .

Putting the above together, the proof is done. □

Theorem 1. *Let \mathbf{M} be a matching returned by Algorithm 1 over a Sinkhorn ranking \mathbf{S} . Then \mathbf{M} is stable w.r.t. \mathbf{S} .*

Proof. We prove it by contradiction. Let us assume that a matching \mathbf{M} returned by Algorithm 1 over a Sinkhorn ranking \mathbf{S} is unstable. Then, by Definition 2, there must exist at least one blocking pair (i, j) with $i \in V_S$ and $j \in V_T$ in \mathbf{M} such that the following conditions are all satisfied:

- (1) $\mathbf{M}_{i,j} = 0$;

- (2) $\exists k \in V_S \setminus \{i\}$ s.t. $\mathbf{M}_{k,j} = 1$ and $\mathbf{S}_{i,j} > \mathbf{S}_{k,j}$;
- (3) $\exists k' \in V_T \setminus \{j\}$ s.t. either $\mathbf{M}_{i,k'} = 1$ and $\mathbf{S}_{i,j} > \mathbf{S}_{i,k'}$, or $\mathbf{M}_{i,k'} = 0$ for all $k' \in V_T \setminus \{j\}$.

By Condition (1), we know that node i is not matched with node j . By Condition (2), there exist another node $k \in V_S$ which is ranked lower than node i in j 's preference order but is matched with node j . As per Algorithm 1 (Lines 8-11), if $i \in V_S$ attempts to match with $j \in V_T$, the matching attempt would be successful as node i is ranked higher than node k in j 's preference order, unless node i is already matched with some node $k' \in V_T \setminus \{j\}$ that is ranked higher than j in i 's preference order. However, by Condition (3), this is not possible.

Therefore, \mathbf{M} does not have any blocking pair, contradicting with the assumption. The proof is done. \square

Theorem 2. *The contrastive loss of Eq. 8 is both positive-informative and negative-informative.*

Proof. We prove that the contrastive loss of Eq. 8 has Property 1 and Property 2 in turn.

- *Property 1:* Given two nodes $\{i, i'\} \subseteq V_S$ with $f^-(i) = f^-(i')$ and $f^+(i) > f^+(i')$, we know that $\frac{f^+(i)}{1+f^-(i)} > \frac{f^+(i')}{1+f^-(i')}$. This leads to

$$-\ln\left(\frac{f^+(i')}{1+f^-(i')}\right) < -\ln\left(\frac{f^+(i)}{1+f^-(i)}\right). \quad (1)$$

Hence, $L_i < L_{i'}$ holds. Property 1 is proved.

- *Property 2:* Given two nodes $\{i, i'\} \subseteq V_S$ with $f^+(i) = f^+(i')$ and $f^-(i) > f^-(i')$, we have $\frac{f^+(i)}{1+f^-(i)} < \frac{f^+(i')}{1+f^-(i')}$, and hence

$$-\ln\left(\frac{f^+(i)}{1+f^-(i)}\right) > -\ln\left(\frac{f^+(i')}{1+f^-(i')}\right) \quad (2)$$

This proves Property 2, i.e., $L_i > L_{i'}$. \square

Theorem 3. *Let \mathbf{M} be a matching produced by StableGM. Then \mathbf{M} is dual-optimal if \mathbf{S} is R1-symmetric.*

Proof. If \mathbf{S} is R1-symmetric, by the first condition of Definition 5, we know that for any given node $i \in V_S$, there must exist one node $j \in V_T$ such that $\mathbf{S}_{i,j} > \mathbf{S}_{i,k}$ holds for all other nodes $k \in V_T \setminus \{j\}$. This implies that node j is the highest ranked node in i 's preference order (Line 2 in Algorithm 1). Then, according to Algorithm 1 (Lines 3-11), when node i attempts to be matched (Line 3), there are two cases:

- if node j is not matched yet, then node i is matched with node j (Line 6 in Algorithm 1);
- if node j is already matched with another node $k \in V_S \setminus \{i\}$, then by the second condition of Definition 5 we know that $\mathbf{S}_{i,j} > \mathbf{S}_{i,k}$. This implies that node i is the highest ranked node in the preference order of j . Hence, node i is also matched with node j (Line 11 in Algorithm 1).

By Definition 3, we know that a matching \mathbf{M} is \mathcal{G}_S -optimal w.r.t. \mathbf{S} if $\mathbf{M}_{i,j} = 1$ implies $\mathbf{S}_{i,j} > \mathbf{S}_{i,k}$ for all $k \in V_T \setminus \{j\}$. From the above two cases, we can see that, when Algorithm 1 matches node i with node j (Lines 6 and 11), $\mathbf{S}_{i,j} > \mathbf{S}_{i,k}$ holds for all $k \in V_T \setminus \{j\}$. Thus \mathbf{M} produced by Algorithm 1 is \mathcal{G}_S -optimal w.r.t. \mathbf{S} .

