



# Notes on Geometric Learning Papers

Matthew Mo

2021 年 4 月 24 日

## 目录

<b>1 NeVAE</b>	<b>11</b>
1.1 Encoder . . . . .	11
1.2 Decoder . . . . .	12
1.3 Training . . . . .	13
1.4 Property Oriented Mol. Gen. . . . .	13
<b>2 Seminar on Self/Un-Supervised Learning @ 2020/9/16</b>	<b>14</b>
2.1 Self-Learning @ Video Learning . . . . .	14
2.2 Transformation Equivariance vs. Invariance @ Visial Repr. Learning . . . . .	16
<b>3 AET, AVT: Autoencoding Transformations</b>	<b>18</b>
<b>4 Flow-Based Generative Models</b>	<b>20</b>
4.1 Outline & Basics . . . . .	20
4.2 MoFlow . . . . .	21
4.2.1 GCF/Graph Conditional Flow . . . . .	23
4.2.2 Validity Correction & Misc . . . . .	24
4.3 GraphNVP . . . . .	25
<b>5 WGAN</b>	<b>25</b>
<b>6 GMMN+AE</b>	<b>26</b>
6.1 Structure & Idea . . . . .	26
6.2 Training . . . . .	27
<b>7 FoldingNet - An AntoEncoder</b>	<b>27</b>

---

<b>8 PointFlow: Flow-based Generative Model on Point Clouds</b>	<b>28</b>
8.1 Continuous Normalizing Flow(CNF) . . . . .	28
8.2 Variational Auto-Encoder . . . . .	28
8.3 Model . . . . .	29
<b>9 FFJORD</b>	<b>31</b>
9.1 CNF . . . . .	31
9.2 Backpropagation through ODE Solutions with Adjoint Method . . . . .	31
9.3 Unbiased Linear-Time Log-Density Estimation . . . . .	31
<b>10 Dequantization to Learn Discrete Distribution</b>	<b>32</b>
10.1 Dequantization as Latent Variable Model . . . . .	32
10.2 Variational Dequantization . . . . .	33
10.3 Importance-Weighted Dequantization . . . . .	33
10.4 Renyi Dequantization . . . . .	34
10.5 Dequantization Distribution . . . . .	34
10.6 (Choice of) Continuous Distribution . . . . .	35
<b>11 DGI: Deep Graph Infomax</b>	<b>35</b>
11.1 Backgrounds, Approach, Math . . . . .	35
11.2 Algorithm . . . . .	36
<b>12 GraphSAGE: Inductive Representation Learning on Graph</b>	<b>37</b>
12.1 Embedding Generation/FP . . . . .	37
12.2 Aggragator Selection . . . . .	37
<b>13 SGC: Simplified Graph Convolution</b>	<b>38</b>
<b>14 FastGCN</b>	<b>38</b>
14.1 Method . . . . .	38
14.2 Variance Reduction . . . . .	39
<b>15 GWNN: Wavelet Transform on Graph</b>	<b>40</b>
15.1 Supplementary Math: Real and Complex Wavelets . . . . .	40
15.2 Graph Wavelets . . . . .	41
15.3 GWNN . . . . .	42
15.3.1 Details . . . . .	42
<b>16 Graph Wavelets</b>	<b>42</b>
16.1 经典小波变换/CWT . . . . .	42
16.2 谱小波变换/SGWT . . . . .	43

---

16.2.1	Scaling Functions . . . . .	44
16.3	SGBT 的性质 . . . . .	44
16.3.1	Inverse SGBT . . . . .	44
16.3.2	局域性 . . . . .	44
16.3.3	Spectral Wavelet Frames . . . . .	45
16.4	Fast SGBT Approximation by Polynomials . . . . .	45
16.4.1	Fast Approximation of Adjoint . . . . .	46
16.4.2	Inverse Calculation . . . . .	47
16.5	Implementations and Details . . . . .	47
<b>17</b>	<b>GMNN: Graph Markov Neural Network</b>	<b>48</b>
17.1	Pseudolikelihood Variational EM . . . . .	48
17.2	Inference . . . . .	49
17.3	Learning . . . . .	50
17.4	Optimization . . . . .	50
<b>18</b>	<b>ClusterGCN: Fast Deep &amp; Large GCNs</b>	<b>51</b>
18.1	Vanilla ClusterGCN: Cluster For Batch . . . . .	51
18.2	Stochastic Multiple Partitions . . . . .	52
18.3	Analysis of Deeper Networks . . . . .	52
<b>19</b>	<b>GAT: Graph Attention Network</b>	<b>53</b>
<b>20</b>	<b>Note on Probabilistic Graphical Models</b>	<b>53</b>
20.1	Bayesian Networks . . . . .	53
20.2	Undirected Networks . . . . .	54
20.3	Local Probabilistic Models   i.e. Specific Models Corresponds to Last 2 Sections . . . . .	56
20.4	Temporal Models . . . . .	57
<b>21</b>	<b>RSCNN(CVPR 19')</b>	<b>57</b>
21.1	Architecture . . . . .	57
21.2	Details & Implementation . . . . .	58
<b>22</b>	<b>SimpleView(ICLR 21' Candidate)</b>	<b>58</b>
22.1	Simple Review of Existing Protocols . . . . .	58
22.2	Model: SimpleView . . . . .	58
<b>23</b>	<b>OT-Flow</b>	<b>59</b>
23.1	Idea & Formulations . . . . .	59
23.2	Parametrization of Model . . . . .	59
23.3	Exact Hessian of Multilayer NN . . . . .	60

---

<b>24 Node2vec: Unsupervised Feature Learning</b>	<b>60</b>
24.1 Basics . . . . .	60
24.2 Biased Random Walk . . . . .	61
24.3 Edge Feature . . . . .	62
<b>25 DeepWalk: Online Representation Learning</b>	<b>62</b>
25.1 DeepWalk . . . . .	62
25.2 SkipGram . . . . .	63
25.3 Hierachichal Softmax . . . . .	63
25.4 Parallelization . . . . .	64
25.5 Variants . . . . .	64
<b>26 DAGNN: Towards Deeper GNN</b>	<b>64</b>
26.1 Smoothness Metrics . . . . .	65
26.2 Convergence of Propagation . . . . .	65
26.3 DAGNN: Deep Adaptive GNN . . . . .	65
<b>27 t-SNE(t-Distributed Stochastic Neighbor Embedding)</b>	<b>66</b>
27.1 SNE . . . . .	66
27.2 UNI-SNE . . . . .	66
27.3 t-SNE . . . . .	66
27.4 Barnes-Hut-SNE . . . . .	67
27.4.1 Approximating Input Similarities by Vantage-point Tree . . . . .	67
27.4.2 Approximating t-SNE Gradients . . . . .	67
<b>28 Autoregressive Flows</b>	<b>68</b>
28.1 Autoregressive Transformation . . . . .	68
28.2 MAF: Masked Autoregressive Flow . . . . .	69
28.3 IAF: Inverse Autoregressive Flow . . . . .	69
28.4 Sylvester NF(UAI 18') . . . . .	69
28.4.1 Idea . . . . .	69
28.4.2 Parametrization of A & B . . . . .	70
28.4.3 Preserving Orthogonality of Q . . . . .	70
<b>29 Contrastive Multi-view Representation Learning/MVRLG(ICML 20?)</b>	<b>71</b>
29.1 Augmentations . . . . .	71
29.2 Encoders . . . . .	71
29.3 Training . . . . .	71

---

<b>30 GCC: Graph Contrastive Coding for GNN Pre-Training(KDD 20')</b>	<b>73</b>
30.1 GCC Pre-Training . . . . .	73
30.2 Finetuning GCC . . . . .	74
<b>31 GIN: Graph Isomorphism Network(ICLR 19')</b>	<b>74</b>
31.1 Weisfeiler-Lehman Test . . . . .	74
31.2 Math Intuitions . . . . .	75
31.3 GIN . . . . .	75
31.4 Graph Readout of GIN . . . . .	75
<b>32 GraphLoG: Self-Supervised Representation Learning with Local &amp; Global Structure</b>	<b>76</b>
32.1 Preliminaries . . . . .	76
32.2 Local-Inst. Stru. Learning . . . . .	76
32.3 Global-Semantic Repr. Learning . . . . .	77
32.3.1 Init. of HP(Hierarchical Prototypes) . . . . .	77
32.3.2 Maintenance of HP . . . . .	77
32.4 Sup-GraphLoG: A Supervised Baseline . . . . .	78
<b>33 Orthogonal Weights in DNNs</b>	<b>78</b>
33.1 Formulation & Good Properties . . . . .	78
33.2 OWN: Orthogonal Weight Normalization . . . . .	78
33.2.1 Backpropagation . . . . .	79
33.2.2 As Convolution . . . . .	79
33.2.3 Group Based Orthogonalization: Divided Filters . . . . .	80
<b>34 OrthDNNs: Orthogonal DNNs(TPAMI 19')</b>	<b>80</b>
34.1 GE Analysis in a Robustness and Isomeric Mapping Perspective . . . . .	80
34.2 GE Analysis of DNN . . . . .	81
34.3 OrthDNN by SVB(Singular Value Bound) . . . . .	81
34.3.1 BN Compatibility . . . . .	82
34.3.2 On CNN: OrthDNN as Convolution . . . . .	82
<b>35 SimCLR: A Simple Framework for Contrastive Learning of Visual Representations</b>	<b>82</b>
35.1 Ideas & Basics . . . . .	82
<b>36 BYOL: Build Your Own Latent</b>	<b>83</b>
36.1 Ideas & Method . . . . .	83

---

<b>37 MoCo: Momentum Contrast for Unsupervised Visual Representation Learning</b>	<b>84</b>
37.1 Ideas & Method . . . . .	84
<b>38 SimSiam: Exploring Simple Siamese Representation Learning</b>	<b>85</b>
38.1 Intuitions . . . . .	85
38.2 Method . . . . .	85
<b>39 Graph Layouts by t-SNE</b>	<b>87</b>
39.1 Backgrounds . . . . .	87
39.2 Method . . . . .	87
39.3 tsNET . . . . .	87
<b>40 细粒度图像数据分类 by Xiangteng He</b>	<b>87</b>
40.1 细粒度图像数据分类 . . . . .	87
40.2 RL-based 图像部件/对象识别 . . . . .	87
40.3 多层注意力区域辨识 . . . . .	87
40.4 多模态 . . . . .	87
<b>41 Dirac Operator for Extrinsic Shape Analysis</b>	<b>88</b>
41.1 Math . . . . .	88
41.2 离散化 . . . . .	89
41.3 实值表示 . . . . .	89
41.4 有界区域: 边界条件 . . . . .	89
<b>42 Mesh-Based Simulation with GNNs</b>	<b>89</b>
42.1 结构 . . . . .	89
42.2 Adaptive Remeshing . . . . .	90
<b>43 SENet</b>	<b>91</b>
<b>44 Continuous-Time Spiking Neural Network</b>	<b>92</b>
44.1 Neurons . . . . .	92
44.2 Network Topology . . . . .	92
44.3 突触塑性规则 . . . . .	92
<b>45 Towards Deep Learning Models Resistant to Adversarial Attacks</b>	<b>93</b>
45.1 Inner Maximize Prob. : 如何提出好的对抗样本 . . . . .	93
45.2 最小化问题 . . . . .	93
45.3 网络能力 & 对抗健壮性 . . . . .	94
<b>46 CAS: Channel-wise Activation Suppressing Module for Adversarial Robustness</b>	<b>94</b>

---

<b>47 Resisting Adversarial Attacks by <math>k</math>-Winners-Takes-All</b>	<b>95</b>
47.1 Related Work: Obfuscated Gradients . . . . .	95
47.2 $k$ -Winners-Takes-All . . . . .	95
47.3 Training . . . . .	96
47.4 Theory Understand of the Discontinuity . . . . .	96
47.5 Related Works . . . . .	96
<b>48 Normalized Loss Functions for Deep Learning with Noisy Labels</b>	<b>97</b>
48.1 Robustness Alone is not Sufficient . . . . .	98
<b>49 A Survey of Label-noise Representation Learning(LNRL): Past, Present and Future</b>	<b>99</b>
49.1 Perspective of Data . . . . .	100
49.2 Perspective of Statistics/Learning Theory . . . . .	100
49.3 Perspective of Opt. Policy . . . . .	101
49.4 Taxonomy . . . . .	101
49.5 Data: Noise Trans. Mat. . . . .	101
49.6 Data: Adapt. Layer . . . . .	102
49.7 Loss Correction: Backward/Forward Correction . . . . .	102
49.8 Loss Correction: Gold Correction . . . . .	103
49.9 Loss Correction: Label Smoothing . . . . .	103
49.10 Prior Knowledge: Human-in-the-Loop Estimation . . . . .	103
49.11 Prior Knowledge: Fine-tuning Revision . . . . .	104
49.12 Regularization: Explicit Regularization . . . . .	104
49.13 Objective Regularization: Implicit . . . . .	105
49.14 Objective Reweighting: Importance Reweighting . . . . .	106
49.15 Objective Reweighting: Bayesian Methods . . . . .	106
49.16 Objective Reweighting: NNs . . . . .	107
49.17 Objective Redesigning . . . . .	107
49.18 Objective Redesigning: Loss Redesign . . . . .	107
49.19 Label Ensemble . . . . .	108
49.20 Optimization Policies . . . . .	109
49.20.1 Memorization Eff. . . . .	110
49.20.2 Self-training . . . . .	110
49.20.3 Co-training . . . . .	111
49.20.4 Beyond Memorization . . . . .	112
49.21 Future . . . . .	112
49.21.1 Build Up New Datasets . . . . .	112
49.21.2 Instance-Dependent LNRL . . . . .	112

---

49.21.3 Adversarial LNRL . . . . .	112
49.21.4 Beyond Labels: Noisy Data . . . . .	113

<b>50 Multimodal Research in Vision and Language: A Review of Current and Emerging Trends</b>	<b>114</b>
50.1 Tasks . . . . .	114
50.1.1 归纳任务 . . . . .	114
50.1.2 分类任务 . . . . .	115
50.1.3 Retrieval 任务 . . . . .	115
50.1.4 其他任务 . . . . .	115
50.2 任务相关 Trends in VisLang . . . . .	115
50.2.1 VC . . . . .	115
<b>51 Answering Questions about Data Visualizations using Efficient Bimodal Fusion</b>	<b>118</b>
51.1 Architecture . . . . .	118
51.1.1 Image Encoder . . . . .	119
51.1.2 Parallel Fusion of I+L . . . . .	119
51.1.3 Recurrent Aggr. of bi-modal features . . . . .	119
51.1.4 OCR Integration for DVQA . . . . .	119
<b>52 Seminar on CVPR2021</b>	<b>119</b>
52.1 Subspace learning, Self-Expressive Model . . . . .	120
52.2 Kaleido-BERT, Fashion Domain VisLang Pre-training . . . . .	120
<b>53 OSCAR: Object-Semantics Aligned Pre-training for Vision-Language Tasks</b>	<b>120</b>
53.1 OSCAR . . . . .	120
53.2 Adapting to V+L Tasks . . . . .	123
<b>54 In Defense of Grid Features for Visual Question Answering</b>	<b>123</b>
54.1 Related Work . . . . .	123
54.2 From Regions to Grids . . . . .	124
54.2.1 Bottom-Up Attention w/ Regions . . . . .	124
54.2.2 Grid Features from the Same Layer . . . . .	124
54.2.3 $1 \times 1$ RoIPool for Improved Grid F . . . . .	124
54.3 Comparison: Region v. Grids . . . . .	125
54.4 Why Work? . . . . .	125
54.5 Generalization . . . . .	125
<b>55 Beyond Bilinear: Generalized Multimodal Factorized High-order Pooling for Visual Question Answering</b>	<b>125</b>
55.1 Related Work: Multi-modal Bilinear Models for VQA . . . . .	125



---

55.2 Generalized Multi-modal Factorized High-order Pooling	126
55.3 Net Architecture	127
<b>56 UNITER UNiversal Image-TExT Representation Learning</b>	<b>129</b>
56.1 Pre-training Tasks	130
<b>57 Vusial Commonsense R-CNN</b>	<b>131</b>





世界がこんなに美しいはずだが、僕は何故か追いつかない。  
あなたと出会い日が来るよう。

*Dedicated to Amatsukaze and Elaina, Fairies From The Ideal World.*



# 1. NeVAE

Idea VAE on graph, node-wise repr, permutation invariant.

## 1.1 Encoder

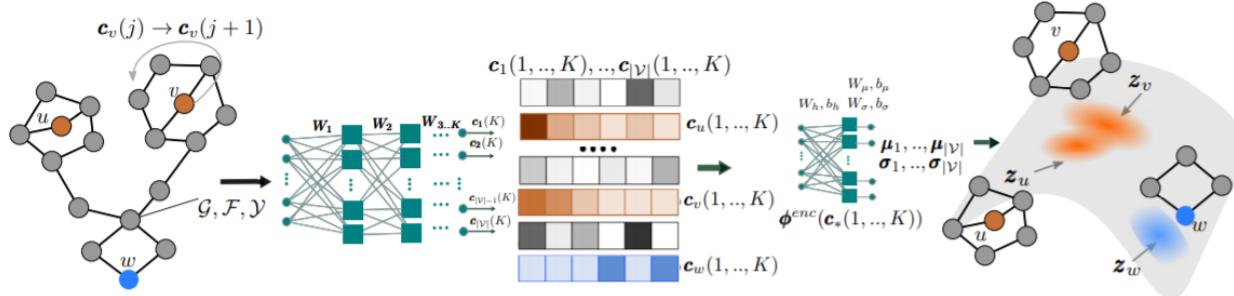


Figure 1: The encoder of our variational autoencoder for molecular graphs. From left to right, given a molecular graph  $\mathcal{G}$  with a set of node features  $\mathcal{F}$  and edge weights  $\mathcal{Y}$ , the encoder aggregates information from a different number of hops  $j \leq K$  away for each node  $v \in \mathcal{G}$  into an embedding vector  $\mathbf{c}_v(j)$ . These embeddings are fed into a differentiable function  $\phi^{enc}$  which parameterizes the posterior distribution  $q_\phi$ , from where the latent representation of each node in the input graph are sampled from.

GNN-like message-passing on k-hops:

$$q_\phi(\mathbf{z}_u | \mathcal{V}, \mathcal{E}, \mathcal{F}, \mathcal{Y}) \sim \mathcal{N}(\boldsymbol{\mu}_u, Diag(\boldsymbol{\sigma}_u)) \quad (1)$$

$$[\boldsymbol{\mu}_u, \boldsymbol{\sigma}_u] = \phi^{enc}(\mathbf{c}_u(k)_{k=1..K}) \quad (2)$$

$$\mathbf{c}_u(k) = \begin{cases} \mathbf{r}(\mathbf{W}_k^T \mathbf{t}_u + \mathbf{W}_k^X \mathbf{x}_u), & k = 1 \\ \mathbf{r}(\mathbf{W}_k^T \mathbf{t}_u + \mathbf{W}_k^X \mathbf{x}_u \odot \Lambda(\{y_{uv} \mathbf{c}_v(k-1)\}_v \in N(u))), & k > 1 \end{cases} \quad (3)$$

## 1.2 Decoder

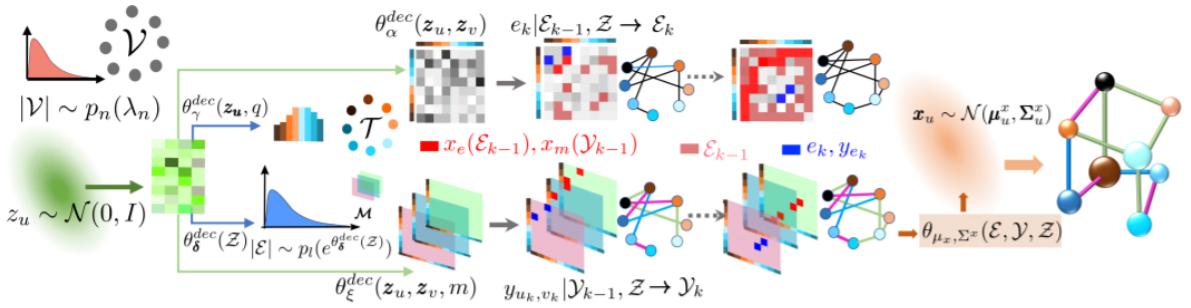


Figure 2: The decoder of our variational autoencoder for molecular graphs. From left to right, the decoder first samples the number of nodes  $n = |\mathcal{V}|$  from a Poisson distribution  $p_n(\lambda_n)$  and it samples a latent vector  $\mathbf{z}_u$  per node  $u \in \mathcal{V}$  from  $\mathcal{N}(\mathbf{0}, \mathbf{I})$ . Then, for each node  $u$ , it represents all potential node feature values as an unnormalized log probability vector (or ‘logits’), where each entry is given by a nonlinearity  $\theta_\gamma^{dec}$  of the corresponding latent representation  $\mathbf{z}_u$ , feeds this logit into a softmax distribution and samples the node features. Next, it feeds all latent vectors  $\mathcal{Z}$  into a nonlinear log intensity function  $\theta_\delta^{dec}(\mathcal{Z})$  which is used to sample the number of edges. Thereafter, on the top row, it constructs a logit for all potential edges  $(u, v)$ , where each entry is given by a nonlinearity  $\theta_\alpha^{dec}$  of the corresponding latent representations  $(\mathbf{z}_u, \mathbf{z}_v)$ . Then, it samples the edges one by one from a soft max distribution depending on the logit and a mask  $\beta_e(\mathcal{E}_{k-1})$ , which gets updated every time it samples a new edge  $e_k$ . On the bottom row, it constructs a logit per edge  $(u, v)$  for all potential edge weight values  $m$ , where each entry is given by a nonlinearity  $\theta_\xi^{dec}$  of the latent representations of the edge and edge weight value  $(\mathbf{z}_u, \mathbf{z}_v, m)$ . Then, every time it samples an edge, it samples the edge weight value from a soft max distribution depending on the corresponding logit and mask  $x_m(u, v)$ , which gets updated every time it samples a new  $y_{u_k v_k}$ . Finally, for each atom  $u$ , it samples its coordinates  $\mathbf{x}_u$  from a multidimensional Gaussian distribution whose mean  $\mu_x$  and variance  $\Sigma_x$  depends on the latent vectors of the corresponding atom and its neighbors and the underlying chemical bonds.

