Mutual CRF-GNN for Few-shot Learning

Shixiang Tang[1†] Dapeng Chen[2] Lei Bai[1] Kaijian Liu[2] Yixiao Ge[3] Wanli Ouyang[1]
[1]The University of Sydney, SenseTime Computer Vision Group, Australia
[2]Sensetime Group Limited, Hong Kong [3]The Chinese University of Hong Kong, Hong Kong
{stan3903, lei.bai, wanli.ouyang}@sydney.edu.au
liukaijian@sensetime.com dapengchenxjtu@yahoo.com yxge@link.cuhk.edu.hk

1. Notation Clarification

To distinguish from the main text, we use S-Fig, S-Tab, and S-Eq to name figures, tables, and equations presented in the supplementary material, correspondingly.

2. Pseudocode of Mutual CRF-GNN

Algorithm 1 and Algorithm 2 summarize the training and evaluation protocol of Mutual CRF-GNN (MCGN), respectively.

Algorithm 1 Training procedure of Mutual CRF-GNN in one episode

Require: a support set
$\mathcal{S} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_{N\times K}, y_{N\times K})\}$;
Require: a query set
$\mathcal{Q} = \{(\mathbf{x}_{N\times K+1}, y_{N\times K+1}), ..., (\mathbf{x}_{N\times K+T}, y_{N\times K+T})\}$;
Require: the backbone of feature extractor f_{emb};
Require: feature transformer σ in GNN;
Initialize features nodes $\mathbf{F}^1$ and affinities $\mathbf{A}^0$ by $\mathbf{F}^1 = f_{emb}(\mathbf{x}_i)$ and Eq. (8);
for l in [1, L] **do**
Compute unary compatibility Ψ_l and binary compatibility Φ_l in $\mathcal{G}_l^{crf}$ by Eq. (3) and Eq. (4), respectively;
Compute marginal distribution $\mathbf{P}(u_i^l)$ in $\mathcal{G}^{crf}$ by Eq. (5);
Compute $\mathbf{A}_l$ by Eq. (6);
Aggregate features $\mathbf{F}^{l+1}$ by Eq. (1);
end for
Compute $\mathcal{L}^{crf}$ by Eq. (9), $\mathcal{L}^{gnn}$ Eq. (10) and $\mathcal{L} = \lambda_{crf}\mathcal{L}^{crf} + \lambda_{gnn}\mathcal{L}^{gnn}$;
Update f_{emb} and σ by backward propagation;

3. Flowcharts of Baseline, GNN-only, CRF-only, CRF+GNN and MCGN

In the ablation study (Sec. 4.4), we introduce five variants of Mutual CRF-GNN Network, namely Baseline, GNN-only, CRF-only, CRF+GNN, and MCGN. The flowcharts of these variants are presented in S-Fig. 1. Specifically, **Baseline** is the MatchingNet [6] where similarities between support samples and query samples are directly calculated from feature embeddings. **GNN-only** is the GNN embedding model which can aggregate features and affinities but the affinity for GNN is defined by the embeddings of two connected nodes. **CRF-only** is the model where a single CRF directly follows the backbone. **CRF+GNN** is the model with two branches. One is the GNN branch which is the same as GNN-only and the other is the CRF branch which is the same as CRF-only. In this setting, CRF and GNN can not mutually contribute to each other. **MCGN** is the proposed method where CRF inference is leveraged to infer the affinity in GNN.