Similarly, Definition 3, we know that a matching \mathbf{M} is \mathcal{G}_T -optimal w.r.t. \mathbf{S} if $\mathbf{M}_{i,j} = 1$ implies $\mathbf{S}_{i,j} > \mathbf{S}_{k,j}$ for all $k \in V_S \setminus \{i\}$. Since Algorithm 1 matches node i with node j (Lines 6 and 11), $\mathbf{S}_{i,j} > \mathbf{S}_{i,k}$ holds for all $k \in V_T \setminus \{j\}$. By the second condition of Definition 5, we can thus have $\mathbf{S}_{i,j} > \mathbf{S}_{k,j}$ holds for all $k \in V_S \setminus \{i\}$. Hence, \mathbf{M} produced by Algorithm 1 is also \mathcal{G}_T -optimal w.r.t. \mathbf{S} .

Since \mathbf{M} is both \mathcal{G}_S -optimal and \mathcal{G}_T -optimal w.r.t. \mathbf{S} , by Definition 4, \mathbf{M} is dual optimal w.r.t. \mathbf{S} . \square

Proposition 1. Let \mathbf{M} be a matching produced by StableGM. Then \mathbf{M} is dual-optimal when the contrastive loss function $L(\cdot)$ is minimized.

Proof. By Eq. 8, we know that $L(\cdot)$ is minimized when the contrastive losses of all nodes in V_S are minimized, and for a given $i \in V_S$, its corresponding contrastive loss L_i is minimized when $L_i = 0$. Further, according to Eq. 9, $L_i = 0$ implies that $f^+(i) = 1$ and $f^-(i) = 0$. Since $f^+(i) = \mathbf{S}_{i,\pi(i)}^2$, we thus know $\mathbf{S}_{i,\pi(i)} = 1$ when $L_i = 0$.

As $\mathbf{S} \in \mathbb{S}$ is a rectangular doubly stochastic matrix satisfying the conditions: $\sum_{k=1}^n \mathbf{S}_{i,k} = 1$ and $\sum_{k=1}^m \mathbf{S}_{k,\pi(i)} \leq 1$, $\mathbf{S}_{i,\pi(i)} = 1$ implies $\mathbf{S}_{i,k} = 0$ for all nodes $k \in V_T \setminus \{\pi(i)\}$ and $\mathbf{S}_{k,\pi(i)} = 0$ for all nodes $k \in V_S \setminus \{i\}$. Hence, from the above, we can see that, when L_i is minimized, the following properties are preserved for $i \in V_S$:

- $\mathbf{S}_{i,\pi(i)} > \mathbf{S}_{i,k}$ holds for all $k \in V_T \setminus \{\pi(i)\}$;
- $\mathbf{S}_{i,\pi(i)} > \mathbf{S}_{k,\pi(i)}$ holds for all $k \in V_S \setminus \{i\}$.

By Definition 5, we know that \mathbf{S} is R1-symmetric in this case because the following conditions hold for $j = \pi(i)$:

- for any $i \in V_S$ there exists some $j \in V_T$ such that $\mathbf{S}_{i,j} > \mathbf{S}_{i,k}$ holds for all $k \in V_T \setminus \{j\}$;
- if $\mathbf{S}_{i,j} > \mathbf{S}_{i,k}$ holds for all $k \in V_T \setminus \{j\}$, then $\mathbf{S}_{i,j} > \mathbf{S}_{k,j}$ also holds for all $k \in V_S \setminus \{i\}$.

By Theorem 3, we know that \mathbf{M} is dual optimal if \mathbf{S} is R1-symmetric. Therefore, \mathbf{M} is also dual optimal when the contrastive loss function $L(\cdot)$ is minimized. □

B ADDITIONAL EXPERIMENTS

B.1 EFFECT OF HARDNESS ATTENTION

β	Pascal VOC		Spair 71K		Willow Object Class	
	StableGM - v1	StableGM - v2	StableGM-v1	StableGM-v2	StableGM-v1	StableGM-v2
0.0	80.88	81.52	82.81	83.14	98.31	98.52
0.1	80.71	81.29	83.05	83.48	98.46	98.37
0.5	80.51	81.03	83.04	83.19	98.27	98.31
1.0	80.18	80.92	82.65	83.15	98.19	98.00

Table 1: Comparison of matching accuracy (%) by StableGM-v1 and StableGM-v2 under different values of $\beta \in \{0, 0.1, 0.5, 1\}$.

When $\beta = 1$, all the negative samples are considered as hard negatives, which is equivalent to a setting without any explicit hard negative sampling. From Table 1, it can be seen that when $\beta = 1.0$, the performance is lower when compared to the settings where an explicit hard negative sampling is performed, demonstrating the importance of the proposed "Hardness Attention" strategy. We can also see that the best performance for Pascal VoC keypoint dataset and the Willow object class dataset is produced when $\beta = 0$ for StableGM-v2. When, $\beta = 0$, only the negative samples with normalized affinities that are greater than that of related positive samples are considered as hard negatives. For the Spair-71K dataset, the best results are produced when $\beta = 0$, demonstrating the importance of the hardness parameter β .