Decoder: gen. logits, softmax edges one-by-one, possible binary mask(for expert exp.).

$$\text{Nodes Count: } |\mathcal{V}| \sim p_l(\lambda_n) \quad (4)$$

$$\text{Latent Repr.: } \mathbf{z}_u \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad (5)$$

$$\text{Node Feat.: } \mathbf{f}_u = \text{softmax}_u(\theta_\gamma^{dec}(\mathbf{z}_u, q)), q \text{ is atom type} \quad (6)$$

$$\text{Edges Count: } |\mathcal{E}| \sim p_l(e^{\theta_\delta^{dec}(\mathcal{Z})}) \quad (7)$$

$$\text{Edges Gen.: } p(e = (u, v) | \mathcal{E}_{k-1}, \mathcal{V}) = \frac{\beta_e e^{\theta_\alpha^{dec}(\mathbf{z}_u, \mathbf{z}_v)}}{\sum_{e' = (u', v') \notin \mathcal{E}_{k-1}} \beta_{e'} e^{\theta_\alpha^{dec}(\mathbf{z}'_u, \mathbf{z}'_v)}} \quad (8)$$

$$\text{E. Feat. Gen.: } p(y_{uv} = m | \mathcal{Y}_{k-1}, \mathcal{V}) = \frac{\beta_m(u, v) e^{\theta_\xi^{dec}(\mathbf{z}_u, \mathbf{z}_v, m)}}{\sum_{m' \neq m} \beta'_{m'}(u, v) e^{\theta_\xi^{dec}(\mathbf{z}_u, \mathbf{z}_v, m')}}, \text{ note: not normal softmax?} \quad (9)$$

$$\text{Pos. Gen.: } p(\mathbf{x}_u | \mathcal{E}, \mathcal{Y}, \mathcal{Z}) = \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x) \quad (10)$$

$$[\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x] = [\theta_{\mu^x}(\mathbf{r}(u)), \theta_{\Sigma^x}(\mathbf{r}(u)) \theta_{\Sigma^x}^T(\mathbf{r}(u))] \quad (11)$$

$$\mathbf{r}(u) = \mathbf{z}_u + \sum_{v \in N(u)} y_{uv} \mathbf{z}_v \quad (12)$$

### 1.3 Training

- Prior:  $\mathcal{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
- maximize evidence lower bound(ELBO)+Poisson max-likelihood:

$$\max_{\phi, \theta, \lambda_n} \frac{1}{N} \sum_{i \in [N]} \mathbb{E}_{q_\phi(\mathcal{Z}_i | \mathcal{Y}_i, \mathcal{E}_i, \mathcal{F}_i, \mathcal{Y}_i)} [\log p_\theta(\mathcal{Y}_i, \mathcal{E}_i, \mathcal{F}_i, |\mathcal{Z}_i)] - KL(q_\phi || p_z) + \log p_{\lambda_n}(n_i) \quad (13)$$

- note term  $E_{q_\phi}[\log p_\theta(\mathcal{Y}_i, \mathcal{E}_i, \mathcal{F}_i, |\mathcal{Z}_i)]$  need the edges sequence specified, use BFS with random tie breaking in child-sel. step, with random selected source node  $s \sim \zeta_s$ ! Thus

$$E_{q_\phi}[\log p_\theta(\mathcal{Y}_i, \mathcal{E}_i, \mathcal{F}_i, |\mathcal{Z}_i)] \approx E_{q_\phi}[\log \mathbb{E}_{s \sim \zeta_s} p_\theta(\mathcal{Y}_i, \mathcal{E}_i, \mathcal{F}_i, |\mathcal{Z}_i)] \quad (14)$$

$$\geq E_{q_\phi, s \sim \zeta_s} [\log p_\theta(\mathcal{Y}_i, \mathcal{E}_i, \mathcal{F}_i, |\mathcal{Z}_i)] \quad (15)$$

- **Theorem** If dist.  $\zeta_s$  is independent to labels of nodes, then the learned model is permutation-invariant.
- **Proposition** Decoder defined is permutation-invariant.

### 1.4 Property Oriented Mol. Gen.

Train variational probabilistic decoder, to maxmize some property of mol.,  $\Rightarrow$ train a supervised decoder  $p^*$  on trained decoder  $p_\theta$ :

$$\min_{p(\cdot|\mathcal{Z})} \mathbb{E}_{\mathcal{Z} \sim p_z(\cdot)} \mathbb{E}_{\mathcal{E}, \mathcal{Y}, \mathcal{F} \sim p(\mathcal{E}, \mathcal{Y}, \mathcal{F} | \mathcal{Z})} [l(\mathcal{E}, \mathcal{Y}, \mathcal{F}) + \rho \log \frac{p(\mathcal{E}, \mathcal{Y}, \mathcal{F} | \mathcal{Z})}{p_\theta(\mathcal{E}, \mathcal{Y}, \mathcal{F} | \mathcal{Z})}] \quad (16)$$

$$\Rightarrow \min \mathbb{E}_{\mathcal{Z} \in p_z} [KL(p(\cdot|\mathcal{Z}) || g_\theta(\cdot|\mathcal{Z}))] \quad (17)$$

$$\text{where } g_\theta(\mathcal{E}, \mathcal{Y}, \mathcal{F} | \mathcal{Z}) = \frac{p_\theta(\mathcal{E}, \mathcal{Y}, \mathcal{F} | \mathcal{Z}) \exp\left(-\frac{l(\mathcal{E}, \mathcal{Y}, \mathcal{F})}{\rho}\right)}{\mathbb{E}_{\mathcal{E}, \mathcal{Y}, \mathcal{F} \sim p_\theta(\cdot|\mathcal{Z})} [p_\theta(\mathcal{E}, \mathcal{Y}, \mathcal{F} | \mathcal{Z}) \exp\left(-\frac{l(\mathcal{E}, \mathcal{Y}, \mathcal{F})}{\rho}\right)]} \quad (18)$$

The above equations has a obvious solution  $p^* \equiv g_\theta$ , however sampling might be too slow for practical use  $\Rightarrow$ A Stochastic Gradient Approach.

**Algorithm 1: PROPERTYORIENTEDDECODER:** it trains a parameterized property-oriented decoder.

- 1: **Given:** The loss function  $\ell(\cdot)$ , parameter  $\rho$ , original decoder  $p_\theta$ , # of iterations  $M$ , mini batch size  $B$ , and learning rate  $\gamma$
- 2:  $\theta'_0 \leftarrow \theta$
- 3: **for**  $j = 1, \dots, M$  **do**
- 4:    $\mathcal{Z}_j \sim p_z(\cdot)$
- 5:    $\mathcal{D} \leftarrow \text{MINIBATCH}(p_{\theta'_j}(\cdot|\mathcal{Z}_j), B)$
- 6:    $\nabla \leftarrow 0$
- 7:   **for**  $(\mathcal{E}_i, \mathcal{Y}_i, \mathcal{F}_i) \in \mathcal{D}$  **do**
- 8:      $S \leftarrow \ell(\mathcal{E}_i, \mathcal{Y}_i, \mathcal{F}_i) + \rho \log \left( p_{\theta'_j}(\mathcal{E}_i, \mathcal{Y}_i, \mathcal{F}_i | \mathcal{Z}_j) / p_\theta(\mathcal{E}_i, \mathcal{Y}_i, \mathcal{F}_i | \mathcal{Z}_j) \right)$
- 9:      $\nabla \leftarrow \nabla + (S + \rho) \nabla_{\theta'} \log p_{\theta'_j}(\mathcal{E}_i, \mathcal{Y}_i, \mathcal{F}_i | \mathcal{Z}_j)$
- 10:    $\theta'_{j+1} \leftarrow \theta'_j + \gamma \frac{\nabla}{B}$
- 11: **Return**  $\theta'_M$

Use SGD to update param.  $\theta'$  of  $p_{\theta'}$ :

$$\Delta\theta' = \alpha \nabla_{\theta'} \mathbb{E}_{\mathcal{Z} \sim p_z(\cdot)} \mathbb{E}_{\mathcal{E}, \mathcal{Y}, \mathcal{F} \sim p(\mathcal{E}, \mathcal{Y}, \mathcal{F} | \mathcal{Z})} [l(\mathcal{E}, \mathcal{Y}, \mathcal{F}) + \rho \log \frac{p(\mathcal{E}, \mathcal{Y}, \mathcal{F} | \mathcal{Z})}{p_{\theta}(\mathcal{E}, \mathcal{Y}, \mathcal{F} | \mathcal{Z})}] \quad (19)$$

$$= \alpha \mathbb{E}_{\mathcal{Z} \sim p_z(\cdot)} \nabla_{\theta'} \mathbb{E}_{\mathcal{E}, \mathcal{Y}, \mathcal{F} \sim p(\mathcal{E}, \mathcal{Y}, \mathcal{F} | \mathcal{Z})} [l(\mathcal{E}, \mathcal{Y}, \mathcal{F}) + \rho \log \frac{p(\mathcal{E}, \mathcal{Y}, \mathcal{F} | \mathcal{Z})}{p_{\theta}(\mathcal{E}, \mathcal{Y}, \mathcal{F} | \mathcal{Z})}] \quad (20)$$

$$(\text{by log-deriv. trick}) = \alpha \mathbb{E}_{\mathcal{Z} \sim p_z(\cdot)} \mathbb{E}_{\mathcal{E}, \mathcal{Y}, \mathcal{F} \sim p(\mathcal{E}, \mathcal{Y}, \mathcal{F} | \mathcal{Z})} \left[ (l(\mathcal{E}, \mathcal{Y}, \mathcal{F}) + \rho \log \frac{p(\mathcal{E}, \mathcal{Y}, \mathcal{F} | \mathcal{Z})}{p_{\theta}(\mathcal{E}, \mathcal{Y}, \mathcal{F} | \mathcal{Z})}) + \rho \nabla_{\theta'} \log p_{\theta'} \right] \quad (21)$$

by a unbiased MC estim.

$$\approx \frac{1}{M} \sum_{i \in [M]} \left[ (l(\mathcal{E}, \mathcal{Y}, \mathcal{F}) + \rho \log \frac{p(\mathcal{E}, \mathcal{Y}, \mathcal{F} | \mathcal{Z})}{p_{\theta}(\mathcal{E}, \mathcal{Y}, \mathcal{F} | \mathcal{Z})}) + \rho \nabla_{\theta'} \log p_{\theta'} \right] \quad (22)$$

## 2 Seminar on Self/Un-Supervised Learning @ 2020/9/16

### 2.1 Self-Learning @ Video Learning

Supervised success: good & sufficient data, a way different from human!  $\Rightarrow$  Linda Smith, *The Dev. of Embodied Cognition*

**Paragidims:**

- Use proxy task(e.g. semantics repr.) for a repr., use linear probing for downstream task.
- Use proxy task(e.g. semantics repr.) for a repr., generalizable with *zero annotation*

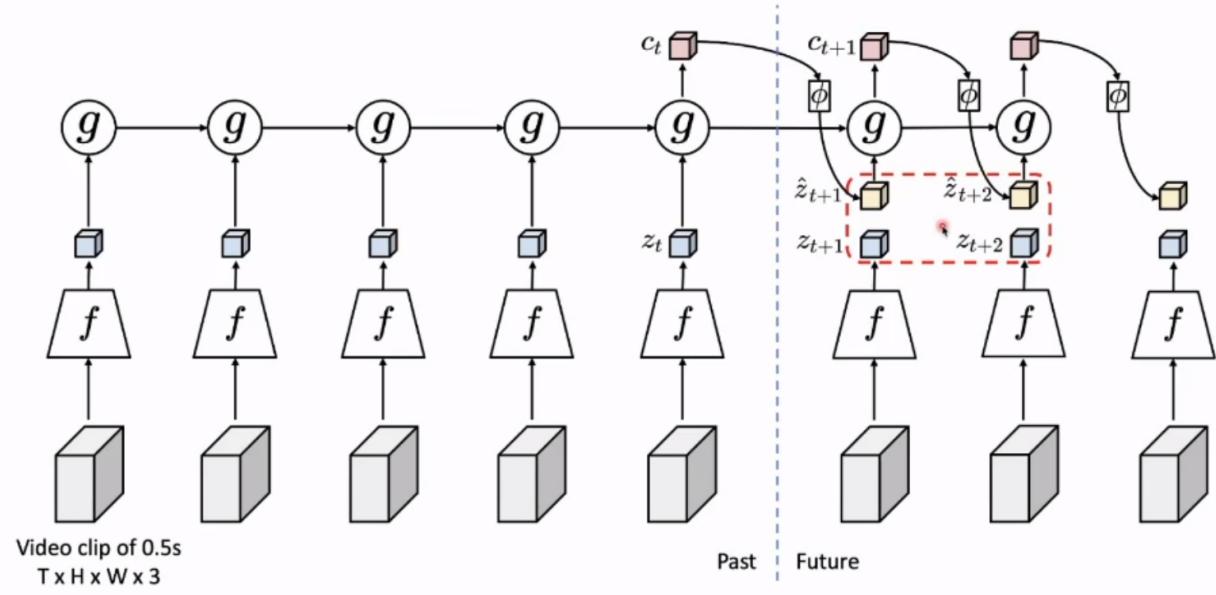
**Why video-based self-supervised:** like what human percepts, rich info; might with audio.

Proxy loss design: temporal info, spatial cohenrence, motions of obj., multimodal

**Temporal:**

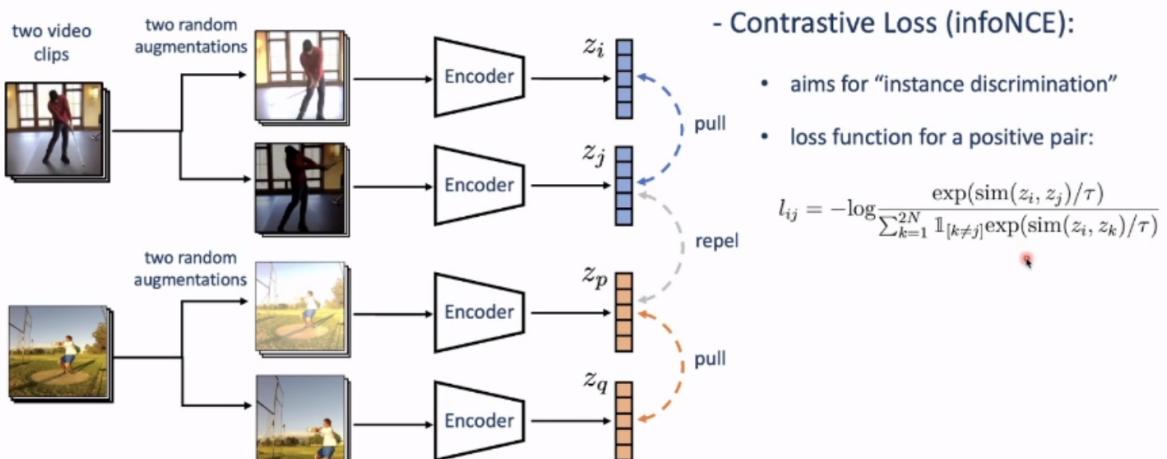
- shuffle & learn
- forward or backward?(arrow of time)
- SpeedNet: which speed(frame-rate) is normal/speed-up
- ===Weak, irrelative with downstream tasks==
- DPC: *learn repr. in predicting future in video*

## Approach - DPC

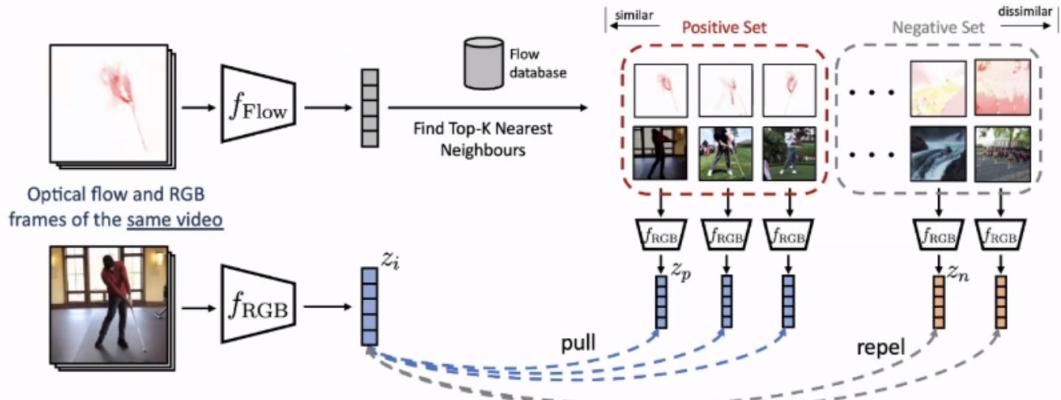


DPC Arch.:encoder-decoder like, contrast learning(infoNCE)

## Self-supervised learning with videos (CoCLR)



## Self-supervised learning with videos (CoCLR)



### Multi-Instance Contrastive Loss (MIL-NCE):

- Features from the positive set are pulled together
- Features **NOT** from the positive set are pushed apart
- Optical flow helps RGB frames to go beyond instance discrimination

$$\mathcal{L}_{\text{CoCLR}} = -\mathbb{E} \left[ \log \frac{\sum_{p \in \mathcal{P}_i} \exp(z_i \cdot z_p)}{\sum_{p \in \mathcal{P}_i} \exp(z_i \cdot z_p) + \sum_{n \in \mathcal{N}_i} \exp(z_i \cdot z_n)} \right]$$

CoCLR: SimCLR like(infoNCE), colearning multimodally with motion flow(MIL-NCE, with noise added)

Audio-Video Co-learning: train a net to check if image/audio clip are same-sourced! Get both video/audio repr.

MAST: self-supervised tracking, give 1st frame mask(seg.), predict sequential segmentations

**Next:**

- More efficient learning
- Scale up model to uncurated data, like GPT-1/2/3
- Design proxy task for obj-centric learning
- Design and understand **effective memory!!!**(for video task especially)
- Hand-crafted proxy task  $\Rightarrow$  Auto proxy task design?
- Theoretic: small or negative improvement, in upstream task to downstream task.
- Are there difficult task for supervised learning but easy for SSL(e.g. unable to label)?

## 2.2 Transformation Equivariance vs. Invariance @ Visial Repr. Learning

**Contents:**

- TER(Transformation Equivariance Repr.)

- AET(AutoEncoding Transformation)
  - AVT(Autoencoding Variational Transformation)
  - SAT(Semi-supervised Autoencoding Transformation)

CNN = Translation Equivariant Repr. + FC Classifier. Go beyond: Transformation Equivariant Repr. + Tranformation Invariant Classifier

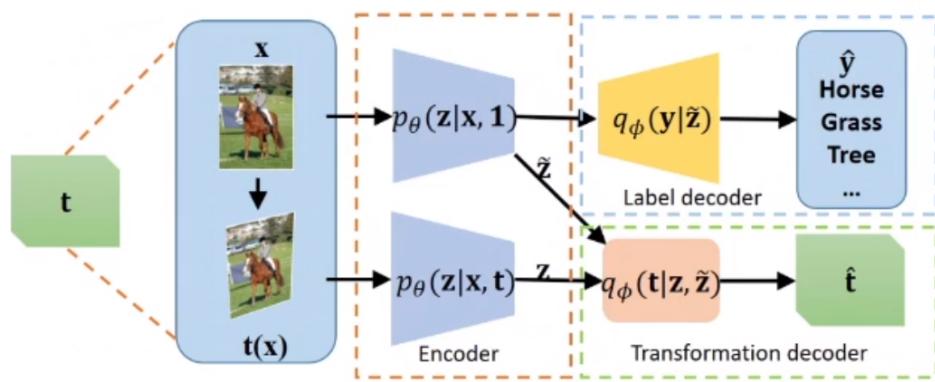
Trans. Equiv.:  $E(t(x)) = \rho(t)[E(x)]$ . Trans. Inv. is when  $\rho \equiv 1_E$ .

Steerability:  $\rho$  is inpend. with sample  $x$ .

Targets: Non-linear  $\rho$ , General Transformation(e.g. recoloring)

SAT: Semi-Supervised Autoencoding Transformation

- Adding a label decoder  $q_\phi(\mathbf{y}|\tilde{\mathbf{z}})$  to approximate the posterior  $p_\theta(\mathbf{y}|\mathbf{x})$



## Variational Bound

- By introducing label decoder and transformation decoder, we have

- Jointly maximizing over encoder  $\theta$  and decoders  $\phi$

$$\max_{\theta, \phi} \mathbb{E}_{p_\theta(\mathbf{y}, \mathbf{z}, \tilde{\mathbf{z}})} \log q_\phi(\mathbf{y} | \mathbf{z}, \tilde{\mathbf{z}}) + \mathbb{E}_{p_\theta(\mathbf{t}, \mathbf{z}, \tilde{\mathbf{z}})} \log q_\phi(\mathbf{t} | \mathbf{z}, \tilde{\mathbf{z}})$$



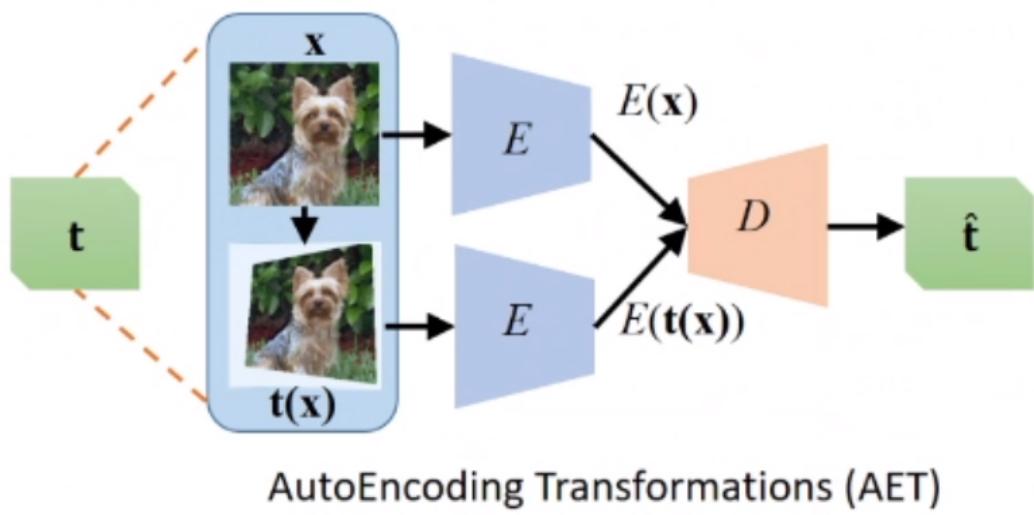
SAT:

- add a label decoder compared to AVT.
- variational surrogate  $\Rightarrow$  cross-entropy loss on supervised data + AVT loss

Contrastive Learning: more utilized trans-invariant repr. Future: unifying trans-inv/equiv repr.

### 3 AET, AVT: Autoencoding Transformations

**Idea** autoencoders used in modeling transformations rather than images in order to learn general repr.



AET:

- use autoencoders to learn **transformations**
- trans. generated randomly for self-supervised learning
- use Siamese net as encoder backbone
- AET loss: parameterized, non-parametric, GAN-induced

Losses in AET:

- parameterized transformations: if trans. are parameterized  $T \in \{t_\theta | \theta \in \Theta\}$ , loss can be defined as norm of param. diff.

$$l(t_\theta, t_{\hat{\theta}}) = \|\theta - \hat{\theta}\|.$$

- for non-parametric trans., use expected distance on source domain

$$l(t, \hat{t}) = \mathbb{E}_{x \sim X} \{ \text{dist}(t(x), \hat{t}(x)) \}$$



- GAN-induced trans.: image transformed in form  $G(x, z)$ , we have loss

$$l(t_z, t_{\hat{z}} = \|z - \hat{z}\|.)$$

Idea of AVT use prob. dist. to model trans., VAE like modeling!