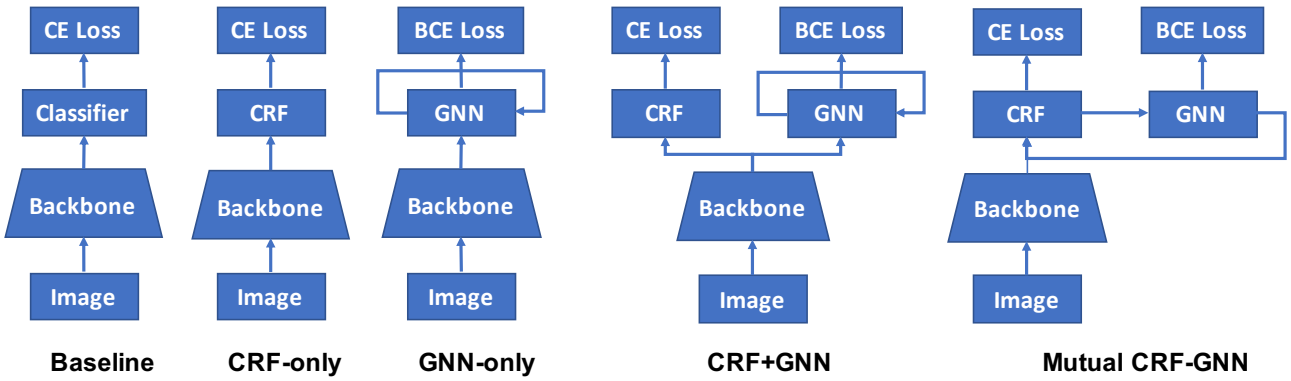
Algorithm 2 Evaluation procedure of Mutual CRF-GNN in one episode

Require: a support set
$\mathcal{S} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_{N\times K}, y_{N\times K})\}$;
Require: a query set
$\mathcal{Q} = \{(\mathbf{x}_{N\times K+1}, y_{N\times K+1}), ..., (\mathbf{x}_{N\times K+T}, y_{N\times K+T})\}$;
Require: the backbone of feature extractor f_{emb};
Require: feature transformer σ in GNN;
Initialize features nodes $\mathbf{F}^1$ and affinities $\mathbf{A}^0$ by $\mathbf{F}^1 = f_{emb}(\mathbf{x}_i)$ and Eq. (8);
for l in [1, L] **do**
Compute unary compatibility Ψ_l and binary compatibility Φ_l in $\mathcal{G}_l^{crf}$ by Eq. (3) and Eq. (4), respectively;
Compute marginal distribution $\mathbf{P}(u_i^l)$ in $\mathcal{G}^{crf}$ by Eq. (5);
Compute $\mathbf{A}_l$ by Eq. (6);
Aggregate features $\mathbf{F}^{l+1}$ by Eq. (1);
end for
Compute prediction of samples in query set by Eq. (11);

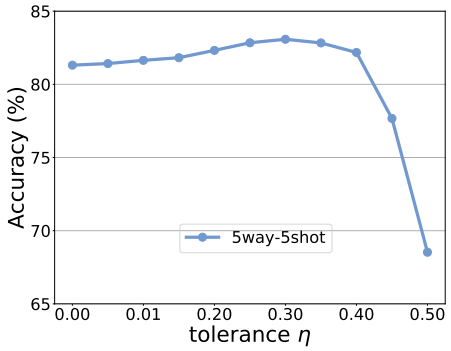
4. Sensitivity of Hyper-parameters

In this paper, we introduce several hyper-parameters, including the tolerance η when constructing unary compatibility ψ, the weight λ^{gnn}, λ^{crf} of loss $\mathcal{L}^{gnn}$ and $\mathcal{L}^{crf}$ in the final loss $\mathcal{L}$.

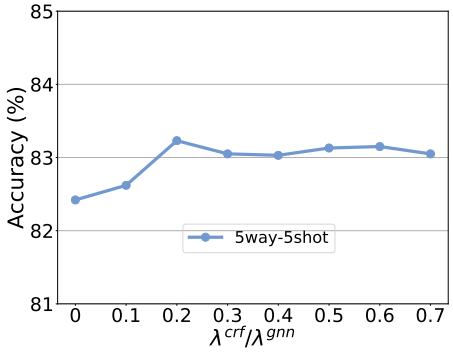
[†]This work was done when Shixiang Tang was an intern at SenseTime.