Hyperparameters: Adam optimizer [Kingma and Ba, 2014] was used for training with an initial learning rate of 1×10^{-5} for VGG₁₆ and 5×10^{-4} for other models. 0.005 was used as the Sinkhorn regularization parameter α .

B.2 COMPARISON OF LOSS FUNCTIONS

We conducted experiments on Spair 71k dataset to observe the impact of different loss functions on our StableGM framework. In the experiments, StableGM-v2 model was modified by replacing the contrastive matching loss with permutation loss (elementwise cross entropy loss) [4] and InfoNCE loss, respectively. The results are given in Table 2. In the table, τ denotes the temperature parameter of InfoNCE loss.

	Permutation Loss	InfoNCE Loss ($\tau=1$)	InfoNCE Loss ($\tau=0.1$)	Contrastive Matching Loss (ours)
StableGM-v2	83.24	81.14	82.35	83.48

Table 2: Comparison of matching accuracy (%) for different loss functions

C RUNTIME ANALYSIS

We conduct experiments to compare the runtime needed for graph matching by different methods. Table 3 presents the average runtime taken to process a pair of graphs (i.e., a pair of keypoint annotated images) in the training and test phases for StableGM-v1 (ours), StableGM-v2 (ours), GCAN [Jiang et al., 2022], NGM-v2 [Wang et al., 2021] and NHGM-v2 [Wang et al., 2021].

Phase	StableGM-v1	StableGM-v2	GCAN	NGM-v2	NHGM-v2
Training (milliseconds)	45.81	48.45	55.01	56.37	322.59
Testing (milliseconds)	20.3	21.80	24.75	25.25	171.23

Table 3: Comparison of average runtime for processing a pair of keypoint annotated images (milliseconds per pair) by different methods on Pascal VoC keypoint dataset.

We can see that StableGM-v1 performs best in both training and testing phases. Further, we have the following observations:

- The time taken by StableGM-v2 is slightly higher than that of StableGM-v1. This is because StableGM-v2 has additional positional encoding and self-attention layers when compared to StableGM-v1.
- Although StableGM-v2 and GCAN implementations employ the same method to derive normalized cross-graph node-to-node affinities, StableGM-v2 performs faster than GCAN. The superior performance of StableGM-v2 against GCAN demonstrates the efficiency of our Stable Matching algorithm over the branch and bound algorithm used by GCAN.
- NGM-v2 and NHGM-v2 first convert two graphs into association graphs and hypergraphs, which results in additional computational overheads. Moreover, NGM-v2 and NHGM-v2 use the Hungarian algorithm to obtain final node correspondence between two graphs, which is slower than our Stable Matching algorithm as discussed in Section B of the main paper.

When evaluating the runtime, the experiments were carried out on a Linux server (Intel Xeon W-2175 with 2.50GHz, 28 cores, NVIDIA RTX A6000 GPU, 512GB of main memory)

D GRADIENT ANALYSIS OF THE STANDARD CONTRASTIVE LOSS FUNCTION

If the standard contrastive loss is adapted to our setting instead of the contrastive matching loss, L_i can be written as,

$$L_i = -\ln\left(\frac{\exp(s_{i,\pi(i)}/\tau)}{\sum_{k=1}^m \sum_{j=1}^n \mathbf{Z}_{k,j}^i \exp(s_{k,j}/\tau) + \exp(s_{i,\pi(i)}/\tau)}\right) \quad (3)$$

Then the gradient w.r.t. the affinity of a negative sample (i, q) will be,

$$\frac{\partial L_i}{\partial \mathbf{S}_{i,q}} = \frac{1}{\tau} P_{i,q}$$

where

$$P_{i,q} = \frac{\mathbf{Z}_{i,q}^i \exp(S_{i,q}/\tau)}{\sum_{k=1}^m \sum_{j=1}^n \mathbf{Z}_{k,j}^i \exp(s_{k,j}/\tau) + \exp(s_{i,\pi(i)}/\tau)}.$$

The gradient w.r.t. a positive sample will be,

$$\frac{\partial L_i}{\partial \mathbf{S}_{i,\pi(i)}} = -\frac{1}{\tau} \sum_{k=1}^m \sum_{j=1}^n P_{k,j}$$

Here τ indicates the temperature parameter used in the standard contrastive loss.

As the magnitude of gradient with respect to a positive sample is equal to the sum of the gradients of all the negative samples that are considered, the gradients propagated from the standard contrastive loss will ignore the affinity between a node pair in a positive sample. Moreover, in an explicit hard negative sampling strategy like ours, the propagated gradient can become 0 even if the affinity between a node pair in a positive sample can be further improved. These are undesirable properties that occur when applying the standard contrastive loss directly to supervised graph matching.

References

- Zheheng Jiang, Hossein Rahmani, Plamen Angelov, Sue Black, and Bryan M Williams. Graph-context attention networks for size-varied deep graph matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2343–2352, 2022.
- Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- Runzhong Wang, Junchi Yan, and Xiaokang Yang. Neural graph matching network: Learning lawler’s quadratic assignment problem with extension to hypergraph and multiple-graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.