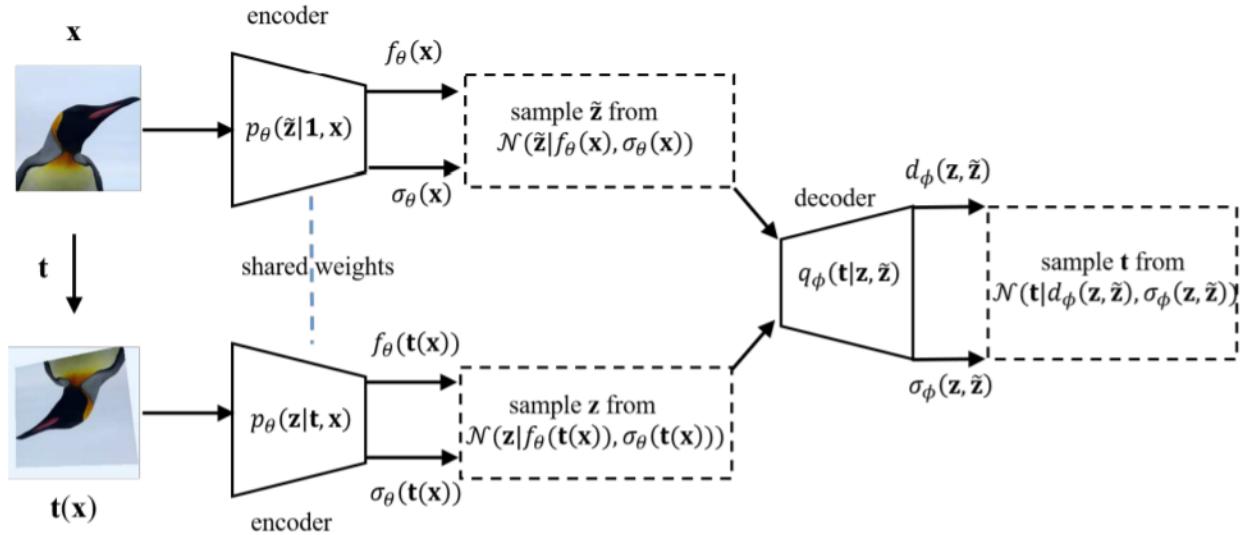


Figure 1: The architecture of the proposed AVT. The original and transformed images are fed through the encoder  $p_\theta$  where  $\mathbf{1}$  denotes an identity transformation to generate the representation of the original image. The resultant representations  $\tilde{\mathbf{z}}$  and  $\mathbf{z}$  of original and transformed images are sampled and fed into the transformation decoder  $q_\phi$  from which the transformation  $\mathbf{t}$  is sampled.

AVT:

- maximize mutual info  $I(t; z|\tilde{z})$
- variational bound, introducing a decoder  $q_\phi(t|z, \tilde{z})$ :

$$I(t; z|\tilde{z}) = H(t|\tilde{z}) - H(t|z, \tilde{z}) \quad (23)$$

$$= H(t|\tilde{z}) + \mathbb{E}_{p_\theta(t, z, \tilde{z})}[p_\theta(t|z, \tilde{z})] \quad (24)$$

$$= H(t|\tilde{z}) + \mathbb{E}_{p(t, z, \tilde{z})}[q_\theta(t|z, \tilde{z})] + \mathbb{E}_{p(z, \tilde{z})}[D(p_\theta(t, z, \tilde{z})||q_\phi(t|z, \tilde{z}))] \quad (25)$$

$$\geq H(t|\tilde{z}) + \mathbb{E}_{p(t, z, \tilde{z})}[q_\phi(t|z, \tilde{z})] \equiv \tilde{I}(t; z|\tilde{z}) \quad (26)$$

$$\Rightarrow \max_{\theta, \phi} \mathbb{E}_{p(t, z, \tilde{z})}[q_\phi(t|z, \tilde{z})] \quad (27)$$

- specifically in batch-wise formulation:

$$\mathbb{E}_{p(t, z, \tilde{z})}[q_\phi(t|z, \tilde{z})] \approx \frac{1}{n} \sum_{i=1}^n \log \mathcal{N}(t^i | d_\phi(z^i, \tilde{z}^i), \sigma_\phi(z^i, \tilde{z}^i)) \quad (28)$$

$$\text{where } z^i = f_\theta(t^i(x^i)) + \sigma_\theta(t^i(x^i)) \odot \epsilon^i \quad (29)$$

$$\text{and } \tilde{z}^i = f_\theta(x^i) + \sigma_\theta(x^i) \odot \tilde{\epsilon}^i \quad (30)$$

$$\text{where } \epsilon^i, \tilde{\epsilon}^i \sim (\epsilon|0, I), t^i \sim p(t) \text{(predifined or so?)} \quad (31)$$

- trick: take 5 samples to full explore the distribution

## 4 Flow-Based Generative Models

### 4.1 Outline & Basics

Two random vector of same dim.:

$$X \sim P_X(x), z \sim \Pi_Z(z), \text{find mapping } f : Z \rightarrow X = x(z), \quad (32)$$

we have

$$\begin{cases} p_X(x) = \pi_Z(f^{-1}(x)) |\det J(f)|^{-1} \\ \pi_Z(z) = p_X(f(z)) |\det J(f)| \end{cases} \quad (33)$$

use a simple dist. on  $Z$  and invertibly generate  $X \sim p_G(x)$ :  $x = G(z)$ . train  $G^{-1}$  as a discriminator.  
keys: invertible, easy-to-compute  $G^{-1}$ , easy-to-compute Jacobian determinant.

“Coupling Layer”:

$$\begin{cases} (\text{copy}) x_i = z_i, i \leq d \\ (\text{affine}) x_i = \beta_i z_i + r_i, d < i \leq D \end{cases} \quad (34)$$

$$\beta_{d+1, \dots, D} = F(z_{d+1, \dots, D}), \gamma_{d+1, \dots, D} = H(z_{d+1, \dots, D}) \quad (35)$$

$$J_G = \left[ \begin{array}{c|c} \mathbf{I}_d & \mathbf{O} \\ \hline M(\text{non-matter}) & D(\text{diagonal}) \end{array} \right] \quad (36)$$

$$\det J_G = \prod_{k=d+1..D} \frac{\partial x_k}{\partial z_k} \quad (37)$$

Use many coupling layer to enhance expressive capability. Parts of image does not change  
⇒exchange copy/affine split:

- exchange within channel
- exchange channel ⇒channel rotation use matrix/ $1 \times 1$  convolution in MoFlow

---

## 4.2 MoFlow

**Summary** Flow-based on molecular graphs, channel rotation as  $1 \times 1$  convolution, relational GCN layer, graph conditional flow(GCF), use sigmoid rather than exp, split dimensions.

“Coupling Layer”:

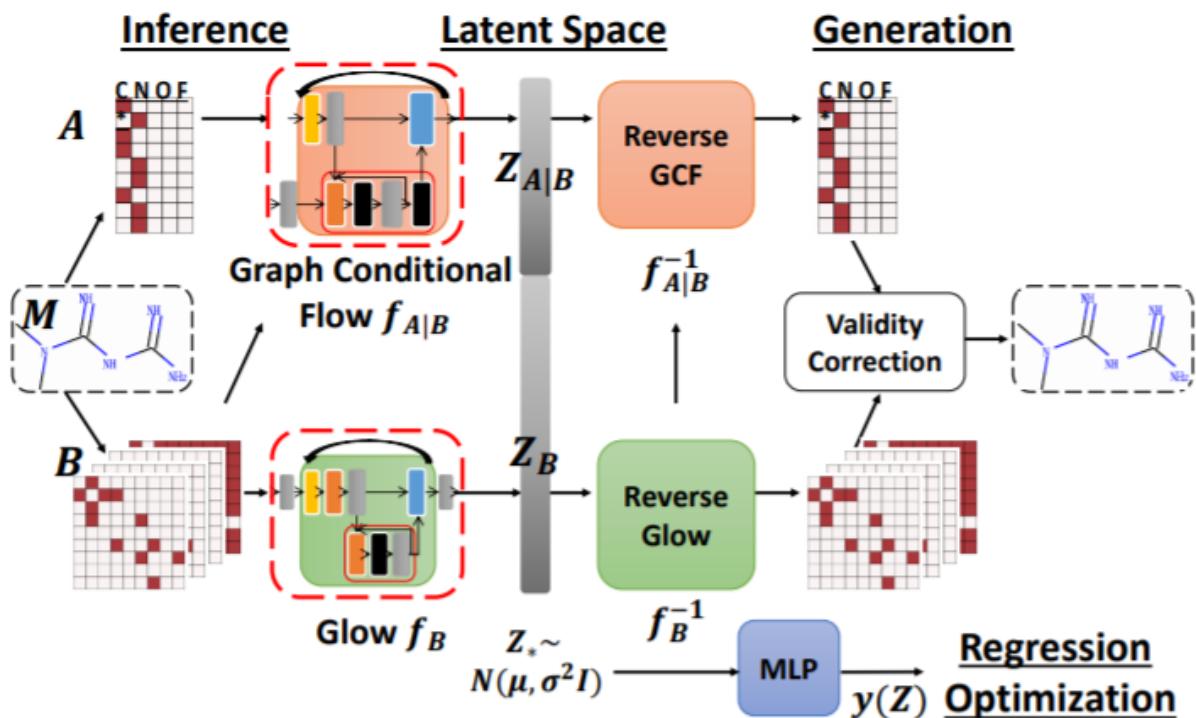
$$Z_{1:d} = X_{1:d} \quad (38)$$

$$Z_{d+1:n} = X_{d+1:n} \odot \text{sigmoid}(S_\Theta(X_{1:d})) + T_\Theta(X_{1:d}) \quad (39)$$

here  $S \sim \text{scaling}$ ,  $T \sim \text{translation}$ , both by DNNs. 每个耦合层交换上一层 copy 的 dims, 通过一个 channel 上的旋转  $W \in \mathbb{R}^{c \times c}$ , 等价于一个  $1*1$  卷积, 变换后的 Y 分为  $(Y_{1:c/2}, Y_{c/2+1,n})$  送入下一层. 采用 split-dims 的 trick, 增加交换 channel 的模型自由度增加:  $X \in \mathbb{R}^{c \times n \times n} \Rightarrow \mathbb{R}^{ch^2 \times n/h \times n/h}$



#### 4.2.1 GCF/Graph Conditional Flow



**Figure 1: The outline of our MoFlow.** A molecular graph  $M$  (e.g. Metformin) is represented by a feature matrix  $A$  for atoms and adjacency tensors  $B$  for bonds. **Inference:** the graph conditional flow (GCF)  $f_{\mathcal{A}|\mathcal{B}}$  for atoms (Sec. 4.2) transforms  $A$  given  $B$  into conditional latent vector  $Z_{A|B}$ , and the Glow  $f_B$  for bonds (Sec. 4.3) transform  $B$  into latent vector  $Z_B$ . The latent space follows a spherical Gaussian distribution. **Generation:** the generation process is the reverse transformations of previous operations, followed by a validity correction (Sec. 4.4) procedure which ensures the chemical validity. We summarize MoFlow in Sec. 4.5. **Regression and optimization:** the mapping  $y(Z)$  between latent space and molecular properties are used for molecular graph optimization and property prediction (Sec. 5.3, Sec. 5.4).

Def.(B conditioned flow) ....

We have

$$J_{A|B} = \frac{\partial f_{A|B}}{\partial(A, B)} = \begin{bmatrix} \frac{\partial f_{A|B}}{\partial A} & \frac{\partial f_{A|B}}{\partial B} \\ \hline O & I \end{bmatrix} \quad (40)$$

$$\det J_{A|B} = \det \frac{\partial f_{A|B}}{\partial A} \quad (41)$$

GCF layer:

$$Z_{A|B} = (Z_{A_1|B}, Z_{A_2|B}), A = (A_1|A_2) \quad (42)$$

$$\begin{cases} \text{copy } Z_{A_1|B} = A_1 \\ \text{affine } Z_{A_2|B} = A_2 \odot \text{sigmoid}(S_\Theta(A_1|B)) + T_\Theta(A_1|B) \end{cases} \quad (43)$$

Special designed  $S, T$  using R-GCN:

$$\text{graphconv}(A_1) = \sum_{i=[C]} \tilde{B}_i(M \odot A)W_i + (M \odot A)W_0, \text{ where } M \text{ is the mask of split,} \quad (44)$$

$$\tilde{B} \text{ is normalized } \mathbf{B}_i : \tilde{B}_i = \mathbf{D}^{-1}\mathbf{B}_i, \quad (45)$$

$$\text{where } \mathbf{D} \text{ is the full deg-mat. } \mathbf{D} = \sum_{c,i} \mathbf{B}_{c,i,j} = \sum_c \mathbf{D}_i, \text{ computed only once!} \quad (46)$$

#### 4.2.2 Validity Correction & Misc

使用价约束

$$\sum_{c,j} c \times B_{c,i,j} \leq Valency_i + Ch_i, \text{ where } Ch_i \text{ is formal charge} \quad (47)$$

$$(48)$$

具体的有效性校验方法:

1. 检查价约束, 满足去 2, 否则去 3
2. 返回最大连通子图
3. 第  $i$  个原子不满足, 对于第  $i$  个原子, 删去最高阶键, 去 1

这种方法试图再分子上做最小修改来满足价约束.

**Note** 为了防止学到的 prob. dist. 退化, 在数据集上增加 dequantization, 每个 dim 加噪声  $\sim U[0, 0.6]$

---

### 4.3 GraphNVP

## 5 WGAN

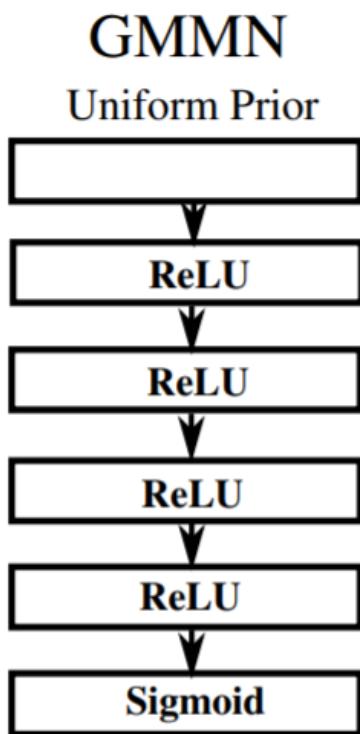
Idea Wasserstain 距离代替 KL/JS 距离.

### Method

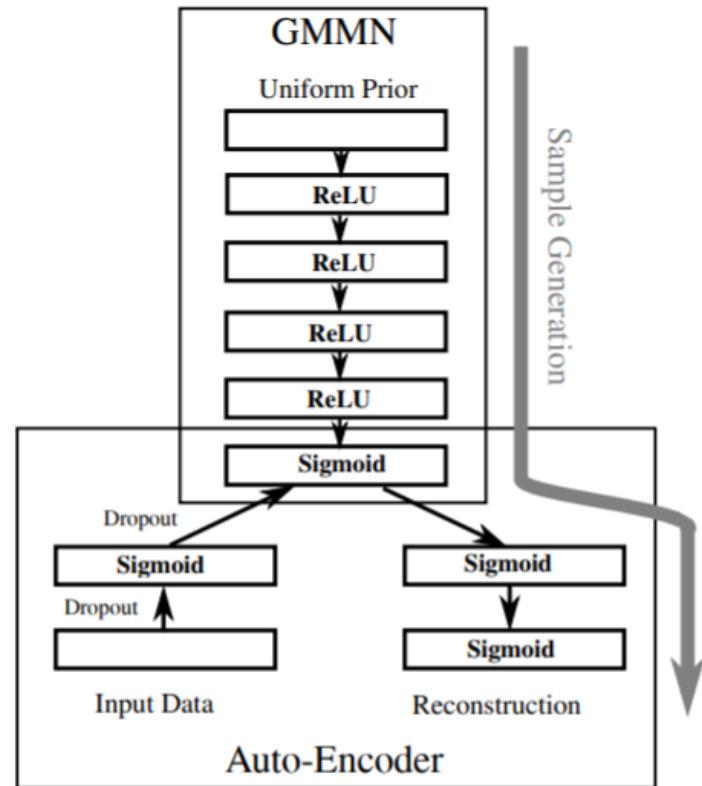
- 判别器不用 sigmoid, loss 不取 log
- 判别器参数截断  $\Rightarrow$  为了让判别器 Lipschitz 连续.
- trick: 不用基于 momentum 的优化器 (Adam etc.), 用 RMSProp/SGD.

## 6 GMMN+AE

### 6.1 Structure & Idea



(a) GMMN



(b) GMMN+AE

Figure 1. Example architectures of our generative moment matching networks. (a) GMMN used in the input data space. (b) GMMN used in the code space of an auto-encoder.

Use MMD(Maximum Mean Discrepancy) loss:

$$L_{MMD^2} = \left\| \frac{1}{N} \sum_i \phi(x_i) - \frac{1}{M} \sum_j \phi(y_j) \right\|^2 \quad (49)$$

$$= \frac{1}{N^2} \sum_{i,i'} K(x_i, x_{i'}) + \frac{1}{M^2} \sum_{i,i'} K(y_i, y_{i'}) - \frac{1}{NM} \sum_{i,j} K(x_i, y_j) \quad (50)$$

使用 k 阶多项式作为核, 则等价于匹配 k 阶矩!  $\Rightarrow$  使用高斯核, 以匹配所有阶矩(看作幂级数), 这也是 GMMN 的名字由来 (Moment-Matching):

$$K(x, y) = \exp\left(-\frac{1}{2\sigma}\|x - y\|^2\right) \quad (51)$$

设生成的数据为  $(x_i^s)$ , gt. 为  $(x_i^d)$ , 则偏导

$$\frac{\partial L_{MMD^2}}{\partial x_{ip}^s} = \frac{1}{\sigma} \left( \frac{2}{M^2} \sum_{j=[M]} K(x_i^s, x_j^s)(x_{jp}^s - x_{ip}^s) - \frac{2}{NM} \sum_{j=[N]} K(x_i^s, x_j^d)(x_{jp}^d - x_{ip}^s) \right) \quad (52)$$

## 6.2 Training

1. 逐层训练 AE
2. Finetune AE
3. 训练 GMMN

## 7 FoldingNet - An AutoEncoder

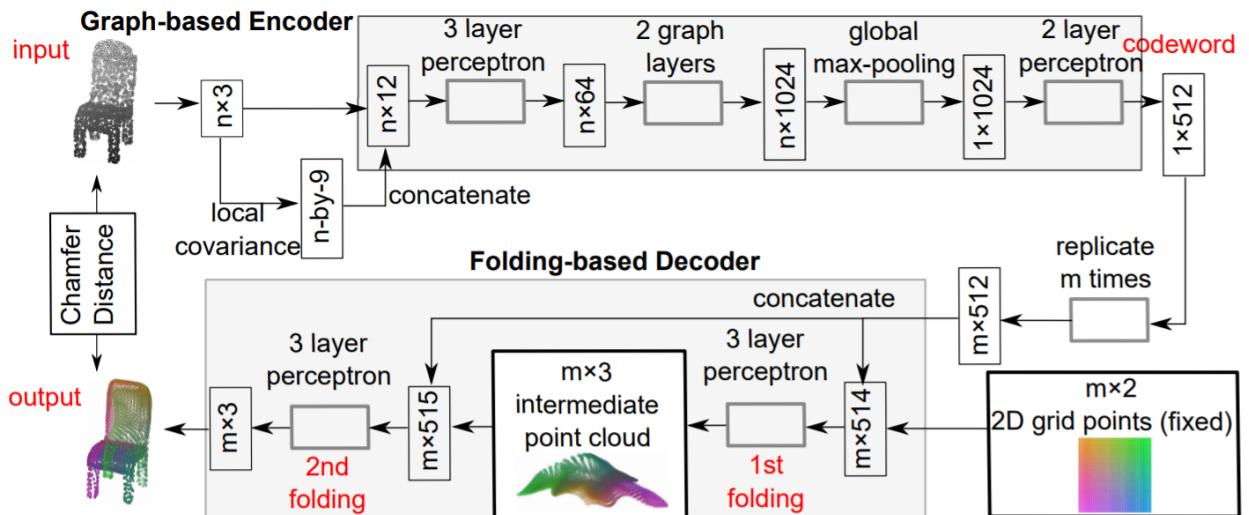


Figure 1. **FoldingNet Architecture.** The graph-layers are the graph-based max-pooling layers mentioned in (2) in Section 2.1. The 1st and the 2nd folding are both implemented by concatenating the codeword to the feature vectors followed by a 3-layer perceptron. Each perceptron independently applies to the feature vector of a single point as in [41], i.e., applies to the rows of the  $m$ -by- $k$  matrix.

使用 (扩展的)Chamfer 距离

$$d_{CH}(S, \hat{S}) = \max\left\{ \frac{1}{|S|} \sum_{x \in S} \min_{x \in \hat{S}} \|x - \hat{x}\|, \frac{1}{|\hat{S}|} \sum_{\hat{x} \in \hat{S}} \min_{x \in S} \|x - \hat{x}\| \right\} \quad (53)$$

这个距离让两个点云的点互相配准.

使用基于图的 Encoder: 使用的特征为局部 (KNN 上的) 协方差<sup>1</sup>+ 位置 ( $n \times 12$ ), 简要结构:  
MLP+GNN-Aggregation+MLP⇒Codeword 其中 Graph Layers

$$\mathbf{Y} = \mathbf{A}_{\max}(\mathbf{X})\mathbf{K} \quad (54)$$

$$\mathbf{A}_{\max}(\mathbf{X})_{ij} = \text{ReLU}(\max_{k \in \mathcal{N}(i)} x_{kj}) \quad (55)$$

基于折叠的 Decoder: 重复  $m$  次 codeword, 和  $2d$  格点 concat 送到 MLP(1st-folding) 得到中间折叠点云, 和 codeword concat 之后再送到第二个 folding-mlp 中得到结果.

**Prop.** Encoder proposed is permutation-invariant.

**Prop.** Decoder proposed can shape arbitrary point cloud.

## 8 PointFlow: Flow-based Generative Model on Point Clouds

Idea As Title

### 8.1 Continuous Normalizing Flow(CNF)

正则化流, 通过一系列可逆变换  $f_i$ :

$$x = f_1 \circ \dots \circ f_n(y) \quad (56)$$

$$\log P(x) = \log P(y) - \sum_i |\log \det \mathcal{J}_{f_i}| \quad (57)$$

离散的正则化流被推广到连续的正则化流—CNFs

$$\frac{\partial y(t)}{t} = f(y(t), t) \quad (58)$$

$$\text{Thus } = y(t_0) + \int_{t_0}^{t_1} f(y(t), t) dt, y(t_0) \sim P(y) \quad (59)$$

$$\log P(x) = \log P(y(t_0)) - \int_{t_0}^{t_1} \mathcal{T}r\left(\frac{\partial f}{\partial y(t)}\right) dt \quad (60)$$

一个黑盒 ODE 求解器可以用于估计流的输出和输入的梯度!