S-Figure 1: The flowchart of five variants, namely Baseline, CRF-only, GNN-only, CRF+GNN and MCGN in the ablation study.



S-Figure 2: Sensitivity of tolerance η. All experiments are tested on *mini*ImageNet in the 5-way 5-shot setting. The number of layers L is fixed to 5 and the maximum round number $R = 7$.



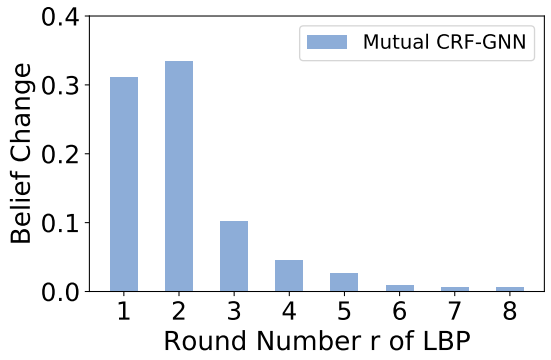
S-Figure 3: Sensitivity of loss weight ratio $\lambda^{crf}/\lambda^{gnn}$ when fixing $\lambda^{gnn} = 1$. All experiments are tested on *mini*ImageNet in the 5-way 5-shot setting. The number of layers L is fixed to 5 and the maximum round number R is fixed to 7.

4.1. Tolerance η

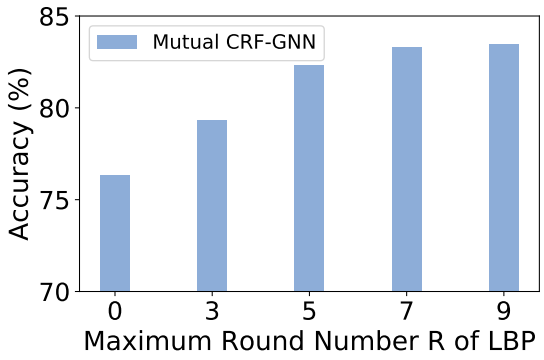
We tune the tolerance in *mini*ImageNet in 5-way 5-shot setting. We set the number of layers $L = 5$ and the maximum round number $R = 7$. The results are presented in S-Fig. 2 When tolerance η is near zero, the accuracy is relative low because it lacks flexibility that the variable has a tiny possibility of mislabelled even though they are observed. Specifically, if the tolerance η is set to 0, the marginal distribution of all random variables corresponding to support samples is deemed to be one-hot. In this scenario, the marginal distribution of random variables is dependent on affinities between query variables and support variables only. When the tolerance η is very high, the accuracy also decreases because there are too much noisy messages for the observations to be delivered to the random variable. We experimentally find that $\eta = 0.3$ is the optimal choice.

4.2. Loss weight ratio $\lambda^{crf}/\lambda^{gnn}$

The loss weight λ^{gnn} and λ^{crf} denotes the importance of $\mathcal{L}^{gnn}$ and $\mathcal{L}^{crf}$, respectively. Following the typical implementations in EGNN [3] and DPGN [7], we set μ_l^{crf} and μ_l^{gnn} to be 0.2 when $l < L+1$ and 1 when $l = L+1$. We fix λ^{gnn} to be 1 and then tuned λ^{crf} from 0 to 0.7. As shown in S-Fig. 3, the performance of MCGN is low when removing $\mathcal{L}^{crf}$ from the loss function (i.e., set $\lambda^{crf} = 0$), which shows the importance of jointly optimizing pairwise relations by $\mathcal{L}^{gnn}$ and class-level relations by $\mathcal{L}^{crf}$. Besides, the accuracy is not sensitive to $\lambda^{crf}/\lambda^{gnn}$ when $\lambda^{crf}/\lambda^{gnn} > 0.1$. Please note that $\lambda^{crf} = 0$ is different from GNN-only in Tab. (3). In GNN-only, we only use GNN and no CRFs are involved. However, when $\lambda^{crf} = 0$, we still incorporate CRF in each GNN layer but do not supervise the marginal distribution of CRF. In this case, GNN can benefit from support labels by CRF although the



S-Figure 4: The impact of the different round number of LBP with the belief change. When the round number $r \geq 6$, the belief change approaches 0, which indicates the convergence of LBP. Please note the number of layer in Mutual CRF-GNN is 5.