### 8.2 Variational Auto-Encoder

Optimize ELBO

$$\log P_\theta(X) \geq \log P_\theta(X) - D_{KL}(Q_\phi(z|X)||P_\theta(z|X)) \quad (61)$$

$$= \mathbb{E}_{Q_\phi(z|X)} [\log P_\theta(X|z)] - D_{KL}(Q_\phi(z|X)||P_\psi\theta(z)) \quad (62)$$

<sup>1</sup>回忆协方差公式

$$\text{Cov}(\mathbf{X}) = \mathbb{E}[(\mathbf{X} - \mathbb{E}\mathbf{X})(\mathbf{X} - \mathbb{E}\mathbf{X})^T]$$

## 8.3 Model

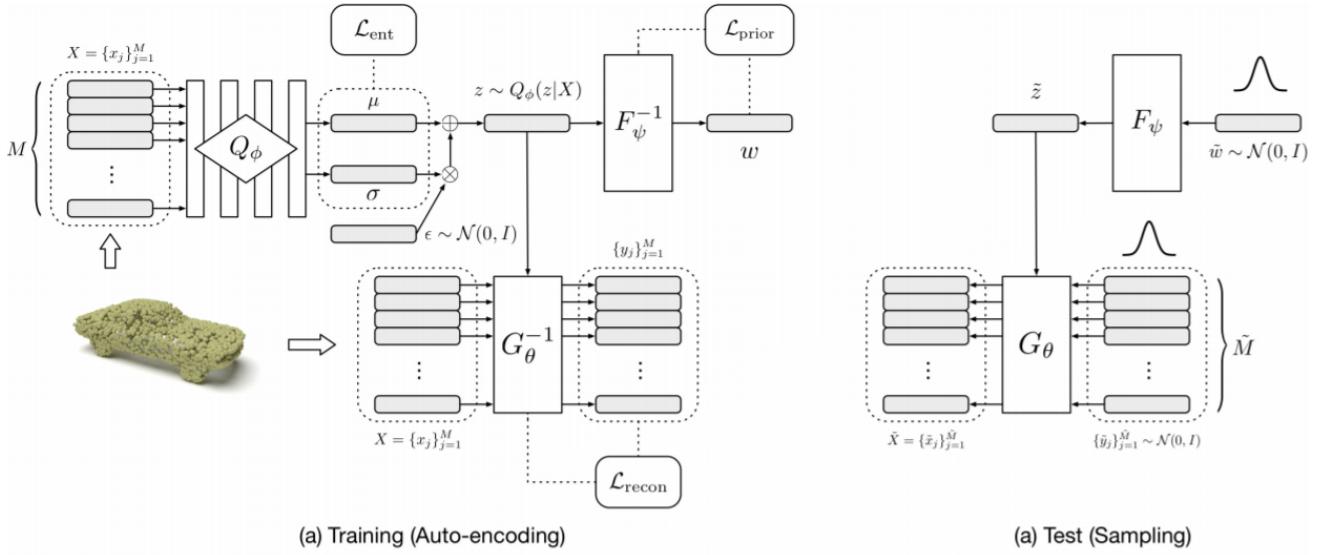


Figure 2: Model architecture. (a) At training time, the encoder  $Q_\phi$  infers a posterior over shape representations given an input point cloud  $X$ , and samples a shape representation  $z$  from it. We then compute the probability of  $z$  in the prior distribution ( $\mathcal{L}_{\text{prior}}$ ) through a inverse CNF  $F_\psi^{-1}$ , and compute the reconstruction likelihood of  $X$  ( $\mathcal{L}_{\text{recon}}$ ) through another inverse CNF  $G_\theta^{-1}$  conditioned on  $z$ . The model is trained end-to-end to maximize the evidence lower bound (ELBO), which is the sum of  $\mathcal{L}_{\text{prior}}$ ,  $\mathcal{L}_{\text{recon}}$ , and  $\mathcal{L}_{\text{ent}}$  (the entropy of the posterior  $Q_\phi(z|X)$ ). (b) At test time, we sample a shape representation  $\tilde{z}$  by sampling  $\tilde{w}$  from a Gaussian prior and transforming it with  $F_\psi$ . To sample points from the shape represented by  $\tilde{z}$ , we first sample points from the 3-D Gaussian prior and then move them according to the CNF parameterized by  $\tilde{z}$ .

**Summary** VAE-like. Decoder: Flow-based, i.e. CNF; Prior: CNF-based; Encoder: some simple permutation-invariant encoder.

### Notations

$$z \sim \text{Latent Repr. for Shape} \quad (63)$$

$$y \sim \text{Simple Distribution/Source Dist. to be Transformed} \quad (64)$$

$$x \sim \text{Point Cloud} \quad (65)$$

$$(66)$$

Point cloud lld

$$\log P_\theta(X|z) = \sum_{x \in X} \log P_\theta(x|z) \quad (67)$$

model  $P(x|z)$  by 条件 CNF

$$x = G_\theta(y(t_0); z) \quad (68)$$

$$= y(t_0) + \int_{t_0}^{t_1} g_\theta(y(t), t; z) dt, y(t_0) \sim P(y) = \mathcal{N}(0, I) \quad (69)$$

reconstruction lld:

$$\log P(x) = \log P(y(t_0)) - \int_{t_0}^{t_1} \mathcal{J}_{g_\theta(t)} dt \quad (70)$$

虽然用高斯分布的先验在 shape repr. 上可行, 但是有证据证明这受限的分布先验在 VAE 中会限制性能. 使用另一个 CNF 来参数化可学习的先验来减少影响

$$D_{KL}(Q_\phi(z|X)||P_\psi\theta(z)) = \mathbb{E}_{Q_\phi(z|X)}[\log P_\psi\theta(z)] - H(P_\psi\theta(z)) \quad (71)$$

obtain  $P_\psi$  by  $P(w) \sim \mathcal{N}(0, I)$  and CNF

$$z = F_\psi(w(t_0)) \quad (72)$$

$$\triangleq w(t_0) + \int_{t_0}^{t_1} f_\psi(w(t), t) dt, w(t_0) \sim P(w) = \mathcal{N}(0, I) \quad (73)$$

log-probability

$$\log P(x) = \log P(F_\psi^{-1}(z)) - \int_{t_0}^{t_1} \mathcal{J}_{f_\psi(t)} dt \quad (74)$$

最终的 loss term(ELBO)

$$\begin{aligned} \mathcal{L}(X; \phi, \psi, \theta) &= \mathbb{E}_{Q_\phi(z|x)} [\log P_\psi(z) + \log P_\theta(X | z)] + H[Q_\phi(z | X)] \\ &= \mathbb{E}_{Q_\phi(z|x)} \left[ \log P(F_\psi^{-1}(z)) - \int_{t_0}^{t_1} \text{Tr} \left( \frac{\partial f_\psi}{\partial w(t)} \right) dt \right. \\ &\quad \left. + \sum_{x \in X} \left( \log P(G_\theta^{-1}(x; z)) - \int_{t_0}^{t_1} \text{Tr} \left( \frac{\partial g_\theta}{\partial y(t)} \right) dt \right) \right] \\ &\quad + H[Q_\phi(z | X)] \end{aligned} \quad (75)$$

can be interpretes in 3 parts:

1. Prior:  $\mathcal{L}_{\text{prior}}(X; \psi, \phi) \triangleq \mathbb{E}_{Q_\phi(z|x)} [\log P_\psi(z)]$ , use reparametrization to MC-sample:

$$\mathbb{E}_{Q_\phi(z|x)} [\log P_\psi(z)] \approx \frac{1}{L} \sum_{l=1}^L \log P_\psi(\mu + \epsilon_l \odot \sigma) \quad (76)$$

2. Recon. ld.:  $\mathcal{L}_{\text{recon}}(X; \theta, \phi) \triangleq \mathbb{E}_{Q_\phi(z|x)} [\log P_\theta(X | z)]$ , 依然使用 MC 采样估计.

3. Posterior Entopy:  $\mathcal{L}_{\text{ent}}(X; \phi) \triangleq H[Q_\phi(z | X)]$ , has form

$$H[Q_\phi(z | X)] = \frac{d}{2}(1 + \ln(2\pi)) + \sum_{i=1}^d \ln \sigma_i \quad (77)$$

## 9 FFJORD

### 9.1 CNF

use some base dist.  $\mathbf{z}_0 \sim p_{z_0}(\mathbf{z}_0)$ , 通过含时 ODE 得到要建模的分布

$$\mathbf{z}(t_0) = \mathbf{z}_0 \quad (78)$$

$$\frac{\partial \mathbf{z}}{\partial t} = f(\mathbf{z}(t), t; \theta) \quad (79)$$

log-pdf 的方程 (*instantaneous change of variables form.*)

$$\frac{\partial \log p(\mathbf{z}(t))}{\partial t} = -\text{Tr}\left(\frac{\partial f}{\partial \mathbf{z}(t)}\right) \quad (80)$$

$$\log p(\mathbf{z}(t_1)) = \log p(\mathbf{z}(t_0)) - \int_{t_0}^{t_1} \text{Tr}\left(\frac{\partial f}{\partial \mathbf{z}(t)}\right) dt \quad (81)$$

$$\underbrace{\begin{bmatrix} \mathbf{z}_0 \\ \log p(\mathbf{x}) - \log p_{z_0}(\mathbf{z}_0) \end{bmatrix}}_{\text{solutions}} = \underbrace{\int_{t_1}^{t_0} \begin{bmatrix} f(\mathbf{z}(t), t; \theta) \\ -\text{Tr}\left(\frac{\partial f}{\partial \mathbf{z}(t)}\right) \end{bmatrix} dt}_{\text{dynamics}}, \underbrace{\begin{bmatrix} \mathbf{z}(t_1) \\ \log p(\mathbf{x}) - \log p(\mathbf{z}(t_1)) \end{bmatrix}}_{\text{initial values}} = \begin{bmatrix} \mathbf{x} \\ 0 \end{bmatrix} \quad (82)$$

### 9.2 Backpropagation through ODE Solutions with Adjoint Method

Problem: calc. deriv. based on loss func.

$$L(\mathbf{z}(t_1)) = L\left(\int_{t_0}^{t_1} f(\mathbf{z}(t), t; \theta) dt\right) \quad (83)$$

Pontryagin(1962) 证明

$$\frac{dL}{d\theta} = - \int_{t_1}^{t_0} \left(\frac{\partial L}{\partial \mathbf{z}(t)}\right)^T \frac{\partial f(\mathbf{z}(t), t; \theta)}{\partial \theta} dt \quad (84)$$

值  $-\partial L / \partial \mathbf{z}(t)$  称为 ODE 的伴随状态 (adjoint state). 使用一个 black-box ODE solver 来计算  $\mathbf{z}(t_1)$ , 再用初值  $\partial L / \partial \mathbf{z}(t_1)$  送进这个 ODE solver 来计算 (84)

### 9.3 Unbiased Linear-Time Log-Density Estimation

Hutchinson Estimator:

$$\text{Tr}(A) = E_{p(\epsilon)} [\epsilon^T A \epsilon] \quad (85)$$

holds if  $\mathbb{E}[\epsilon] = 0$ ,  $\text{Cov}\epsilon = I$  to avoid randomness, fix noise at each round of solving ODE

$$\begin{aligned} \log p(\mathbf{z}(t_1)) &= \log p(\mathbf{z}(t_0)) - \int_{t_0}^{t_1} \text{Tr}\left(\frac{\partial f}{\partial \mathbf{z}(t)}\right) dt \\ &= \log p(\mathbf{z}(t_0)) - \int_{t_0}^{t_1} \mathbb{E}_{p(\epsilon)} \left[ \epsilon^T \frac{\partial f}{\partial \mathbf{z}(t)} \epsilon \right] dt \\ &= \log p(\mathbf{z}(t_0)) - \mathbb{E}_{p(\epsilon)} \left[ \int_{t_0}^{t_1} \epsilon^T \frac{\partial f}{\partial \mathbf{z}(t)} \epsilon dt \right] \end{aligned} \quad (86)$$

噪声分布可以选为高斯分布/Rademacher 分布<sup>2</sup> 并且向量和 Jacobian 的乘积, i.e.  $\epsilon \frac{\partial f}{\partial z(t)}$ , 可以快速算出 (通过 auto-diff)

Trick: Bottleneck width  $H$  to reduce variance of estimator.

---

### Algorithm 1 Unbiased stochastic log-density estimation using the FFJORD model

---

**Require:** dynamics  $f_\theta$ , start time  $t_0$ , stop time  $t_1$ , minibatch of samples  $\mathbf{x}$ .

```

 $\epsilon \leftarrow \text{sample\_unit\_variance}(\mathbf{x}.\text{shape})$                                 ▷ Sample  $\epsilon$  outside of the integral
function  $f_{aug}([\mathbf{z}_t, \log p_t], t)$ :                                         ▷ Augment  $f$  with log-density dynamics.
     $f_t \leftarrow f_\theta(\mathbf{z}(t), t)$                                               ▷ Evaluate dynamics
     $g \leftarrow \epsilon^T \frac{\partial f}{\partial \mathbf{z}}|_{\mathbf{z}(t)}$                                ▷ Compute vector-Jacobian product with automatic differentiation
     $\tilde{\text{Tr}} = \text{matrix\_multiply}(g, \epsilon)$                                      ▷ Unbiased estimate of  $\text{Tr}(\frac{\partial f}{\partial \mathbf{z}})$  with  $\epsilon^T \frac{\partial f}{\partial \mathbf{z}} \epsilon$ 
    return  $[f_t, -\tilde{\text{Tr}}]$                                                  ▷ Concatenate dynamics of state and log-density
end function
 $[\mathbf{z}, \Delta_{logp}] \leftarrow \text{odeint}(f_{aug}, [\mathbf{x}, \vec{0}], t_0, t_1)$    ▷ Solve the ODE, ie.  $\int_{t_0}^{t_1} f_{aug}([\mathbf{z}(t), \log p(\mathbf{z}(t))], t) dt$ 
 $\log \hat{p}(\mathbf{x}) \leftarrow \log p_{\mathbf{z}_0}(\mathbf{z}) - \Delta_{logp}$                            ▷ Add change in log-density
return  $\log \hat{p}(\mathbf{x})$ 
```

---

## 10 Dequantization to Learn Discrete Distribution

为了近似一个离散空间上的 pd., 需要通过在数据点上加入噪声, 使用“去量化”技巧 (dequantization). 可变性更好的 noise  $\Rightarrow$  更紧的下界  $\Rightarrow$  learned noise?.

**Theorem** 加入合适的噪声后的连续随机变量的 ld(likelihood) 是对应离散随机变量 ld 的下界.

### 10.1 Dequantization as Latent Variable Model

$$P_{model}(x) = \int P_\theta(x|v)p(v)dv, \quad (88)$$

$$\text{where } P_\theta(x|v) = \mathbb{1}[v \in B_\theta(x)] \quad (89)$$

称  $P_\theta(x|v)$  是量化子 (quantizer). 不同的量化子导致了不同的去量化方法. Half-infinite dequant. for bin. var.:  $B(x) = \{x \cdot u | u \in \mathbb{R}_+^D\}$ ,  $x \in -1, 1$ ; Hypercube dequant. for grid var.(images etc.):  $B(x) = \{x + u | u \in [0, 1]^D\}$

上述积分难以计算, 引入去量化子  $q_\phi(v|x)$ , 注意它具有不重叠的紧支撑集, 为此标记  $u = v + x$

<sup>2</sup> 在  $\{-1, 1\}$  上均匀分布的离散分布

$$f(k) = \begin{cases} 1/2 & \text{if } k = -1 \\ 1/2 & \text{if } k = +1 \\ 0 & \text{otherwise} \end{cases} \quad (87)$$



$$P_{model}(x) = \int \frac{q_\phi(u|x) P_\theta(x|v)p(v)}{q_\phi(u|x)} dv \quad (90)$$

$$= \mathbb{E}_{u \sim q_\phi(u|x)} \left[ \frac{P_\theta(x|v)p(v)}{q_\phi(u|x)} \right] \quad (91)$$

现有的方法常常使用格点积分作为离散和连续模型的区分

$$P(x) = \int_{[0,1]^D} p(x+u) du \quad (92)$$

下面提出三种 dequant. : i) variational inference, ii) weighted importance sampling, iii) variational Renyi approx.

## 10.2 Variational Dequantization

根据 Jensen 不等式, 得到 lld 的变分代理函数

$$\log P_{model}(x) \geq \mathbb{E}_{u \sim q_\phi(u|x)} \left[ \log \frac{P_\theta(x|v)p_\theta(v)}{q_\phi(u|x)} \right] \quad (93)$$

注意去量化子要有紧支撑集, 所以对其输出进行 sigmoid; 并且进而有

$$\log P_{model}(x) \geq \mathbb{E}_{u \sim q_\phi(u|x)} [\log p_\theta(v)] + \mathbb{H}[q_\phi] \quad (94)$$

熵一项防止了概率分布退化到离散点上的 delta-peak, 从而推出了变分去量化 (vi dequant.)

## 10.3 Importance-Weighted Dequantization

除此之外, 还可以把去量化分布看作 proposal dist., 用采样多次替代 Jensen 不等式:

$$\log P_{model}(x) \geq \log \left[ \frac{1}{K} \sum_{k \in [K]} \frac{P_\theta(x|v_k)p_\theta(v_k)}{q_\phi(u_k|x)} \right] \quad (95)$$

若提案分布限制于紧支撑集上, 则有

$$\log P_{model}(x) \geq \log \left[ \frac{1}{K} \sum_{k \in [K]} \frac{p_\theta(v_k)}{q_\phi(u_k|x)} \right] \quad (96)$$

$$= \log [w_k(x)] \quad (97)$$

$$\text{where } w_k(x) \triangleq \frac{p_\theta(v_k)}{q_\phi(u_k|x)} \quad (98)$$

若  $K \rightarrow \infty$ , 则取等号, 否则给出了 lld 的一个下界 (iw-bound), 故给出了 vi 界的更好估计  $\Rightarrow$  iw-dequant.

## 10.4 Renyi Dequantization

vi/iw-去量化都可以看作变分 Renyi 去量化的特例. lld 可以用 Renyi Divergence 提供下界

$$\log P_{model}(x) \geq \frac{1}{1-\alpha} \log \left[ \frac{1}{K} \sum_{k \in [K]} \left( \frac{P_\theta(x|v_k)p_\theta(v_k)}{q_\phi(u_k|x)} \right)^{1-\alpha} \right] \quad (99)$$

$$\text{where } \alpha \in [0, 1) \quad (100)$$

vi-bound  $\alpha \rightarrow 1$ , iw-bound  $\alpha = 0$ . [Li & Turner, 2016] 考虑小于 0 的  $\alpha$ , 这可能在低采样数时提供更紧的下界 (当  $K \rightarrow \infty$ , 实际上提供了一个上界). 令  $\alpha = -\infty$ , 得到 VR-max(variational Renyi max-approximation), 一个 iw-bound 的快速估计

$$\log P_{model}(x) \approx \log \max_k w_k(x) \quad (101)$$

**Detail** 用 Cholesky 分解计算协方差矩阵  $\Lambda = \Gamma\Gamma^T$ ,  $\Gamma^T$  可学习.

## 10.5 Dequantization Distribution

**Uniform Dequant.**  $q_\phi(u|x)$  is uniform in  $\mathcal{B}(x)$

**Gaussian Dequant.** 更具表达力的是条件 logit-正态分布 (cond. logit-normal dist.)

$$q_\phi(u|x) = \text{sigmoid}(\mathcal{N}(\mu_\phi(x), \Sigma_\phi(x))) \quad (102)$$

**Flow-based Dequant.**

$$q_\phi(u|x) = q_\phi(\varepsilon = f_\phi(\text{sigmoid}^{-1}(u); x)|x) \det \mathbf{J} \quad (103)$$

由基分布  $q_\phi(\varepsilon|x)$  和流双射  $f \in \mathbb{R}^D \rightarrow \mathbb{R}^D$  组成. 这里的基分布采用对角高斯分布, 以及两种双射: coupling layer/flow/bipartite 和 autoregressive.

**Bipartite Dequant.** (Dinh et al., 2017) 使用流模型的耦合层:

$$\begin{cases} (\text{copy}) u_1 = \varepsilon + 1 \\ (\text{affine}) u_2 = \varepsilon_2 \odot s_\phi(\varepsilon_1; x) + t_\phi(\varepsilon_1) \end{cases} \quad (104)$$

$$(105)$$

为了保证所有分量都被变换, 应用另一个更改了 copy 层位置的耦合层.

**Autoregressive Dequant./ARD** (Kingma et al., 2016) 使用一个自回归模型

$$[m, s] = ARM_\phi(\varepsilon, h) \quad (106)$$

$$u = s \odot \varepsilon + m \quad (107)$$

其中  $h$  是上下文变量, 基于条件变量  $x$ , 通过网络  $s$  计算出来.

## 10.6 (Choice of) Continuous Distribution

可以按前一节那样任意地选择量化子  $p_\theta(v)$ . 但是在训练中需要采样  $v \sim p_\theta(v)$ , 故使用自回归模型是很慢的, 故只考虑对角协方差/正常协方差的高斯分布和二分 flow-based 模型.

# 11 DGI: Deep Graph Infomax

## 11.1 Backgrounds, Approach, Math

**Target** Learn a encoder  $\mathcal{E} = \mathcal{E}(\mathbf{X}, \mathbf{A}) : \mathbb{R}^{N \times F} \times \mathbb{R}^{N \times N} \rightarrow \mathbb{R}^{N \times F}$ .

**Approach** 最大化局部互信息. 使用 *Readout* 函数来获得全局图特征  $\vec{s} = \mathcal{R}(\mathcal{E}(\mathbf{X}, \mathbf{A}))$ . 为了能够计算 MI, 引入判别器  $\mathcal{D} : \mathbb{R}^F \times \mathbb{R}^F \rightarrow \mathbb{R}$ ,  $\mathcal{D}(\vec{h}_i, \vec{s})$  代表了两个图的 (repr.) 相似度. 判别器的负样本通过把一个图和一个不同的图联系在一起组成. 对于多图场景 (ModelNet/Molecule Graphs) 这可以通过采样其他图得到; 对于单图情景 (Cora etc.), 必须定义一个 (随机) 损坏函数  $\mathcal{C} : \mathbb{R}^{N \times F} \times \mathbb{R}^{N \times N} \rightarrow \mathbb{R}^{M \times F} \times \mathbb{R}^{M \times M}$

为此, 使用 contrastive loss

$$\mathcal{L} = \frac{1}{N+M} \left( \sum_{i=1}^N \mathbb{E}_{(\mathbf{X}, \mathbf{A})} \left[ \log \mathcal{D}(\vec{h}_i, \vec{s}) \right] + \sum_{j=1}^M \mathbb{E}_{(\tilde{\mathbf{X}}, \tilde{\mathbf{A}})} \left[ \log \left( 1 - \mathcal{D}(\vec{h}_j, \vec{s}) \right) \right] \right) \quad (108)$$

这个 Jensen-Shannon divergence 本质上是互信息的 estimator!

**Lemma 1.**  $\{\mathbf{X}^{(k)}\}_{k \in [|X|]}$  从  $p(\mathbf{X})$  中取出的一系列节点表示, 且  $p(\mathbf{X}^{(k)}) = p(\mathbf{X}^{(k')}) \forall k, k'$ , 并且  $\mathcal{R}(\odot)$  是确定性 Readout 函数,  $\vec{s}^{(k)} = \mathcal{R}(\mathbf{X}^{(k)})$ , 具有边缘分布  $p(\vec{s})$ . 则基于联合分布的最优分类器  $p(\mathbf{X}, \vec{s})$  和边缘分布的乘积  $p(\mathbf{X})p(\vec{s})$  的误差有上界  $\text{Err}^* = \frac{1}{2} \sum_{k=1}^{|X|} p(\vec{s}^{(k)})^2$ . 当  $\mathcal{R}$  是单射时达到上界.

**Corollary 1.** 此后都假设  $\mathcal{R}$  是单射, 假设  $\vec{s}$  的状态不少于  $|\mathbf{X}|$ , 则最优全局表示满足  $|\vec{s}^*| = |\mathbf{X}|$ .

**Theorem 1.**  $\vec{s}^* = \operatorname{argmax}_{\vec{s}} I(\mathbf{X}; \vec{s})$

**Theorem 2.** 令  $\mathbf{X}_i^{(k)} = \{\vec{x}_j\}_{j \in n(\mathbf{X}^{(k)}, i)}$ , 是第  $k$  层图卷积的特征,  $\vec{h}_i = \mathcal{E}(\mathbf{X}_i^{(k)})$ , 假设  $|\mathbf{X}_i| = |\mathbf{X}| = |\vec{s}| \geq |\vec{h}_i|$ , 则最小化  $p(\vec{h}_i, \vec{s})$  and  $p(\vec{h}_i)p(\vec{s})$  的  $\vec{h}_i$  也让 MI 最大化.

## 11.2 Algorithm

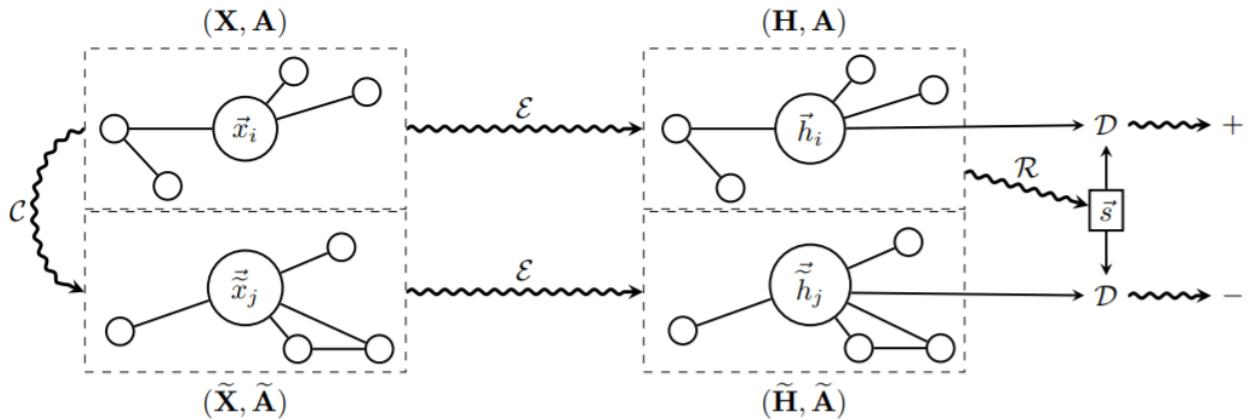


Figure 1: A high-level overview of Deep Graph Infomax. Refer to Section 3.4 for more details.

1. 从损坏函数中采样  $(\tilde{\mathbf{X}}, \tilde{\mathbf{A}}) \sim \mathcal{C}(\mathbf{X}, \mathbf{A})$
2. 获得正负样本的 patch node-repr.,  $\mathbf{H} = \mathcal{E}(\mathbf{X}, \mathbf{A}) = \{\vec{h}_1, \vec{h}_2, \dots, \vec{h}_N\}$ ,  $\tilde{\mathbf{H}} = \mathcal{E}(\tilde{\mathbf{X}}, \tilde{\mathbf{A}}) = \{\tilde{h}_1, \tilde{h}_2, \dots, \tilde{h}_M\}$
3. 获得全局特征表示  $\vec{s} = \mathcal{R}(\mathbf{H})$ .
4. 根据方程 (108) 更新  $\mathcal{R}, \mathcal{D}, \mathcal{E}$  参数.

**Details** 使用 PReLU, 迁移学习任务上 (transductive, Cora, Citeseer, PubMed) 使用 GCN

$$\mathcal{E}(\mathbf{X}, \mathbf{A}) = \sigma \left( \hat{\mathbf{D}}^{-\frac{1}{2}} \hat{\mathbf{A}} \hat{\mathbf{D}}^{-\frac{1}{2}} \mathbf{X} \Theta \right) \quad (109)$$

在推断任务上 (inductive, Reddit) 使用 mean-aggr 和 GraphSAGE-GCN

$$\text{MP}(\mathbf{X}, \mathbf{A}) = \hat{\mathbf{D}}^{-1} \hat{\mathbf{A}} \mathbf{X} \Theta \quad (110)$$

在多图任务上 (PPI) 使用三层带有 dense skip conn. 的 mean-pooling 层

$$\begin{aligned} \mathbf{H}_1 &= \sigma(\text{MP}_1(\mathbf{X}, \mathbf{A})) \\ \mathbf{H}_2 &= \sigma(\text{MP}_2(\mathbf{H}_1 + \mathbf{X} \mathbf{W}_{\text{skip}}, \mathbf{A})) \\ \mathcal{E}(\mathbf{X}, \mathbf{A}) &= \sigma(\text{MP}_3(\mathbf{H}_2 + \mathbf{H}_1 + \mathbf{X} \mathbf{W}_{\text{skip}}, \mathbf{A})) \end{aligned} \quad (111)$$

在 Readout 函数上使用简单的 graph-mean-aggr

$$\mathcal{R}(\mathbf{H}) = \sigma \left( \frac{1}{N} \sum_{i=1}^N \vec{h}_i \right) \quad (112)$$

在判别器上使用简单的双线性打分函数

$$\mathcal{D}(\vec{h}_i, \vec{s}) = \sigma(\vec{h}_i^T \mathbf{W} \vec{s}) \quad (113)$$

## 12 GraphSAGE: Inductive Representation Learning on Graph

### 12.1 Embedding Generation/FP

Idea 在 k-hops 上逐层做 aggr.! Weisfeiler-Lehman 图同构检验的连续推广.

---

#### Algorithm 1: GraphSAGE embedding generation (i.e., forward propagation) algorithm

---

**Input :** Graph  $\mathcal{G}(\mathcal{V}, \mathcal{E})$ ; input features  $\{\mathbf{x}_v, \forall v \in \mathcal{V}\}$ ; depth  $K$ ; weight matrices  $\mathbf{W}^k, \forall k \in \{1, \dots, K\}$ ; non-linearity  $\sigma$ ; differentiable aggregator functions  $\text{AGGREGATE}_k, \forall k \in \{1, \dots, K\}$ ; neighborhood function  $\mathcal{N} : v \rightarrow 2^{\mathcal{V}}$

**Output:** Vector representations  $\mathbf{z}_v$  for all  $v \in \mathcal{V}$

```
1  $\mathbf{h}_v^0 \leftarrow \mathbf{x}_v, \forall v \in \mathcal{V}$ ;  
2 for  $k = 1 \dots K$  do  
3   for  $v \in \mathcal{V}$  do  
4      $\mathbf{h}_{\mathcal{N}(v)}^k \leftarrow \text{AGGREGATE}_k(\{\mathbf{h}_u^{k-1}, \forall u \in \mathcal{N}(v)\})$ ;  
5      $\mathbf{h}_v^k \leftarrow \sigma(\mathbf{W}^k \cdot \text{CONCAT}(\mathbf{h}_v^{k-1}, \mathbf{h}_{\mathcal{N}(v)}^k))$   
6   end  
7    $\mathbf{h}_v^k \leftarrow \mathbf{h}_v^k / \|\mathbf{h}_v^k\|_2, \forall v \in \mathcal{V}$   
8 end  
9  $\mathbf{z}_v \leftarrow \mathbf{h}_v^K, \forall v \in \mathcal{V}$ 
```

---

使用固定大小的邻域函数  $\mathcal{N}(v)$  以使用固定大小的权重  $\mathbf{W}$ , 本工作使用邻域上的均匀采样. (?) 那么非均匀或者随时间变化的采样呢? ) 为了进行图上的无监督学习, 引入 graph-loss(鼓励相邻节点具有相似的学到的表示)

$$J_{\mathcal{G}}(\mathbf{z}_u) = -\log(\sigma(\mathbf{z}_u^\top \mathbf{z}_v)) - Q \cdot \mathbb{E}_{v_n \sim P_n(v)} \log(\sigma(-\mathbf{z}_u^\top \mathbf{z}_{v_n})) \quad (114)$$

这里  $\sigma$  是 sigmoid 函数,  $v$  是从  $u$  开始的固定长的随机游走序列上的节点,  $P_n$  是负样本分布.

### 12.2 Aggregator Selection

**Mean Aggregator** 和 GCN 不同, mean-aggr. 的 repr. 和上一层的表示 concat, 可以看作 skip-conn., 大幅改善了性能.

$$\mathbf{h}_v^k \leftarrow \sigma(\mathbf{W} \cdot \text{MEAN}(\{\mathbf{h}_v^{k-1}\} \cup \{\mathbf{h}_u^{k-1}, \forall u \in \mathcal{N}(v)\})) \quad (115)$$

**LSTM Aggregator** 由于 LSTM 并不是内蕴轮换不变的, 所以使用结点的随机打乱作为输入.

**Pooling Aggregator**

$$\text{AGGREGATE}_k^{\text{pool}} = \max(\{\sigma(\mathbf{W}_{\text{pool}} \mathbf{h}_{u_i}^k + \mathbf{b}), \forall u_i \in \mathcal{N}(v)\}) \quad (116)$$

注意是 MLP+max-pooling.

## 13 SGC: Simplified Graph Convolution

回顾 GCN 中的图卷积, node-wise

$$\mathbf{h}_i^{(k)} \leftarrow \frac{1}{d_i + 1} \mathbf{h}_i^{(k-1)} + \sum_{j=1}^n \frac{a_{ij}}{\sqrt{(d_i + 1)(d_j + 1)}} \mathbf{h}_j^{(k-1)} \quad (117)$$

matrix-repr.

$$\begin{aligned} \mathbf{S} &= \tilde{\mathbf{D}}^{-\frac{1}{2}} \tilde{\mathbf{A}} \tilde{\mathbf{D}}^{-\frac{1}{2}} \\ \bar{\mathbf{H}}^{(k)} &\leftarrow \mathbf{S} \mathbf{H}^{(k-1)} \end{aligned} \quad (118)$$

每一层的 feat.-trans. 和最后的分类器

$$\begin{aligned} \mathbf{H}^{(k)} &\leftarrow \text{ReLU}(\bar{\mathbf{H}}^{(k)} \Theta^{(k)}) \\ \hat{\mathbf{Y}}_{\text{GCN}} &= \text{softmax}(\mathbf{S} \mathbf{H}^{(K-1)} \Theta^{(K)}) \end{aligned} \quad (119)$$

SGC 直接在 k-hops 上聚合 (可以看作在 k-hop 连接图上聚集)

$$\hat{\mathbf{Y}}_{\text{SGC}} = \text{softmax}(\mathbf{S}^K \mathbf{X} \Theta) \quad (120)$$

这是一个凸优化问题, 可以通过二阶方法或者 SGD 来求解.

回顾 Ch

Net,

$$\hat{\mathbf{U}} \mathbf{G} \mathbf{U}^\top \mathbf{x} \approx \sum_{i=0}^k \theta_i \Delta^i \mathbf{x} = \mathbf{U} \left( \sum_{i=0}^k \theta_i \Delta^i \right) \mathbf{U}^\top \mathbf{x} \quad (121)$$

$$(122)$$

## 14 FastGCN

### 14.1 Method

回忆 GCN

$$\tilde{H}^{(l+1)} = \hat{A} H^{(l)} W^{(l)}, \quad H^{(l+1)} = \sigma(\tilde{H}^{(l+1)}), \quad l = 0, \dots, M-1, \quad L = \frac{1}{n} \sum_{i=1}^n g(H^{(M)}(i, :)) \quad (123)$$

写成泛函/积分变换的形式

$$\begin{aligned} \tilde{h}^{(l+1)}(v) &= \int \hat{A}(v, u) h^{(l)}(u) W^{(l)} dP(u), \quad h^{(l+1)}(v) = \sigma(\tilde{h}^{(l+1)}(v)), \quad l = 0, \dots, M-1 \\ L &= \mathbb{E}_{v \sim P}[g(h^{(M)}(v))] = \int g(h^{(M)}(v)) dP(v) \end{aligned} \quad (124)$$

把每个节点看作是 (连续 iid) 随机变量! 写成这种形式可以便利地使用 Monte-Carlo estimator 来估计, 每层使用  $t_l$  个采样来计算

$$\tilde{h}_{t_{l+1}}^{(l+1)}(v) := \frac{1}{t_l} \sum_{j=1}^{t_l} \hat{A}(v, u_j^{(l)}) h_{t_l}^{(l)}(u_j^{(l)}) W^{(l)}, \quad h_{t_{l+1}}^{(l+1)}(v) := \sigma(\tilde{h}_{t_{l+1}}^{(l+1)}(v)), \quad l = 0, \dots, M-1 \quad (125)$$

损失的估计 (这个估计是相容的 (以 1 概率收敛至真实值))

$$L_{t_0, t_1, \dots, t_M} := \frac{1}{t_M} \sum_{i=1}^{t_M} g\left(h_{t_M}^{(M)}\left(u_i^{(M)}\right)\right) \quad (126)$$

对于 mini-batch

$$L_{\text{batch}} = \frac{1}{t_M} \sum_{i=1}^{t_M} g\left(H^{(M)}\left(u_i^{(M)}, : \right)\right) \quad (127)$$

以及每一层的 FP

$$H^{(l+1)}(v, :) = \sigma\left(\frac{n}{t_l} \sum_{j=1}^{t_l} \hat{A}\left(v, u_j^{(l)}\right) H^{(l)}\left(u_j^{(l)}, : \right) W^{(l)}\right), \quad l = 0, \dots, M-1 \quad (128)$$

其中  $n$  是图节点数量, 作为正则化系数 (从矩阵形式到积分形式).

## 14.2 Variance Reduction

**Summary** Utilize Importance Sampling, Degree Weighted.

Use notations

	Function	Samples	Num. samples	
Layer $l+1$ ; random variable $v$	$\tilde{h}_{t_{l+1}}^{(l+1)}(v) \rightarrow y(v)$	$u_i^{(l+1)} \rightarrow v_i$	$t_{l+1} \rightarrow s$	(129)
Layer $l$ ; random variable $u$	$h_{t_l}^{(l)}(u)W^{(l)} \rightarrow x(u)$	$u_j^{(l)} \rightarrow u_j$	$t_l \rightarrow t$	

consider layer repr.

$$G := \frac{1}{s} \sum_{i=1}^s y(v_i) = \frac{1}{s} \sum_{i=1}^s \left( \frac{1}{t} \sum_{j=1}^t \hat{A}(v_i, u_j) x(u_j) \right) \quad (130)$$

compute it's variance

$$\text{Var}\{G\} = R + \frac{1}{st} \iint \hat{A}(v, u)^2 x(u)^2 dP(u) dP(v) \quad (131)$$

$$\text{where } R = \frac{1}{s} \left(1 - \frac{1}{t}\right) \int e(v)^2 dP(v) - \frac{1}{s} \left(\int e(v) dP(v)\right)^2, e(v) = \int \hat{A}(v, u) x(u) dP(u) \quad (132)$$

第一项很难再优化, 由于取决于下一层的采样. 优化第二项, 引入新的采样分布, 则为了保持  $G$  期望不变

$$y_Q(v) := \frac{1}{t} \sum_{j=1}^t \hat{A}(v, u_j) x(u_j) \left( \frac{dP(u)}{dQ(u)} \Big|_{u_j} \right), \quad u_1, \dots, u_t \sim Q \quad (133)$$

此时

$$G_Q := \frac{1}{s} \sum_{i=1}^s y_Q(v_i) = \frac{1}{s} \sum_{i=1}^s \left( \frac{1}{t} \sum_{j=1}^t \hat{A}(v_i, u_j) x(u_j) \left( \frac{dP(u)}{dQ(u)} \Big|_{u_j} \right) \right) \quad (134)$$

**Theorem** 当

$$dQ(u) = \frac{b(u)|x(u)|dP(u)}{\int b(u)|x(u)|dP(u)} \quad \text{where} \quad b(u) = \left[ \int \hat{A}(v, u)^2 dP(v) \right]^{\frac{1}{2}} \quad (135)$$

时, 方差最小, 为

$$\text{Var}\{G_Q\} = R + \frac{1}{st} \left[ \int b(u)|x(u)|dP(u) \right]^2 \quad (136)$$

然而实际上  $|x(u)|$  会变化且难以计算, 直接用 (... 那你论证半天为个啥...)

$$dQ(u) = \frac{b(u)^2 dP(u)}{\int b(u)^2 dP(u)} \quad (137)$$

MC 形式

$$q(u) = \|\hat{A}(:, u)\|^2 / \sum_{u' \in V} \|\hat{A}(:, u')\|^2, \quad u \in V \quad (138)$$

即和节点度数正比, 此时的每一层 FP 公式

$$H^{(l+1)}(v, :) = \sigma \left( \frac{1}{t_l} \sum_{j=1}^{t_l} \frac{\hat{A}(v, u_j^{(l)}) H^{(l)}(u_j^{(l)}, :) W^{(l)}}{q(u_j^{(l)})} \right), \quad u_j^{(l)} \sim q, \quad l = 0, \dots, M-1 \quad (139)$$

## 15 GWNN: Wavelet Transform on Graph

### 15.1 Supplementary Math: Real and Complex Wavelets

Function  $\psi \in L^2(\mathbb{R})$  called **orthogonal wavelet**, if it could be used to define a orthogonal complete basis of Hilbert space  $L^2(\mathbb{R})$ . Given  $\psi$ , the basis are

$$\psi_{jk}(x) = 2^{\frac{j}{2}} \psi(2^j x - k) \quad (140)$$

under normal inner-product on  $L^2(\mathbb{R})$ , it's orthogonal

$$\begin{aligned} \langle \psi_{jk}, \psi_{lm} \rangle &= \int_{-\infty}^{\infty} \psi_{jk}(x) \overline{\psi_{lm}(x)} dx \\ &= \delta_{jl} \delta_{km} \end{aligned} \quad (141)$$

#### Integral Wavelet Transform

$$[W_\psi f](a, b) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} \overline{\psi\left(\frac{x-b}{a}\right)} f(x) dx \quad (142)$$

wavelet coefficient given by

$$c_{jk} = [W_\psi f](2^{-j}, k2^{-j}) \quad (143)$$

Meyer Wavelet in frequency-domain defined

$$\Psi(\omega) := \begin{cases} \frac{1}{\sqrt{2\pi}} \sin\left(\frac{\pi}{2} \nu \left(\frac{3|\omega|}{2\pi} - 1\right)\right) e^{j\omega/2} & \text{if } 2\pi/3 < |\omega| < 4\pi/3 \\ \frac{1}{\sqrt{2\pi}} \cos\left(\frac{\pi}{2} \nu \left(\frac{3|\omega|}{4\pi} - 1\right)\right) e^{j\omega/2} & \text{if } 4\pi/3 < |\omega| < 8\pi/3 \\ 0 & \text{otherwise} \end{cases} \quad (144)$$

where (standard impl.)

$$\nu(x) := \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 < x < 1 \\ 1 & \text{if } x > 1 \end{cases} \quad (145)$$

it can also be

$$\nu(x) := \begin{cases} x^4(35 - 84x + 70x^2 - 20x^3) & \text{if } 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases} \quad (146)$$

in time-domain a close form is obtained

$$\phi(t) = \begin{cases} \frac{2}{3} + \frac{4}{3\pi} \sin\left(\frac{2\pi}{3}t\right) + \frac{4}{3}t \cos\left(\frac{4\pi}{3}t\right) & t = 0 \\ \frac{\sin\left(\frac{2\pi}{3}t\right) + \frac{4}{3}t \cos\left(\frac{4\pi}{3}t\right)}{\pi t - \frac{16\pi}{9}t^3} & \text{otherwise} \end{cases} \quad (147)$$

and  $\psi(t) = \psi_1(t) + \psi_2(t)$ ,

$$\begin{aligned} \psi_1(t) &= \frac{\frac{4}{3}\left(t-\frac{1}{2}\right)\cos\left[\frac{2\pi}{3}\left(t-\frac{1}{2}\right)\right] - \frac{1}{\pi}\sin\left[\frac{4\pi}{3}\left(t-\frac{1}{2}\right)\right]}{\left(t-\frac{1}{2}\right) - \frac{16}{9}\left(t-\frac{1}{2}\right)^3} \\ \psi_2(t) &= \frac{\frac{8}{3}\pi\left(t-\frac{1}{2}\right)\cos\left[\frac{8\pi}{3}\left(t-\frac{1}{2}\right)\right] + \frac{1}{\pi}\sin\left[\frac{4\pi}{3}\left(t-\frac{1}{2}\right)\right]}{\left(t-\frac{1}{2}\right) - \frac{64}{9}\left(t-\frac{1}{2}\right)^3} \end{aligned} \quad (148)$$

**Mexican Hat Wavelet** 1d-form Ricker Wavelet, 2nd deriv. of Gaussian dist.

$$\psi(t) = \frac{2}{\sqrt{3\sigma}\pi^{1/4}} \left(1 - \left(\frac{t}{\sigma}\right)^2\right) e^{-\frac{t^2}{2\sigma^2}} \quad (149)$$

2d-form Marr Wavelet

$$\psi(x, y) = \frac{1}{\pi\sigma^4} \left(1 - \frac{1}{2} \left(\frac{x^2 + y^2}{\sigma^2}\right)\right) e^{-\frac{x^2 + y^2}{2\sigma^2}} \quad (150)$$

**Morlet Wavelet**

$$\Psi_\sigma(t) = c_\sigma \pi^{-\frac{1}{4}} e^{-\frac{1}{2}t^2} (e^{i\sigma t} - \kappa_\sigma) \quad (151)$$

scale factor

$$c_\sigma = \left(1 + e^{-\sigma^2} - 2e^{-\frac{3}{4}\sigma^2}\right)^{-\frac{1}{2}} \quad (152)$$

## 15.2 Graph Wavelets

定义一系列图上的小波  $\psi_s = \{\psi_{si}\}$ ,  $\psi_{si}$  代表以结点  $i$  为中心, 尺度为  $s$  的小波, 数学上可以写成

$$\psi_s = \mathbf{U} \mathbf{G}_s \mathbf{U}^\top \quad (153)$$

其中  $\mathbf{U}$  是拉普拉斯矩阵的特征向量,  $\mathbf{G}_s = \text{diag}(g(s\lambda_1), \dots, g(s\lambda_n))$ ,  $g(s\lambda_i) = e^{s\lambda_i}$  (...就这? 这不是说  $\mathbf{G}_s = \exp(s\Lambda)$ ?) 图上的小波变换

$$\hat{\mathbf{x}} = \psi_s^{-1} \mathbf{x} \quad (154)$$

小波基的卷积

$$\mathbf{x} *_{\mathcal{G}} \mathbf{y} = \psi_s ((\psi_s^{-1} \mathbf{y}) \odot (\psi_s^{-1} \mathbf{x})) \quad (155)$$

### 15.3 GWNN

GWNN Layer

$$\mathbf{X}_{[:,j]}^{m+1} = h \left( \psi_s \sum_{i=1}^p \mathbf{F}_{i,j}^m \psi_s^{-1} \mathbf{X}_{[:,i]}^m \right) \quad j = 1, \dots, q \quad (156)$$

in node-wise favor

$$\mathbf{x}_j^{m+1} = h \left( \psi_s \sum_{i=1}^p \mathbf{F}_{i,j}^m \psi_s^{-1} \mathbf{x}_i^m \right) \quad j = 1, \dots, q \quad (157)$$

where  $\mathbf{F}$  is diagonal. On inductive missions(Cora etc.), 使用两层 (ReLU,softmax)GWNN. Parameters  $O(npq)$ , bad! Detach feat. trans. and graph conv.(as if GCN)

$$\mathbf{X}^{m+1} = h \left( \psi_s \mathbf{F}^m \psi_s^{-1} \mathbf{X}^m \mathbf{W} \right) \quad (158)$$

**Advantages**

1. 高效性: 小波基可以通过快速方法得到 (Chebyshev 估计,m 阶对应复杂度  $O(m|E|)$ ), 无需昂贵的 EVD).
2. 高稀疏性.
3. 局部化卷积.
4. 可变的邻域.