S-Figure 5: The impact of maximum round number of belief propagation. We perform the experiments on *mini*ImageNet in 5-way 5-shot setting. When the maximum round number $R < 7$, the accuracy increases with larger maximum round number. Please note the number of layer in Mutual CRF-GNN is 5.

marginal distribution can not be directly supervised.

5. Marginal Distribution Inference by Loopy Belief Propagation

The marginal distribution $\mathbf{P}(u_i^l|\mathbf{F}^l, \mathcal{Y}_s)$ of variable u_i^l is obtained by marginalizing out all random variables other than u_i^l in CRF. Mathematically, i.e.,

$$\mathbf{P}(u_i^l|\mathbf{F}^l, \mathcal{Y}_s) \propto \sum_{\mathcal{V}_l^{crf} \setminus \{u_i^l\}} \mathbf{P}(u_1^l, u_2^l, \ldots, u_{NK+T}^l|\mathbf{F}^l, \mathcal{Y}_s), \quad (1)$$

where $\mathbf{P}(u_i^l = m|\mathbf{F}^l, \mathcal{Y}_s) = p_{i,m}^l$, in which $p_{i,m}^l$ represents the possibility of u_i^l assigned label m. Marginal distribution requires the summation of all possible configures and can give a better prediction for each variable. In this paper, we adopt the loopy belief propagation [8, 4] to calculate marginal distribution.

5.1. Loopy Belief Propagation

In the following, we briefly introduce Loopy Belief Propagation (LBP) for inferring the marginal distributions of random variables $\mathbf{P}(u_i^l|\mathbf{F}^l, \mathcal{Y}_s)$. LBP maintains a belief $\mathbf{b}'_{l,i}$ of random variable u_i to represent the marginal distribution $\mathbf{P}(u_i^l|\mathbf{F}^l, \mathcal{Y}_s)$. $\mathbf{b}'_{l,i} \in \mathbb{R}^{1\times N}$ is a column vector and its j-th element is the marginal probability of u_i^l taking value j. According to LBP [8], given a initial $(\mathbf{b}_{l,i})^0$, the belief $\mathbf{b}'_{l,i}$ is obtained by running the following update rules until convergence,

$$\mathbf{m}_{l,i\to j}^r = [\phi(u_i^l, u_j^l)((\mathbf{b}_{l,i})^{r-1} \oslash \mathbf{m}_{l,j\to i}^{r-1})], \quad (2)$$

$$(\mathbf{b}_{l,j})^r \propto \begin{cases} \psi(u_j^l) \prod_{i\in\mathcal{N}_j} \mathbf{m}_{l,i\to j}^r & \text{if } j \leq N\times K, \\ \prod_{i\in\mathcal{N}_j} \mathbf{m}_{l,i\to j}^r & \text{if } j > N\times K. \end{cases} \quad (3)$$

where r denotes the round index of belief propagation and $r \in [0, R]$ with R as the maximum round number, $\mathbf{m}_{l,i\to j}^r$ is the message from u_i^l to u_j^l, ϕ_{ij}^l is the compatibility between variables for the l-th layer obtained using Eq. (4), $[\cdot]$ represents a normalization function that divides a vector by the sum of its elements, $\oslash$ represents the element-wise division between two vectors, $\mathcal{N}_j$ represents the neighbors of node j and the product of messages $\prod_{i\in\mathcal{N}_j} \mathbf{m}_{l,i\to j}^r$ means element-wise multiplication. We do not have unary compatibility ψ_j^l when $j > N \times K$ because query samples have no observations (labels). We get $\mathbf{b}'_{l,i} = (\mathbf{b}_{l,i})^R$, where R is the index of last iteration before LBP stops.