#### 15.3.1 Details

1.  $\mathbf{F}$  是一个对角阵 (特征向量的滤波器)
2. 只用了一个尺度 (严格地说是两个  $s, -s$ ), 核函数是 heat kernel:  $e^{-t}$
3. 可以用pygsp包的内建函数来计算 Chebyshev 系数
  - `pygsp.filters.approximations.compute_cheby_coeff(filter, order)`
  - `pygsp.filters.approximations.cheby_op(G, c, signal)`
4. 源代码里用了一个 trick, 即在  $N \times N$  单位阵上应用`cheby_op(G, c, I)`来得到  $\psi_s$  的稀疏表示. 最后还用 L1 范数归一化.
5. Shapes:  $\mathbf{X}^m, \mathbf{X}^{m'} \in \mathbb{R}^{N \times F}, \psi_s \in \mathbb{R}^{F \times N}, \psi_s^{-1} \in \mathbb{R}^{N \times F}, \mathbf{F} = \text{Diag}(\mathbf{f}) \in \mathbb{R}^{N \times N},$

## 16 Graph Wavelets

### 16.1 经典小波变换/CWT

小波

$$\psi_{s,a}(x) = \frac{1}{s} \psi \left( \frac{x-a}{s} \right) \quad (159)$$



(经典) 小波变换/CWT

$$W_f(s, a) = \int_{-\infty}^{\infty} \frac{1}{s} \psi^* \left( \frac{x-a}{s} \right) f(x) dx \quad (160)$$

可逆, 若满足 admissibility cond.

$$\int_0^{\infty} \frac{|\hat{\psi}(\omega)|^2}{\omega} d\omega = C_{\psi} < \infty \quad (161)$$

逆变换/IWT

$$f(x) = \frac{1}{C_{\psi}} \int_0^{\infty} \int_{-\infty}^{\infty} W_f(s, a) \psi_{s,a}(x) \frac{das}{s} \quad (162)$$

定义算子

$$T^s f(a) = W_f(s, a) \quad (163)$$

有

$$\bar{\psi}_s(x) = \frac{1}{s} \psi^* \left( \frac{-x}{s} \right) \quad (164)$$

则有

$$\begin{aligned} (T^s f)(a) &= \int_{-\infty}^{\infty} \frac{1}{s} \psi^* \left( \frac{x-a}{s} \right) f(x) dx = \int_{-\infty}^{\infty} \bar{\psi}_s(a-x) f(x) dx \\ &= (\bar{\psi}_s \star f)(a) \end{aligned} \quad (165)$$

频域上有

$$\widehat{T^s f}(\omega) = \hat{\bar{\psi}}_s(\omega) \hat{f}(\omega) \quad (166)$$

以及

$$\hat{\psi}_s(\omega) = \hat{\psi}^*(s\omega) \quad (167)$$

那么

$$(T^s f)(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{i\omega x} \hat{\psi}^*(s\omega) \hat{f}(\omega) d\omega \quad (168)$$

## 16.2 谱小波变换/SGWT

SGWT 核  $g \Rightarrow T_g = g(\mathcal{L})$ , 有频谱

$$\widehat{T_g f}(\ell) = g(\lambda_\ell) \hat{f}(\ell) \quad (169)$$

使用 IFT

$$(T_g f)(m) = \sum_{\ell=0}^{N-1} g(\lambda_\ell) \hat{f}(\ell) \chi_\ell(m) \quad (170)$$

局域化的图小波  $\psi_{t,n} = T_g^t \delta_n$ , 展开得

$$\psi_{t,n}(m) = \sum_{\ell=0}^{N-1} g(t\lambda_\ell) \chi_\ell^*(n) \chi_\ell(m) \quad (171)$$

小波系数

$$W_f(t, n) = (T_g^t f)(n) = \sum_{\ell=0}^{N-1} g(t\lambda_\ell) \hat{f}(\ell) \chi_\ell(n) \quad (172)$$

### 16.2.1 Scaling Functions

小波都和第一特征向量  $\chi_0$  正交, 并和特征值接近 0 的 eig-vec 几乎正交. 于是引入尺度函数, 类似地通过一个函数  $h : \mathbb{R}^+ \rightarrow \mathbb{R}$  定义, 满足  $h(0) = 0, h(\infty) = 0, \phi_n = T_h \delta_n = h(\mathcal{L}) \delta_n$ , 系数  $S_f(n) = \langle \phi_n, f \rangle$ .

将会在之后看到, 当  $G(\lambda) = h(\lambda)^2 + \sum_{j=1}^J g(t_j \lambda)^2$  有界且离开 0 时, 可以达到稳定近似.

## 16.3 SGWT 的性质

### 16.3.1 Inverse SGWT

**Lemma** 若 SGWT 核满足 admissibility cond.

$$\int_0^\infty \frac{g^2(x)}{x} dx = C_g < \infty \quad (173)$$

且  $g(0) = 0$ , 则

$$\frac{1}{C_g} \sum_{n=1}^N \int_0^\infty W_f(t, n) \psi_{t,n}(m) \frac{dt}{t} = f^\#(m) \quad (174)$$

且

$$f = f^\# + \hat{f}(0) \chi_0 \quad (175)$$

### 16.3.2 局域性

**Lemma** 定义  $d_G(m, n)$  为结点最短路径长度 (不考虑边权). 若  $d_G(m, n) > s$ ,  $(\mathcal{L}^s)_{m,n} = 0$

**Lemma** Let  $\psi_{t,n} = T_g^t \delta_n$  and  $\tilde{\psi}_{t,n} = T_{\tilde{g}}^t \delta_n$  be the wavelets at scale  $t$  generated by the kernels  $g$  and  $\tilde{g}$ . If  $|g(t\lambda) - \tilde{g}(t\lambda)| \leq M(t)$  for all  $\lambda \in [0, \lambda_{N-1}]$ , then  $|\psi_{t,n}(m) - \tilde{\psi}_{t,n}(m)| \leq M(t)$  for each vertex  $m$ . Additionally,  $\|\psi_{t,n} - \tilde{\psi}_{t,n}\|_2 \leq \sqrt{N} M(t)$

**Lemma** Let  $g$  be  $K+1$  times continuously differentiable, satisfying  $g(0) = 0, g^{(r)}(0) = 0$  for all  $r < K$ , and  $g^{(K)}(0) = C \neq 0$ . Assume that there is some  $t' > 0$  such that  $|g^{(K+1)}(\lambda)| \leq B$  for all  $\lambda \in [0, t' \lambda_{N-1}]$ . Then, for  $\tilde{g}(t\lambda) = (C/K!)(t\lambda)^K$  we have

$$M(t) = \sup_{\lambda \in [0, \lambda_{N-1}]} |g(t\lambda) - \tilde{g}(t\lambda)| \leq t^{K+1} \frac{\lambda_{N-1}^{K+1}}{(K+1)!} B$$

for all  $t < t'$

**Theorem** Let  $G$  be a weighted graph with Laplacian  $\mathcal{L}$ . Let  $g$  be a kernel satisfying the hypothesis of Lemma 5.4, with constants  $t'$  and  $B$ . Let  $m$  and  $n$  be vertices of  $G$  such that  $d_G(m, n) > K$ . Then there exist constants  $D$  and  $t''$ , such that

$$\frac{\psi_{t,n}(m)}{\|\psi_{t,n}\|} \leq Dt$$

for all  $t < \min(t', t'')$

### 16.3.3 Spectral Wavelet Frames

使用中必然使用  $J$  个  $t$  的离散采样, 导致  $NJ$  个小波和  $N$  个伸缩函数(尺度函数). 我们称一个在离散化的尺度上的小波为一个帧. 一些 Hilbert 空间上的向量组成的帧  $\Gamma_k \in \mathcal{H}$ , 不等式

$$A\|f\|^2 \leq \sum_k |\langle f, \Gamma_k \rangle|^2 \leq B\|f\|^2$$

控制了数值稳定性.

**Theorem** Given a set of scales  $\{t_j\}_{j=1}^J$ , the set  $F = \{\phi_n\}_{n=1}^N \cup \{\psi_{t_j, n}\}_{j=1}^J N_{n=1}$  forms a frame with bounds  $A, B$  given by

$$A = \min_{\lambda \in [0, \lambda_{N-1}]} G(\lambda)$$

$$B = \max_{\lambda \in [0, \lambda_{N-1}]} G(\lambda)$$

where  $G(\lambda) = h^2(\lambda) + \sum_j g(t_j \lambda)^2$

## 16.4 Fast SGWT Approximation by Polynomials

**Lemma** (多项式逼近的有效性) Let  $\lambda_{\max} \geq \lambda_{N-1}$  be any upper bound on the spectrum of  $\mathcal{L}$ . For fixed  $t > 0$ , let  $p(x)$  be a polynomial approximant of  $g(tx)$  with  $L_\infty$  error  $B = \sup_{x \in [0, \lambda_{\max}]} |g(tx) - p(x)|$ . Then the approximate wavelet coefficients  $\tilde{W}_f(t, n) = (p(\mathcal{L})f)_n$  satisfy

$$|W_f(t, n) - \tilde{W}_f(t, n)| \leq B\|f\|$$

获得这么一个估计在使用时往往需要知道一个特征值上界的估计  $\lambda_{\max}$ , 但这是个很容易的问题, 只需要做一些矩阵-向量乘法即可, 比如 Arnoldi 迭代或者 Jacobi-Davidson 算法.

使用 Chebyshev 多项式逼近: 由数值分析得知 Chebyshev 时同阶多项式逼近性能最好的.

$$T_0(\lambda) = 1, T_1(\lambda) = \lambda, \quad (176)$$

$$T_j(\lambda) = 2\lambda T_{j-1}(\lambda) - T_{j-2}(\lambda) \quad (177)$$

$$T_n(x) = \cos(n \arccos(x)) \quad (178)$$

使用变换  $x = a(y+1), a = \lambda_{\max}/2$  来把  $x$  变换到  $[-1, 1]$  上. 假设使用一系列离散化的尺度  $t_n$ , 记偏移的 CP  $\bar{T}(x) = T_k\left(\frac{x-a}{a}\right)$ , 可写

$$g(t_n x) = \frac{1}{2} c_{n,0} + \sum_{k=1}^{\infty} c_{n,k} \bar{T}_k(x) \quad (179)$$

系数

$$c_{n,k} = \frac{2}{\pi} \int_0^\pi \cos(k\theta) g(t_n(a(\cos(\theta)+1))) d\theta \quad (180)$$

(简单的数值积分即可, 计算快速). 对于任何尺度系  $t_j$ , 截断级数到  $M_j$  项来逼近核函数  $g$ , 我们



$$\tilde{W}_f(t_j, n) = \left( \frac{1}{2} c_{j,0} f + \sum_{k=1}^{M_j} c_{j,k} \bar{T}_k(\mathcal{L}) f \right)_n \quad (181)$$

$$\Rightarrow \tilde{W}_f(t_j, :) = \frac{1}{2} c_{j,0} f + \sum_{k=1}^{M_j} c_{j,k} \bar{T}_k(\mathcal{L}) f \quad (182)$$

$$\tilde{S}_f(n) = \left( \frac{1}{2} c_{0,0} f + \sum_{k=1}^{M_0} c_{0,k} \bar{T}_k(\mathcal{L}) f \right)_n \quad (183)$$

$$\Rightarrow \tilde{S}_f = \frac{1}{2} c_{0,0} f + \sum_{k=1}^{M_0} c_{0,k} \bar{T}_k(\mathcal{L}) f \quad (184)$$

$$\text{with efficient comp. of } \bar{T}_k(\mathcal{L}) f = \frac{2}{a} (\mathcal{L} - I) (\bar{T}_{k-1}(\mathcal{L}) f) - \bar{T}_{k-2}(\mathcal{L}) f \quad (185)$$

#### 16.4.1 Fast Approximation of Adjoint

可以认为  $W : \mathbb{R}^N \rightarrow \mathbb{R}^{N(J+1)}$  是一个线性变换, 且  $Wf = ((T_h f)^T, (T_g^{t_1} f)^T, \dots, (T_g^{t_J} f)^T)^T$ , 考虑其多项式估计  $\tilde{W} = ((p_0(\mathcal{L}) f)^T, (p_1(\mathcal{L}) f)^T, \dots, (p_J(\mathcal{L}) f)^T)^T$ , 这里展示其伴随算子的快速近似算法. 有

$$\begin{aligned} \langle \eta, Wf \rangle_{N(J+1)} &= \langle \eta_0, T_h f \rangle + \sum_{j=1}^J \langle \eta_j, T_g^{t_j} f \rangle_N \\ &= \langle T_h^* \eta_0, f \rangle + \left\langle \sum_{j=1}^J (T_g^{t_j})^* \eta_j, f \right\rangle_N = \left\langle T_h \eta_0 + \sum_{j=1}^J T_g^{t_j} \eta_j, f \right\rangle_N \end{aligned} \quad (186)$$

这表明

$$W^* \eta = T_h \eta_0 + \sum_{j=1}^J T_g^{t_j} \eta_j \quad (187)$$

为了计算逆变换 (伪逆  $L = (W^* W)^{-1} W^*$  是差异 norm 最小的逆变换), 计算变换和其伴随算子的乘积

$$\tilde{W}^* \tilde{W} f = \sum_{j=0}^J p_j(\mathcal{L}) (p_j(\mathcal{L}) f) = \left( \sum_{j=0}^J (p_j(\mathcal{L}))^2 \right) f \quad (188)$$

记  $P(x) = \sum_{j=0}^J (p_j(x))^2 = \frac{1}{2} d_0 + \sum_{k=1}^{M^*} d_k \bar{T}_k(x)$ ,  $M^* \max(M_j)$ , 有公式

$$T_k(x) T_l(x) = \frac{1}{2} (T_{k+l}(x) + T_{|k-l|}(x)) \quad (189)$$

设  $c'_{j,k} = c_{j,k}$  for  $k \geq 1$  and  $c'_{j,0} = \frac{1}{2} c_{j,0}$  有

$$d'_{j,k} = \begin{cases} \frac{1}{2} \left( c'_{j,0} 2 + \sum_{i=0}^{M_n} c'_{j,i} 2 \right) & \text{if } k = 0 \\ \frac{1}{2} \left( \sum_{i=0}^k c'_{j,i} c'_{j,k-i} + \sum_{i=0}^{M_j-k} c'_{j,i} c'_{j,k+i} + \sum_{i=k}^{M_j} c'_{j,i} c'_{j,i-k} \right) & \text{if } 0 < k \leq M_j \\ \frac{1}{2} \left( \sum_{i=k-M_j}^{M_j} c'_{j,i} c'_{j,k-i} \right) & \text{if } M_j < k \leq 2M_j \end{cases} \quad (190)$$



则有

$$d_{n,0} = 2d'_{j,0} \text{ and } d_{j,k} = d'_{j,k} \text{ for } k \geq 1, \text{ and setting } d_k = \sum_{j=0}^J d_{j,k} \quad (191)$$

$$\tilde{W}^* \tilde{W} f = P(\mathcal{L})f = \frac{1}{2} d_0 f + \sum_{k=1}^{M^*} d_k \bar{T}_k(\mathcal{L})f \quad (192)$$

### 16.4.2 Inverse Calculation

使用伪逆  $L = (W^* W)^{-1} W^*$ , 给定小波系数  $c$ , 可以通过方程

$$(W^* W) f = W^* c \quad (193)$$

计算原信号. 直接解是困难的, 使用快速共轭梯度法, 或者经典的帧算法 (frame algorithm).

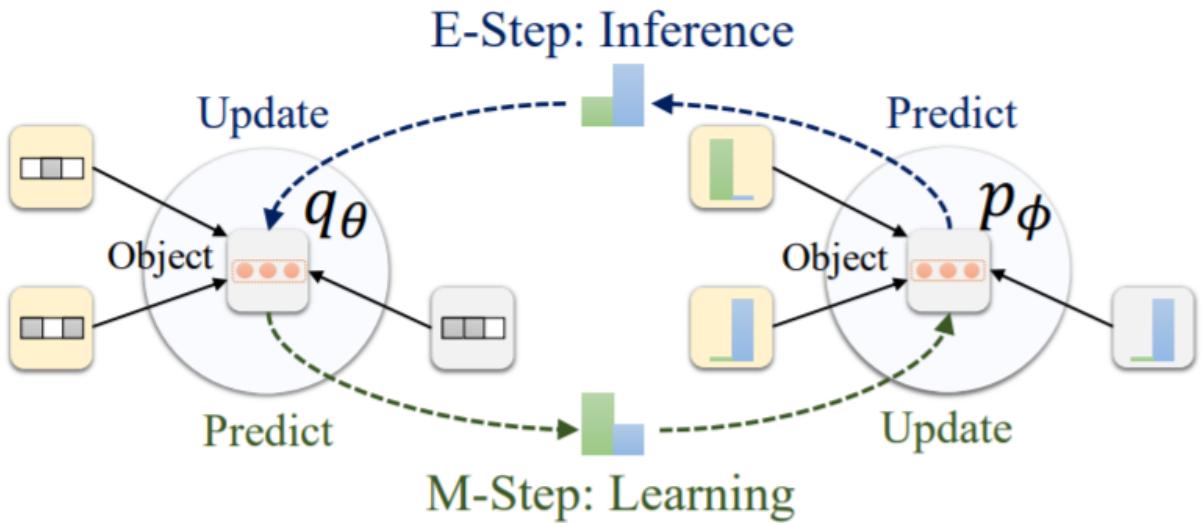
## 16.5 Implementations and Details

A good example:

$$g(x; \alpha, \beta, x_1, x_2) = \begin{cases} x_1^\alpha x^{-\alpha} & \text{for } x < x_1 \\ s(x) & \text{for } x_1 \leq x \leq x_2 \\ x_2^\beta x^{-\beta} & \text{for } x > x_2 \end{cases} \quad (194)$$

其中  $s(x)$  是三次样条. 伸缩函数  $h(x) = h(0) \exp\left(-\left(\frac{x}{0.6\lambda_{\min}}\right)^4\right)$ , 尺度按从小到大对数线性间隔选取,  $t_1 = x_2/\lambda_{\min}$ ,  $t_I = x_2/\lambda_{\max}$ ,  $\lambda_{\min} = \lambda_{\max}/K$

## 17 GMNN: Graph Markov Neural Network



**Figure 1.** Framework overview. Yellow and grey squares are labeled and unlabeled objects. Grey/white grids are attributes. Histograms are label distributions of objects. Orange triple circles are object representations. GMNN is trained by alternating between an E-step and an M-step. See Sec. 4.4 for the detailed explanation.

Idea 使用统计关系学习 (SRL) 建模

$$p(\mathbf{y}_V | \mathbf{x}_V) = \frac{1}{Z(\mathbf{x}_V)} \prod_{(n_i, n_j) \in E} \psi_{i,j}(\mathbf{y}_{n_i}, \mathbf{y}_{n_j}, \mathbf{x}_V) \quad (195)$$

而 GNN 模型则忽略 labels 之间的关系

$$p(\mathbf{y}_V | \mathbf{x}_V) = \prod_{n \in V} p(\mathbf{y}_n | \mathbf{x}_V) \quad (196)$$

具体上讲, GMNN 使用一个条件随机场 (CRF)+ 平均场近似 (mean-field approx.) 来建模, 并用 EM 算法来优化.

### 17.1 Pseudolikelihood Variational EM

优化 ELBO

$$\log p_\phi(\mathbf{y}_L | \mathbf{x}_V) \geq \mathbb{E}_{q_\theta(\mathbf{y}_U | \mathbf{x}_V)} [\log p_\phi(\mathbf{y}_L, \mathbf{y}_U | \mathbf{x}_V) - \log q_\theta(\mathbf{y}_U | \mathbf{x}_V)] \quad (197)$$

这里  $q_\theta$  是任意分布, 当

$$q_\theta(\mathbf{y}_U \mid \mathbf{x}_V) = p_\phi(\mathbf{y}_U \mid \mathbf{y}_L, \mathbf{x}_V) \quad (198)$$

取等号. 使用经典的 EM 算法来学习! 然而  $p_\phi$  中的配分函数难以计算, 使用以下 psedo-ld

$$\begin{aligned} \ell_{PL}(\phi) &\triangleq \mathbb{E}_{q_\theta(\mathbf{y}_U \mid \mathbf{x}_V)} \left[ \sum_{n \in V} \log p_\phi(\mathbf{y}_n \mid \mathbf{y}_{V \setminus n}, \mathbf{x}_V) \right] \\ &= \mathbb{E}_{q_\theta(\mathbf{y}_U \mid \mathbf{x}_V)} \left[ \sum_{n \in V} \log p_\phi(\mathbf{y}_n \mid \mathbf{y}_{\text{NB}(n)}, \mathbf{x}_V) \right] \end{aligned} \quad (199)$$

以上伪似然函数广泛应用于 Markov 学习中.

## 17.2 Inference

这一步设计计算后验分布  $p_\phi(\mathbf{y}_U \mid \mathbf{y}_L, \mathbf{x}_V)$ , 但这是困难的, 使用另一个变分分布来计算, 并使用平均场近似

$$q_\theta(\mathbf{y}_U \mid \mathbf{x}_V) = \prod_{n \in U} q_\theta(\mathbf{y}_n \mid \mathbf{x}_V) \quad (200)$$

使用一个 GNN 来参数化上述公式的每一项

$$q_\theta(\mathbf{y}_n \mid \mathbf{x}_V) = \text{MLP}[\text{Cat}(\mathbf{y}_n \mid \text{softmax}(W_\theta \mathbf{h}_{\theta,n}))] \quad (201)$$

根据平均场近似, 最优值为

$$\begin{aligned} \log q^*(\mathbf{y}_n \mid \mathbf{x}_V) &= \\ \mathbb{E}_{q_\theta(\mathbf{y}_{\text{NB}(n) \cap U} \mid \mathbf{x}_V)} [\log p_\phi(\mathbf{y}_n \mid \mathbf{y}_{\text{NB}(n)}, \mathbf{x}_V)] + \text{const.} \end{aligned} \quad (202)$$

使用 Monte-Carlo 估计

$$\begin{aligned} \mathbb{E}_{q_\theta(\mathbf{y}_{\text{NB}(n) \cap U} \mid \mathbf{x}_V)} [\log p_\phi(\mathbf{y}_n \mid \mathbf{y}_{\text{NB}(n)}, \mathbf{x}_V)] \\ \simeq \log p_\phi(\mathbf{y}_n \mid \hat{\mathbf{y}}_{\text{NB}(n)}, \mathbf{x}_V) \end{aligned} \quad (203)$$

其中  $\hat{\mathbf{y}}_{\text{NB}(n)} = \{\hat{\mathbf{y}}_{n'}\}_{n' \in \text{NB}(n)}$ , 且对于任何 unlabeled neighbors, 使用采样的标签  $\hat{\mathbf{y}}_{n'} \sim q_\theta(\mathbf{y}_{n'} \mid \mathbf{x}_V)$ , 实践中发现只取一个 (unlabeled) 样本几乎和取很多样本效果相当 (!), 效率考虑只取一个, 综上,

$$q^*(\mathbf{y}_n \mid \mathbf{x}_V) \approx p_\phi(\mathbf{y}_n \mid \hat{\mathbf{y}}_{\text{NB}(n)}, \mathbf{x}_V) \quad (204)$$

那么我们可以把后者作为 (最大化) 目标, 然后最小化 KL 散度

$$\text{KL}(p_\phi(\mathbf{y}_n \mid \hat{\mathbf{y}}_{\text{NB}(n)}, \mathbf{x}_V) \parallel q_\theta(\mathbf{y}_n \mid \mathbf{x}_V)) \quad (205)$$

进一步还是用并行更新策略, 独立的为每个 unlabeled node 优化

$$O_{\theta,U} = \sum_{n \in U} \mathbb{E}_{p_\phi(\mathbf{y}_n \mid \hat{\mathbf{y}}_{\text{NB}(n)}, \mathbf{x}_V)} [\log q_\theta(\mathbf{y}_n \mid \mathbf{x}_V)] \quad (206)$$

以及在 labeled node 上优化

$$O_{\theta,L} = \sum_{n \in L} \log q_\theta(\mathbf{y}_n \mid \mathbf{x}_V) \quad (207)$$

最终的 loss

$$O_\theta = O_{\theta,U} + O_{\theta,L} \quad (208)$$

---

## Algorithm 1 Optimization Algorithm

---

**Input:** A graph  $G$ , some labeled objects  $(L, \mathbf{y}_L)$ .

**Output:** Object labels  $\mathbf{y}_U$  for unlabeled objects  $U$ .

Pre-train  $q_\theta$  with  $\mathbf{y}_L$  according to Eq. (11).

**while** not converge **do**

**□ M-Step: Learning Procedure**

    Annotate unlabeled objects with  $q_\theta$ .

    Denote the sampled labels as  $\hat{\mathbf{y}}_U$ .

    Set  $\hat{\mathbf{y}}_V = (\mathbf{y}_L, \hat{\mathbf{y}}_U)$  and update  $p_\phi$  with Eq. (14).

**□ E-Step: Inference Procedure**

    Annotate unlabeled objects with  $p_\phi$  and  $\hat{\mathbf{y}}_V$ .

    Denote the predicted label distribution as  $p_\phi(\mathbf{y}_U)$ .

    Update  $q_\theta$  with Eq. (10), (11) based on  $p_\phi(\mathbf{y}_U), \mathbf{y}_L$ .

**end while**

Classify each unlabeled object  $n$  based on  $q_\theta(\mathbf{y}_n | \mathbf{x}_V)$ .

---

### 17.3 Learning

直接使用 GNN 来建模, 而非使用势函数

$$p_\phi(\mathbf{y}_n | \mathbf{y}_{NB(n)}, \mathbf{x}_V) = \text{Cat}(\mathbf{y}_n | \text{softmax}(W_\phi \mathbf{h}_{\phi,n})) \quad (209)$$

还可以使用 SRL 中的 techniques, 同时把  $\mathbf{y}_{NB(n)}, \mathbf{x}_{NB(n)}$  送到 GNN 中作为 in-feature. 最终的优化目标

$$O_\phi = \sum_{n \in V} \log p_\phi(\hat{\mathbf{y}}_n | \hat{\mathbf{y}}_{NB(n)}, \mathbf{x}_V) \quad (210)$$

### 17.4 Optimization

在有标签数据上预训练  $q_\theta$ , 再 EM. 最后用  $q_\theta$  来预测 (往往比用  $p_\phi$  准确率高)

	GCN [9]	Vanilla SGD	GraphSAGE [5]	FastGCN [1]	VR-GCN [2]
Time complexity	$O(L\ A\ _0 F + LNF^2)$	$O(d^L NF^2)$	$O(r^L NF^2)$	$O(rLNF^2)$	$O(L\ A\ _0 F + LNF^2 + r^L NF^2)$
Memory complexity	$O(LNF + LF^2)$	$O(bd^L F + LF^2)$	$O(br^L F + LF^2)$	$O(brLF + LF^2)$	$O(LNF + LF^2)$

## 18 ClusterGCN: Fast Deep & Large GCNs

### 18.1 Vanilla ClusterGCN: Cluster For Batch

GCN 需要整个 epoch 来更新一次梯度, 使用 mini-batch SGD 可能可以增加性能, 为此使用 batch-estimator

$$\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \nabla \text{loss}(y_i, z_i^{(L)}) \quad (211)$$

估计 loss. 但这会增加一整个 epoch 的计算时间, SGD 导致了 node-repr. 聚合了  $O(d^L)$  个邻居的信息, 导致 BP 复杂度很高. 为此定义 embedding utilization(嵌入效用), 为一个节点的表示在 BP 中被重复利用的次数. 在 GCN 中很高, 每层都为  $d$ , 但是在 GraphSAGE/FastGCN 中是一个很低的常数, 由于 k-hops 很难重叠.

为此, 考虑到一个 batch 的 emb. util. 是其中的边数  $\|A_{\mathcal{B}, \mathcal{B}}\|_0$ , 故提出想法: 每次取出边数最大的(导出)子图. 对于一个图  $G$ , 有分割

$$\bar{G} = [G_1, \dots, G_c] = [\{\mathcal{V}_1, \mathcal{E}_1\}, \dots, \{\mathcal{V}_c, \mathcal{E}_c\}] \quad (212)$$

据此, 有

$$A = \bar{A} + \Delta = \begin{bmatrix} A_{11} & \cdots & A_{1c} \\ \vdots & \ddots & \vdots \\ A_{c1} & \cdots & A_{cc} \end{bmatrix} \quad (213)$$

以及

$$\bar{A} = \begin{bmatrix} A_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & A_{cc} \end{bmatrix}, \Delta = \begin{bmatrix} 0 & \cdots & A_{1c} \\ \vdots & \ddots & \vdots \\ A_{c1} & \cdots & 0 \end{bmatrix} \quad (214)$$

令  $\bar{A}'$  是正则化后的邻接矩阵, FP 公式变得对于 cluster 分离

$$\begin{aligned} Z^{(L)} &= \bar{A}' \sigma(\bar{A}' \sigma(\cdots \sigma(\bar{A}' X W^{(0)}) W^{(1)}) \cdots) W^{(L-1)} \\ &= \begin{bmatrix} \bar{A}'_{11} \sigma(\bar{A}'_{11} \sigma(\cdots \sigma(\bar{A}'_{11} X_1 W^{(0)}) W^{(1)}) \cdots) W^{(L-1)} \\ \vdots \\ \bar{A}'_{cc} \sigma(\bar{A}'_{cc} \sigma(\cdots \sigma(\bar{A}'_{cc} X_c W^{(0)}) W^{(1)}) \cdots) W^{(L-1)} \end{bmatrix} \end{aligned} \quad (215)$$

loss 同理

$$\mathcal{L}_{\bar{A}'} = \sum_t \frac{|\mathcal{V}_t|}{N} \mathcal{L}_{\bar{A}'_{tt}}, \mathcal{L}_{\bar{A}'_{tt}} = \frac{1}{|\mathcal{V}_t|} \sum_{i \in \mathcal{V}_t} \text{loss}(y_i, z_i^{(L)}) \quad (216)$$

使用图节点聚类方法来产生分割 (Metis or Graclus), 本作中使用的是 METIS 算法<sup>3</sup>

<sup>3</sup>

## 18.2 Stochastic Multiple Partitions

以上分割算法的问题：固定地排除了一些边；并且倾向于把相似的结点放在一起，可能引入 bias。解决方案：先分割出相对大量的聚类，再随机选取一些聚类并在一起作为 batch。加快收敛。

### Algorithm 1: Cluster GCN

**Input:** Graph  $A$ , feature  $X$ , label  $Y$ ;

**Output:** Node representation  $\bar{X}$

- 1 Partition graph nodes into  $c$  clusters  $\mathcal{V}_1, \mathcal{V}_2, \dots, \mathcal{V}_c$  by METIS;
- 2 **for**  $iter = 1, \dots, max\_iter$  **do**
- 3     Randomly choose  $q$  clusters,  $t_1, \dots, t_q$  from  $\mathcal{V}$  without replacement;
- 4     Form the subgraph  $\bar{G}$  with nodes  $\bar{\mathcal{V}} = [\mathcal{V}_{t_1}, \mathcal{V}_{t_2}, \dots, \mathcal{V}_{t_q}]$  and links  $A_{\bar{\mathcal{V}}, \bar{\mathcal{V}}}$ ;
- 5     Compute  $g \leftarrow \nabla \mathcal{L}_{A_{\bar{\mathcal{V}}, \bar{\mathcal{V}}}}$  (loss on the subgraph  $A_{\bar{\mathcal{V}}, \bar{\mathcal{V}}}$ ) ;
- 6     Conduct Adam update using gradient estimator  $g$
- 7 Output:  $\{W_l\}_{l=1}^L$

## 18.3 Analysis of Deeper Networks

一个方法是增加 residual-links

$$X^{(l+1)} = \sigma(A' X^{(l)} W^{(l)}) + X^{(l)} \quad (217)$$

另一个想法是

$$X^{(l+1)} = \sigma((A' + I) X^{(l)} W^{(l)}) \quad (218)$$

(from Wikipedia) METIS is a software package for graph partitioning that implements various multilevel algorithms. METIS' multilevel approach has three phases and comes with several algorithms for each phase:

1. Coarsen the graph by generating a sequence of graphs  $G_0, G_1, \dots, G_N$ , where  $G_0$  is the original graph and for each  $0 \leq i \leq j \leq N$ , the number of vertices in  $G_i$  is greater than the number of vertices in  $G_j$ .
2. Compute a partition of  $G_N$
3. Project the partition back through the sequence in the order of  $G_N, \dots, G_0$ , refining it with respect to each graph.

The final partition computed during the third phase (the refined partition projected onto  $G_0$ ) is a partition of the original graph.



用于强调上一层的 embedding, 为了提供数值稳定性, 使用度正则化 (区别于 GCN 的对称正则化)

$$\tilde{A} = (D + I)^{-1}(A + I) \quad (219)$$

以及 FP 公式

$$X^{(l+1)} = \sigma \left( (\tilde{A} + \lambda \text{diag}(\tilde{A})) X^{(l)} W^{(l)} \right) \quad (220)$$

实验证明这提高了深层网络的性能.

## 19 GAT: Graph Attention Network

GAT 层:

1. feat. trans.  $\mathbf{h}' = \mathbf{W}\mathbf{h}$
2. atten. coeff.  $e_{ij} = a(\mathbf{h}'_i, \mathbf{h}'_j)$
3. atten. on neighbors  $\alpha_i = \text{softmax}(\mathbf{e}_i), \alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}_i} \exp(e_{ik})}$ , 本文中使用单层 MLP+concat  
feat. 作为注意力层, 则有

$$\alpha_{ij} = \frac{\exp \left( \text{LeakyReLU} \left( \vec{\mathbf{a}}^T [\mathbf{W}\vec{\mathbf{h}}_i \| \mathbf{W}\vec{\mathbf{h}}_j] \right) \right)}{\sum_{k \in \mathcal{N}_i} \exp \left( \text{LeakyReLU} \left( \vec{\mathbf{a}}^T [\mathbf{W}\vec{\mathbf{h}}_i \| \mathbf{W}\vec{\mathbf{h}}_k] \right) \right)} \quad (221)$$

4. 进一步, 使用 multi-head atten.

$$\vec{\mathbf{h}}'_i = \parallel_{k=1}^K \sigma \left( \sum_{j \in \mathcal{N}_i} \alpha_{ij}^k \mathbf{W}^k \vec{\mathbf{h}}_j \right) \quad (222)$$

最终层则使用 mean-aggr 而非 concat

$$\vec{\mathbf{h}}'_i = \sigma \left( \frac{1}{K} \sum_{k=1}^K \sum_{j \in \mathcal{N}_i} \alpha_{ij}^k \mathbf{W}^k \vec{\mathbf{h}}_j \right) \quad (223)$$

5. trick: transductive task 上使用随机采样的固定大小的邻域结点

## 20 Note on Probabilistic Graphical Models

### 20.1 Bayesian Networks

**Theorem** (d-分离的完备性) 几乎所有 (在测度意义上) 的能被 BN 表征的概率分布  $P(\text{CSDs})$  都满足: 若两节点 d-分离, 则它们条件独立.

**Theorem** (I-等价判定) 若两个 BN 有相同的骨架 (无向图基底) 和相同的 v-结构 ( $X \rightarrow Z \leftarrow Y$ )  
朴素贝叶斯, 贝叶斯网络 (一个 DAG)

## 20.2 Undirected Networks

**Definition** 一个 (或一些) r.v.  $D$  的因子是一个函数  $\phi : \text{dom}(D) \rightarrow \mathbb{R}$ . 并且定义因子的乘积,  $\phi_1 : \text{dom}((X_i) \cup (Y_j)) \rightarrow \mathbb{R}, \phi_2 : \text{dom}((Y_j) \cup (Z_k)) \rightarrow \mathbb{R}, \psi(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = \phi_1(\mathbf{X}, \mathbf{Y}) \times \phi_2(\mathbf{Y}, \mathbf{Z})$ .

**Definition** 一个被

$$\Phi = \{\phi_1(\mathbf{D}_1), \dots, \phi_K(\mathbf{D}_K)\}$$

参数化的 Gibbs 分布  $P_\Phi$  满足

$$P_\Phi(X_1, \dots, X_n) = \frac{1}{Z} \tilde{P}_\Phi(X_1, \dots, X_n) \quad (224)$$

且

$$\tilde{P}_\Phi(X_1, \dots, X_n) = \phi_1(\mathbf{D}_1) \times \phi_2(\mathbf{D}_2) \times \dots \times \phi_m(\mathbf{D}_m) \quad (225)$$

其中配分函数

$$Z = \sum_{X_1, \dots, X_n} \tilde{P}_\Phi(X_1, \dots, X_n) \quad (226)$$

**Definition** MN 的约化 (reduction)  $\mathcal{H}[\mathbf{u}]$  和  $P_\Phi[\mathbf{u}]$  定义为在变量集合  $\mathbf{U}$  上取值后的分布/图. 且他们是一一对应的.

**Definition**  $\mathbf{X}, \mathbf{Y}$  关于  $\mathbf{Z}$  分离, 若前二者之间没有不通过  $\mathbf{Z}$  的路径.

**Theorem**  $P$  是 Gibbs 分布, factorize  $MN\mathcal{H}$ , 则后者是前者的 I-map(即独立关系包含前者).

**Theorem** (Hammersley-Clifford)  $\mathcal{H}$  是 MN, 是前者结点上的正分布  $P$  的 I-map, 则  $P$  是 Gibbs 分布, 且 factorize  $MN\mathcal{H}$ .

**Theorem** 若  $\mathbf{X}, \mathbf{Y}$  关于  $\mathbf{Z}$  不分离, 那么  $\mathbf{X}, \mathbf{Y}$  关于  $\mathbf{Z}$  不独立.

类似的, 我们可以说在几乎所有分布上独立可以推出在图上分离.

**Definition**

$$\mathcal{I}_p(\mathcal{H}) = \{(X \perp Y \mid \mathcal{X} - \{X, Y\}) : X - Y \notin \mathcal{H}\} \quad (227)$$

是 pairwise-separation of  $\mathcal{H}$ ,

$$\mathcal{I}_\ell(\mathcal{H}) = \{(X \perp \mathcal{X} - \{X\} - \text{MB}_{\mathcal{H}}(X) \mid \text{MB}_{\mathcal{H}}(X)) : X \in \mathcal{X}\} \quad (228)$$

是 markov-blanket of  $\mathcal{H}$ .

**Theorem** 以下结论等价 1.  $P \models \mathcal{I}_\ell(\mathcal{H})$ . 2.  $P \models \mathcal{I}_p(\mathcal{H})$ . 3.  $P \models \mathcal{I}(\mathcal{H})$ .

**Definition**  $\phi(\mathbf{D}) = \exp(-\epsilon(\mathbf{D}))$ ,  $\epsilon(\mathbf{D})$  是能量函数.

**Definition 20.1** 一个分布  $P$  是一个 log-linear 模型, 在  $\mathcal{H}$  上, 若它和以下参数关联:

1. a set of features  $\mathcal{F} = \{f_1(\mathbf{D}_1), \dots, f_k(\mathbf{D}_k)\}$ , where each  $\mathbf{D}_i$  is a complete subgraph in  $\mathcal{H}$ ,
2. a set of weights  $w_1, \dots, w_k$  such that

$$P(X_1, \dots, X_n) = \frac{1}{Z} \exp \left[ - \sum_{i=1}^k w_i f_i(\mathbf{D}_i) \right]$$

**Example** • Ising Model: 二元 r.v.  $X_i \in \{-1, +1\}$ ,  $\epsilon_{i,j}(x_i, x_j) = w_{i,j} x_i x_j$ , 有能量函数

$$P(\xi) = \frac{1}{Z} \exp \left( - \sum_{i < j} w_{i,j} x_i x_j - \sum_i u_i x_i \right) \quad (229)$$

- Boltzmann Dist.: 二元 r.v.  $X_i \in \{0, 1\}$ , 边上的能量如同 Ising 模型, 但每个随机变量都分配了 pdf sigmoid( $z$ ),  $z = -\left(\sum_j w_{i,j} x_j\right) - w_i$
- Metric CRF: 使用 CRF 来标注图节点, 有能量函数

$$E(x_1, \dots, x_n) = \sum_i \epsilon_i(x_i) + \sum_{(i,j) \in \mathcal{E}} \epsilon_{i,j}(x_i, x_j) \quad (230)$$

其中能量函数的取法导致了不同的模型, Ising Model:

$$\epsilon_{i,j}(x_i, x_j) = \begin{cases} 0 & x_i = x_j \\ \lambda_{i,j} & x_i \neq x_j \end{cases} = \delta_{x_i, x_j} \lambda_{i,j} \quad (231)$$

Potts Model 定义了结点的度规函数  $\mu$ (需要满足非负性, 自反性, 三角不等式) 用于能量函数, 并适用于多种标签的情况. 若一个度规满足前二者 (非负性, 自反性), 则称其为 semi-metric/半度规的. CV 中常用的能量函数截断范数

$$\epsilon(x_i, x_j) = \min\left(c \|x_i - x_j\|_p, \text{dist}_{\max}\right) \quad (232)$$

**Definition 20.2** 令  $\ell(\xi) = \log P(\xi)$  Canonical energy on clique, 关于一个特定的赋值

$$\xi^* = (x_1^*, \dots, x_n^*)$$

$$\epsilon_D^*(d) = \sum_{Z \subset D} (-1)^{|D-Z|} \ell(d_Z, \xi_{-Z}^*) \quad (233)$$

**Proposition 20.3** Let  $\mathcal{B}$  be a Bayesian network over  $\mathcal{X}$  and  $\mathbf{E} = e$  an observation. Let  $\mathbf{W} = \mathcal{X} - \mathbf{E}$ . Then  $P_{\mathcal{B}}(\mathbf{W} | e)$  is a Gibbs distribution defined by the factors  $\Phi = \{\phi_{X_i}\}_{X_i \in \mathcal{X}}$ , where

$$\phi_{X_i} = P_{\mathcal{B}}(X_i | \text{Pa}_{X_i}) [\mathbf{E} = e]$$

The partition function for this Gibbs distribution is  $P(e)$

**Definition 20.4** Moralized map for BNG 定义为一个同样节点的无向图  $M[\mathcal{G}]$ , 其中一条边  $(X, Y)$  存在若在  $\mathcal{G}$  中有一条有向边连接, 或者他们是 moral 的 (具有相同的子结点).

**Proposition 20.5** For BN  $\mathcal{G}$ ,  $M[\mathcal{G}]$  是极小 I-map.

**Proposition 20.6** Let  $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$  be three disjoint sets of nodes in a Bayesian network  $\mathcal{G}$ . Let  $\mathbf{U} = \mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z}$ , and let  $\mathcal{G}' = \mathcal{G}^+[\mathbf{U}]$  be the induced Bayesian network over  $\mathbf{U} \cup$  Ancestors  $_{\mathbf{U}}$ . Let  $\mathcal{H}$  be the moralized graph  $M[\mathcal{G}']$ . Then  $d - \text{sep}_{\mathcal{G}}(\mathbf{X}; \mathbf{Y} | \mathbf{Z})$  if and only if  $\text{sep}_{\mathcal{H}}(\mathbf{X}; \mathbf{Y} | \mathbf{Z})$

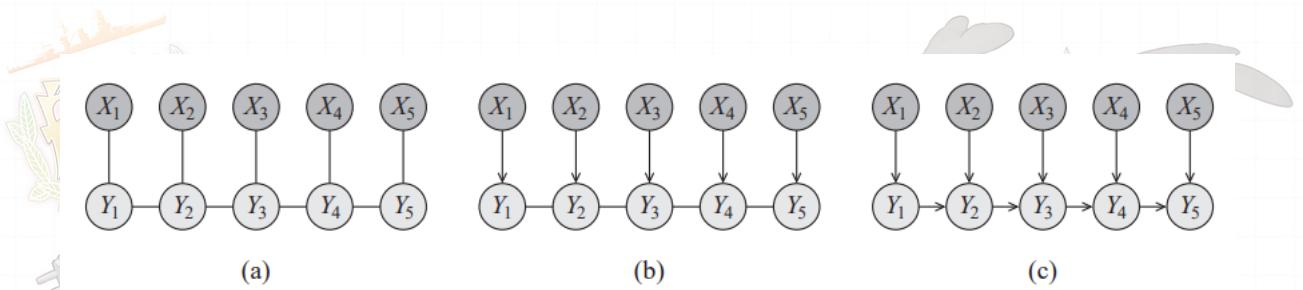
**Theorem 20.7** 若  $MN\mathcal{H}$  是弦图, 则存在 BN  $\mathcal{G}$  such that  $I(\mathcal{H}) = I(\mathcal{G})$

**Definition 20.8** CRF 是一个无向图  $\mathcal{H}$ , 节点为  $\mathbf{X} \cup \mathbf{Y}$ , 带有因子  $\phi_1(\mathbf{D}_1), \dots, \phi_m(\mathbf{D}_m)$  such that each  $\mathbf{D}_i \not\subseteq \mathbf{X}$ , models dist. such as

$$P(\mathbf{Y} | \mathbf{X}) = \frac{1}{Z(\mathbf{X})} \tilde{P}(\mathbf{Y}, \mathbf{X})$$

$$\tilde{P}(\mathbf{Y}, \mathbf{X}) = \prod_{i=1}^m \phi_i(\mathbf{D}_i)$$

$$Z(\mathbf{X}) = \sum_Y \tilde{P}(\mathbf{Y}, \mathbf{X}) \quad (234)$$



**Figure 4.14 Different linear-chain graphical models:** (a) a linear-chain-structured conditional random field, where the feature variables are denoted using grayed-out ovals; (b) a partially directed variant; (c) a fully directed, non-equivalent model. The  $X_i$ 's are assumed to be always observed when the network is used, and hence they are shown as darker gray.

**Example** 考虑只有一个  $Y$  的 CRF(朴素 Markov 模型), 能量函数

$$\phi_i(X_i, Y) = \exp\{w_i I\{X_i = 1, Y = 1\}\} \quad (235)$$

我们可以得到

$$P(Y = 1 | x_1, \dots, x_k) = \text{sigmoid}\left(w_0 + \sum_{i=1}^k w_i x_i\right) \quad (236)$$

a sigmoid-regression model!

### 20.3 Local Probabilistic Models | i.e. Specific Models Corresponds to Last 2 Sections

表式  $\Rightarrow$  复杂度极高!