5.2. Convergence of LBP for Inference

Loopy belief propagation is a standard method of marginal distribution inference. It converges in most cases but cannot be theoretically confirmed, which leads to divergence in some cases [2, 1]. To explore the convergence empirically, we define belief change in MCGN as

$$\Delta_r = \frac{1}{L} \sum_{l=1}^{L} \sum_{i=1}^{N\times K+T} ||(\mathbf{b}_{l,i})^{r+1} - (\mathbf{b}_{l,i})^r||, \quad (4)$$

where L is number of layers in Mutual CRF-GNN. We report Δ_r with the r-th round of belief propagation in S-Fig. 4. Δ_r is large when the round index r is less than 3 and will decay from $r = 3$ to $r = 5$. Finally, it will converge to 0 when the round index is larger than 5. The diminish of Δ_r illustrates the convergence of belief $\{(\mathbf{b}_{l,i})^r\}_{i=1}^{N\times K+T}$ when $r \geq 6$.

5.3. Maximum Round Number R in LBP

To illustrate the improvement by the maximum round number R of belief propagation, we initialize the belief ($\mathbf{b}_{l,i}^0$ in S-Eq. 3) by the cosine similarity of the features $\mathbf{F}^l = \{\mathbf{f}_i^l\}_{i=1}^{N\times K+T}$ and the prototypes $\{\mathbf{c}_i^l\}_{i=1}^N$, *i.e.*, $(\mathbf{b}_{l,j})^0 =$

$(\mathbf{c}_1^l{}^\top \mathbf{f}_j^l, ..., \mathbf{c}_N^l{}^\top \mathbf{f}_j^l)$, where $\mathbf{c}_i^l = \frac{1}{K} \sum_{y_m=i} \mathbf{f}_m^l$ and K is the number of shots. In this paper, we treat the maximum round number as a hyperparameter and perform experiments on it. The results are reported in S-Fig. 5. The testing accuracy raises from 76.34% to 83.03% when R increases and it comes to converge if $R \geq 7$. The convergence can be explained by the convergence of LBP. As illustrated in S-Fig. 4, when $r \geq 6$, the belief change approaches 0, which means LBP converges when we set the maximum round index $R \geq 7$.

5.4. Time Complexity of LBP

The computational cost of LBP is $|E|S^2r$ [5], where $|E|$ is the number of edges in CRF, S is number of possible variable states, R is the maximum round number of LBP. Here, $|E| \approx (N \times K + T)^2$, $S = N$, where N is the number of classes and K is the number of shots. So the time complexity is $\mathrm{O}\left((N \times K + T)^2 N^2 R\right)$.

References

[1] Alexander T Ihler, John W Fisher, and Alan S Willsky. Message errors in belief propagation. In *Advances in Neural Information Processing Systems*, pages 609–616, 2005. 3

[2] Alexander T Ihler, Alan S Willsky, et al. Loopy belief propagation: Convergence and effects of message errors. *Journal of Machine Learning Research*, 6(May):905–936, 2005. 3

[3] Jongmin Kim, Taesup Kim, Sungwoong Kim, and Chang D Yoo. Edge-labeling graph neural network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11–20, 2019. 2

[4] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009. 3

[5] Koichi Ogawara. Approximate belief propagation by hierarchical averaging of outgoing messages. In *2010 20th International Conference on Pattern Recognition*, pages 1368–1372. IEEE, 2010. 4

[6] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638, 2016. 1

[7] Ling Yang, Liangliang Li, Zilun Zhang, Xinyu Zhou, Erjin Zhou, and Yu Liu. Dpgn: Distribution propagation graph network for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[8] Jaemin Yoo, Hyunsik Jeon, and U Kang. Belief propagation network for hard inductive semi-supervised learning. In *International Joint Conference on Artificial Intelligence*, pages 4178–4184, 2019. 3