确定性 CPD 由父节点的函数决定

$$P(x | \text{pa}_X) = \begin{cases} 1 & x = f(\text{pa}_X) \\ 0 & \text{otherwise} \end{cases} \quad (237)$$

树形 CPD: 类似于决策树, 但每个节点都 annotate 一个子结点上的分布

基于规则的 CPD

noisy-or CPD

$$P(y^0 | X_1, \dots, X_k) = (1 - \lambda_0) \prod_{i:X_i=x_i^0} (1 - \lambda_i) \\ P(y^1 | X_1, \dots, X_k) = 1 - \left[ (1 - \lambda_0) \prod_{i:X_i=x_i^1} (1 - \lambda_i) \right] \quad (238)$$

sigmoid CPD

$$P(y^1 | X_1, \dots, X_k) = \text{sigmoid}\left(w_0 + \sum_{i=1}^k w_i X_i\right) \quad (239)$$

Gaussian CPD

$$p(Y | x_1, \dots, x_k) = \mathcal{N}(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k; \sigma^2) \quad (240)$$

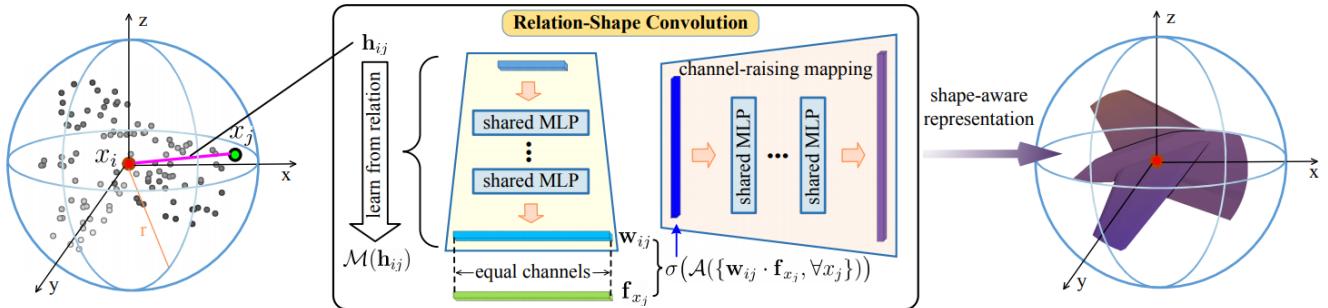


Figure 2. Overview of relation-shape convolution (RS-Conv). The key is to learn from relation. Specifically, the convolutional weight for  $x_j$  is converted to  $\mathbf{w}_{ij}$ , which learns a mapping  $\mathcal{M}$  (Eq. (2)) on predefined geometric relation vector  $\mathbf{h}_{ij}$ . In this way, the inductive convolutional representation  $\sigma(\mathcal{A}(\{\mathbf{w}_{ij} \cdot \mathbf{f}_{x_j}, \forall x_j\}))$  (Eq. (3)) can expressively reason the spatial layout of points, resulting in discriminative shape awareness. As in image CNN [34], further channel-raising mapping is conducted for a more powerful shape-aware representation.

写成向量，则为

$$p(Y | x) = \mathcal{N}(\beta_0 + \beta^T x; \sigma^2) \quad (241)$$

条件线性高斯模型 (CLG)

$$p(X | \mathbf{u}, \mathbf{y}) = \mathcal{N}\left(a_{\mathbf{u}, 0} + \sum_{i=1}^k a_{\mathbf{u}, i} y_i; \sigma_{\mathbf{u}}^2\right) \quad (242)$$

**Definition 20.9** (*conditional Bayesian networks*) 条件贝叶斯网络  $\mathcal{G}$  是一个 DAG, 节点是分离的三个集合的并  $X \cup Y \cup Z$ ,  $X$  没有父节点, 称作输入,  $Y$  称作输出, 且条件概率分布由链式法则定义

$$P_{\mathcal{B}}(Y, Z | X) = \prod_{X \in Y \cup Z} P(X | \text{Pa}_X^{\mathcal{G}}) \quad (243)$$

. 边缘分布由求和给出

$$P_{\mathcal{B}}(Y | X) = \sum_Z P_{\mathcal{B}}(Y, Z | X) \quad (244)$$

## 20.4 Temporal Models

# 21 RSCNN(CVPR 19')

## 21.1 Architecture

**Idea** 使用空间卷积/spatial conv., 在球形邻域上.

一个广义卷积

$$\mathbf{f}_{P_{\text{sub}}} = \sigma(\mathcal{A}(\{\mathcal{T}(\mathbf{f}_{x_j}), \forall x_j\})), d_{ij} < r \forall x_j \in \mathcal{N}(x_i) \quad (245)$$

要想是这个卷积 permut.-invar., 函数  $\mathcal{A}, \mathcal{T}$  必须分别是对称的和 shared.

使用 shape-aware/geometric info 函数  $\mathcal{M}$ (shared MLP 建模) 代替传统卷积

$$\mathcal{T}(\mathbf{f}_{x_j}) = \mathbf{w}_{ij} \cdot \mathbf{f}_{x_j} = \mathcal{M}(\mathbf{h}_{ij}) \cdot \mathbf{f}_{x_j} \quad (246)$$

则卷积形式变为

$$\mathbf{f}_{P_{\text{sub}}} = \sigma(\mathcal{A}(\{\mathcal{M}(\mathbf{h}_{ij}) \cdot \mathbf{f}_{x_j}, \forall x_j\})) \quad (247)$$

为了和 CNN 相对应, 使用 channel-raising MLP 来增多 channels.

最终, 这个卷积具有以下性质: permut. invar., 对于刚性变换的健壮性, shared weights, interacted point geometric.

## 21.2 Details & Implementation

使用 ReLU 激活函数, 使用 BN,  $\mathcal{M}$  使用三层 MLP, aggr. f. 为 max-pooling. Low-level 几何表示  $\mathbf{h}_{ij} = [\mathbf{x}_i - \mathbf{x}_j, \mathbf{x}_i, \mathbf{x}_j]$ , channel-raising 使用单层 MLP.

点云采样方面, 使用原点云的 FPS 采样. 使用 3-scale 邻域 (不同于 PN++ 的 MSG)

## 22 SimpleView(ICLR 21' Candidate)

### 22.1 Simple Review of Existing Protocols

数据增强: 包括抖动, y-轴随机旋转, 随即平移和缩放. ModelNet40 由于已经对齐, y-轴随机旋转会降低性能.

Model	PointNet(++)	DGCNN	RSCNN
All	随机旋转/平移	随机旋转/平移	

输入点数 PN(++) 使用 1024 个固定输入. PointCNN, RSCNN 使用每个 epoch 重采样的点.

Loss 大多数方法使用交叉熵, DGCNN 使用了平滑了的交叉熵 (label 经过平滑, 这个方法在所有结构上提高了性能)

模型选择 PN(++) 使用最终收敛的模型, DGCNN/RSCNN 使用测试集上的最好模型.

模型聚合 PN(++) 在 inference-time 把最终模型在不同旋转角度的输入上做判定 (10 次), 然后投票决定. RSCNN, DensePoint 在不同尺寸和角度的输入上判定 (300 次), 然后投票决定. DGCNN 完全没有投票.

比较性能, 本文提出的方法使用随机平移/缩放强化和 smooth-loss, 并且为了不利用任何测试集的信息, 使用 final model sel.

### 22.2 Model: SimpleView

Idea 使用多个视角的深度图像!

具体上, 使用六个 view(水平面四个, z 轴两个, 实验上这样性能最好), 并且在每张深度图上使用 ResNet18/4 骨架 (ResNet18, 滤波器数量为 1/4), concat 连接特征.

## 23 OT-Flow

### 23.1 Idea & Formulations

Formulation(based on FFJORD)

$$\partial_t \begin{bmatrix} z(\mathbf{x}, t) \\ \ell(\mathbf{x}, t) \end{bmatrix} = \begin{bmatrix} \mathbf{v}(z(\mathbf{x}, t), t; \boldsymbol{\theta}) \\ \text{tr}(\nabla \mathbf{v}(z(\mathbf{x}, t), t; \boldsymbol{\theta})) \end{bmatrix}, \quad \begin{bmatrix} z(\mathbf{x}, 0) \\ \ell(\mathbf{x}, 0) \end{bmatrix} = \begin{bmatrix} \mathbf{x} \\ 0 \end{bmatrix} \quad (248)$$

在 FFJORD 的基础

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{\rho_0(\mathbf{x})} \left\{ C(\mathbf{x}, T) := \frac{1}{2} \|z(\mathbf{x}, T)\|^2 - \ell(\mathbf{x}, T) + \frac{d}{2} \log(2\pi) \right\} \quad (249)$$

上增加最优输运代价

$$L(\mathbf{x}, T) = \int_0^T \frac{1}{2} \|\mathbf{v}(z(\mathbf{x}, t), t)\|^2 dt \quad (250)$$

满足上两个 cost 的和最小化时, 则必定存在势函数

$$\mathbf{v}(\mathbf{x}, t; \boldsymbol{\theta}) = -\nabla \Phi(\mathbf{x}, t; \boldsymbol{\theta}) \quad (251)$$

并且满足 HJB 方程 (Hamilton-Jacobi-Bellman Eq.)

$$-\partial_t \Phi(\mathbf{x}, t) + \frac{1}{2} \|\nabla \Phi(z(\mathbf{x}, t), t)\|^2 = 0, \quad \Phi(\mathbf{x}, T) = G(\mathbf{x}) \quad (252)$$

故引入惩罚项

$$R(\mathbf{x}, T) = \int_0^T \left| \partial_t \Phi(z(\mathbf{x}, t), t) - \frac{1}{2} \|\nabla \Phi(z(\mathbf{x}, t), t)\|^2 \right| dt \quad (253)$$

本工作直接不建模梯度函数  $\mathbf{v}$ , 而是直接建模势函数  $\Phi$ .

### 23.2 Parametrization of Model

势函数

$$\Phi(\mathbf{s}; \boldsymbol{\theta}) = \mathbf{w}^\top N(\mathbf{s}; \boldsymbol{\theta}_N) + \frac{1}{2} \mathbf{s}^\top (\mathbf{A}^\top \mathbf{A}) \mathbf{s} + \mathbf{b}^\top \mathbf{s} + c, \quad \text{where } \boldsymbol{\theta} = (\mathbf{w}, \boldsymbol{\theta}_N, \mathbf{A}, \mathbf{b}, c) \quad (254)$$

其中  $N$  是一个 NN(这里用的是一个简单的两层 ResNet),  $\mathbf{A} \in \mathbb{R}^{r \times (d+1)}$ , 且限制 rank  $r = \max(10, d)$  这里后面三项建模了一个二次势函数, 也即一个线性动力系统, NN 则建模了非线性部分.

ResNet 结构

$$\begin{aligned} \mathbf{u}_0 &= \sigma(\mathbf{K}_0 \mathbf{s} + \mathbf{b}_0) \\ N(\mathbf{s}; \boldsymbol{\theta}_N) &= \mathbf{u}_1 = \mathbf{u}_0 + h \sigma(\mathbf{K}_1 \mathbf{u}_0 + \mathbf{b}_1) \end{aligned} \quad (255)$$

梯度计算

$$\nabla_s \Phi(\mathbf{s}; \boldsymbol{\theta}) = \nabla_s N(\mathbf{s}; \boldsymbol{\theta}_N) \mathbf{w} + (\mathbf{A}^\top \mathbf{A}) \mathbf{s} + \mathbf{b} \quad (256)$$

Hessian Trace 计算

$$\text{tr}(\nabla^2 \Phi(\mathbf{s}; \boldsymbol{\theta})) = \text{tr}(\mathbf{E}^\top \nabla_s^2(N(\mathbf{s}; \boldsymbol{\theta}_N) \mathbf{w}) \mathbf{E}) + \text{tr}(\mathbf{E}^\top (\mathbf{A}^\top \mathbf{A}) \mathbf{E}) \quad (257)$$

后一项是 trivial 的,  $\mathbf{E}$  是  $\mathbb{R}^{(d+1)}$  标准正交基的前 d 项, ResNet 项可以得到一个闭形式

$$\begin{aligned}\text{tr}(\mathbf{E}^\top \nabla_s^2(N(\mathbf{s}; \boldsymbol{\theta}_N) \mathbf{w}) \mathbf{E}) &= t_0 + ht_1, \quad \text{where} \\ t_0 &= (\sigma''(\mathbf{K}_0 \mathbf{s} + \mathbf{b}_0) \odot \mathbf{z}_1)^\top ((\mathbf{K}_0 \mathbf{E}) \odot (\mathbf{K}_0 \mathbf{E})) \mathbf{1} \\ t_1 &= (\sigma''(\mathbf{K}_1 \mathbf{u}_0 + \mathbf{b}_1) \odot \mathbf{w})^\top ((\mathbf{K}_1 \nabla_s \mathbf{u}_0^\top) \odot (\mathbf{K}_1 \nabla_s \mathbf{u}_0^\top)) \mathbf{1}\end{aligned}\tag{258}$$

第一层的 Hessian 计算复杂度  $O(md)$ , 之后每多一层为  $O(m^2d)$ , 总复杂度则为  $O(d)$

### 23.3 Exact Hessian of Multilayer NN

Exact Trace Computation Using (13) and the same  $E$ , we compute the trace in one forward pass through the layers. The trace of the first ResNet layer is

$$\begin{aligned}t_0 &= \text{tr}(\mathbf{E}^\top \nabla_s (\mathbf{K}_0^\top \text{diag}(\sigma''(\mathbf{K}_0 \mathbf{s} + \mathbf{b}_0)) \mathbf{z}_1) \mathbf{E}) \\ &= \text{tr}(\mathbf{E}^\top \mathbf{K}_0^\top \text{diag}(\sigma''(\mathbf{K}_0 \mathbf{s} + \mathbf{b}_0) \odot \mathbf{z}_1) \mathbf{K}_0 \mathbf{E}) \\ &= (\sigma''(\mathbf{K}_0 \mathbf{s} + \mathbf{b}_0) \odot \mathbf{z}_1)^\top ((\mathbf{K}_0 \mathbf{E}) \odot (\mathbf{K}_0 \mathbf{E})) \mathbf{1}\end{aligned}$$

using the same notation as (14). For the last step, we used the diagonality of the middle matrix. Computing  $t_0$  requires  $\mathcal{O}(m \cdot d)$  FLOPS when first squaring the elements in the first  $d$  columns of  $\mathbf{K}_0$ , then summing those columns, and finally one inner product.

To compute the trace of the entire ResNet, we continue with the remaining rows in (27) in reverse order to obtain

$$\text{tr}(\mathbf{E}^\top \nabla_s^2(N(\mathbf{s}; \boldsymbol{\theta}_N) \mathbf{w}) \mathbf{E}) = t_0 + h \sum_{i=1}^M t_i$$

where  $t_i$  is computed as

$$\begin{aligned}t_i &= \text{tr}(J_{i-1}^\top \nabla_s (\mathbf{K}_i^\top \text{diag}(\sigma''(\mathbf{K}_i \mathbf{u}_{i-1}(\mathbf{s}) + \mathbf{b}_i)) \mathbf{z}_{i+1}) J_{i-1}) \\ &= \text{tr}(J_{i-1}^\top \mathbf{K}_i^\top \text{diag}(\sigma''(\mathbf{K}_i \mathbf{u}_{i-1} + \mathbf{b}_i) \odot \mathbf{z}_{i+1}) \mathbf{K}_i J_{i-1}) \\ &= (\sigma''(\mathbf{K}_i \mathbf{u}_{i-1} + \mathbf{b}_i) \odot \mathbf{z}_{i+1})^\top ((\mathbf{K}_i J_{i-1}) \odot (\mathbf{K}_i J_{i-1})) \mathbf{1}\end{aligned}$$

Here,  $J_{i-1} = \nabla_s \mathbf{u}_{i-1}^\top \in \mathbb{R}^{m \times d}$  is a Jacobian matrix, which can be updated and over-written in the forward pass at a computational cost of  $\mathcal{O}(m^2 \cdot d)$  FLOPS. The  $J$  update follows:

$$\begin{aligned}\nabla_s \mathbf{u}_i^\top &= \nabla_s \mathbf{u}_{i-1} + h \sigma'(\mathbf{K}_i \mathbf{u}_{i-1} + \mathbf{b}_i) \mathbf{K}_i^\top \nabla_s \mathbf{u}_{i-1} \\ J &\leftarrow J + h \sigma'(\mathbf{K}_i \mathbf{u}_{i-1} + \mathbf{b}_i) \mathbf{K}_i^\top J\end{aligned}$$

## 24 Node2vec: Unsupervised Feature Learning

### 24.1 Basics

MF approx. + lld optimization, 在某种采样策略  $S$  下:

$$\max_f \sum_{u \in V} \log \Pr(N_S(u) | f(u))\tag{259}$$

---

**Algorithm 1** The node2vec algorithm.

**LearnFeatures** (Graph  $G = (V, E, W)$ , Dimensions  $d$ , Walks per node  $r$ , Walk length  $l$ , Context size  $k$ , Return  $p$ , In-out  $q$ )

$\pi = \text{PreprocessModifiedWeights}(G, p, q)$

$G' = (V, E, \pi)$

Initialize  $walks$  to Empty

**for**  $iter = 1$  **to**  $r$  **do**

**for all** nodes  $u \in V$  **do**

$walk = \text{node2vecWalk}(G', u, l)$

        Append  $walk$  to  $walks$

$f = \text{StochasticGradientDescent}(k, d, walks)$

**return**  $f$

---

**node2vecWalk** (Graph  $G' = (V, E, \pi)$ , Start node  $u$ , Length  $l$ )

    Initialize  $walk$  to  $[u]$

**for**  $walk\_iter = 1$  **to**  $l$  **do**

$curr = walk[-1]$

$V_{curr} = \text{GetNeighbors}(curr, G')$

$s = \text{AliasSample}(V_{curr}, \pi)$

        Append  $s$  to  $walk$

**return**  $walk$

---

进一步使用 MF(邻域内条件独立)

$$\Pr(N_S(u) | f(u)) = \prod_{n_i \in N_S(u)} \Pr(n_i | f(u)) \quad (260)$$

特征空间对称性(点之间)给出

$$\Pr(n_i | f(u)) = \frac{\exp(f(n_i) \cdot f(u))}{\sum_{v \in V} \exp(f(v) \cdot f(u))} \quad (261)$$

最后优化目标为

$$\max_f \sum_{u \in V} \left[ -\log Z_u + \sum_{n_i \in N_S(u)} f(n_i) \cdot f(u) \right] \quad (262)$$

## 24.2 Biased Random Walk

不同与传统的BFS/DFS, 采用一种折衷的方法(二阶Markov随机游走), 设上一步为  $(t \rightarrow v)$ , 则下一步的转移概率为  $\pi_{vx} = \alpha_{pq}(t, x)$ , 其中

$$\alpha_{pq}(t, x) = \begin{cases} \frac{1}{p} & \text{if } d_{tx} = 0 \\ 1 & \text{if } d_{tx} = 1 \\ \frac{1}{q} & \text{if } d_{tx} = 2 \end{cases} \quad (263)$$

其中  $d_{uv}$  是节点最短路径.

考虑其中的参数

1. Return Param.  $p$ , 控制了回到之前的点的概率. 设为较高的值 ( $> \max(q, 1)$ ) 可以防止回到原点, 设为较低的值 ( $< \max(p, 1)$ ) 则会鼓励在原点附近探索.

2. In-out Param.  $q$ . 大于 1 的值偏向于探索原点附近的节点, 小于 1 的值偏向于 DFS 那样的远离探索.

$l$  长度的游走可以为  $k$  个节点生成  $l - k$  大小的邻域, 总时间复杂度为  $O(\frac{l}{k(l-k)})$

### 24.3 Edge Feature

简单的在节点间使用 edge feature generator 即可

Operator	Symbol	Definition	
Average	$\oplus$	$[f(u) \oplus f(v)]_i = \frac{f_i(u) + f_i(v)}{2}$	
Hadamard	$\square$	$[f(u) \square f(v)]_i = f_i(u) * f_i(v)$	(264)
Weighted-L1	$\ \cdot\ _1$	$\ f(u) \cdot f(v)\ _1 =  f_i(u) - f_i(v) $	
Weighted-L2	$\ \cdot\ _2$	$\ f(u) \cdot f(v)\ _2 =  f_i(u) - f_i(v) ^2$	

## 25 DeepWalk: Online Representation Learning

**Note** 自然语言中词语的出现 pdf 和社交图中节点在短随机游走中出现的概率都近似服从幂律分布.

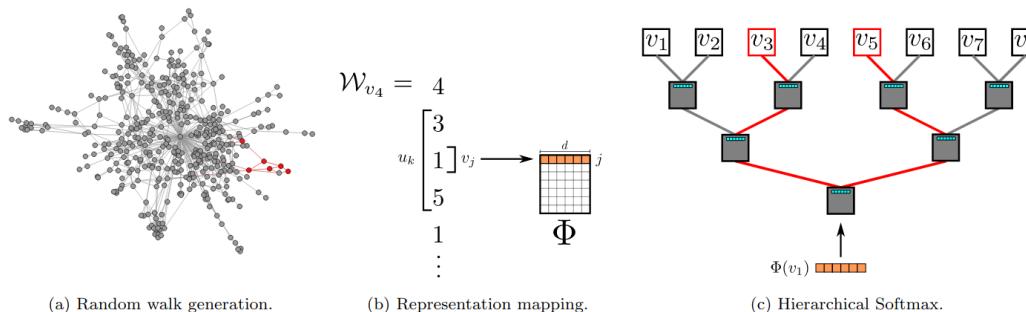


Figure 3: Overview of DEEPWALK. We slide a window of length  $2w + 1$  over the random walk  $\mathcal{W}_{v_4}$ , mapping the central vertex  $v_1$  to its representation  $\Phi(v_1)$ . Hierarchical Softmax factors out  $\Pr(v_3 \mid \Phi(v_1))$  and  $\Pr(v_5 \mid \Phi(v_1))$  over sequences of probability distributions corresponding to the paths starting at the root and ending at  $v_3$  and  $v_5$ . The representation  $\Phi$  is updated to maximize the probability of  $v_1$  co-occurring with its context  $\{v_3, v_5\}$ .

### 25.1 DeepWalk

从图中的每个节点开始 (使用一个随机生成的二叉生成树来指定顺序), 进行随机游走, 长度不定, 在邻域上均匀采样决定下一个节点, 接着使用 SkipGram 算法来更新节点表示. 每一个生成的随机游走为  $\mathcal{W}_{v_i}$ , 长度为  $t$ . 注意, 特征的形式为表式  $\Phi \in \mathbb{R}^{|V| \times d}$



---

**Algorithm 1** DEEPWALK( $G, w, d, \gamma, t$ )

---

**Input:** graph  $G(V, E)$ window size  $w$ embedding size  $d$ walks per vertex  $\gamma$ walk length  $t$ **Output:** matrix of vertex representations  $\Phi \in \mathbb{R}^{|V| \times d}$ 1: Initialization: Sample  $\Phi$  from  $\mathcal{U}^{|V| \times d}$ 2: Build a binary Tree  $T$  from  $V$ 3: **for**  $i = 0$  to  $\gamma$  **do**4:   **O** = Shuffle( $V$ )5:   **for each**  $v_i \in \mathcal{O}$  **do**6:      $\mathcal{W}_{v_i} = \text{RandomWalk}(G, v_i, t)$ 7:     SkipGram( $\Phi, \mathcal{W}_{v_i}, w$ )8:   **end for**9: **end for**

---

## 25.2 SkipGram

---

**Algorithm 2** SkipGram( $\Phi, \mathcal{W}_{v_i}, w$ )

---

1: **for each**  $v_j \in \mathcal{W}_{v_i}$  **do**2:   **for each**  $u_k \in \mathcal{W}_{v_i}[j - w : j + w]$  **do**3:      $J(\Phi) = -\log \Pr(u_k | \Phi(v_j))$ 4:      $\Phi = \Phi - \alpha * \frac{\partial J}{\partial \Phi}$ 5:   **end for**6: **end for**

---

SkipGram 是一个最大化一句话中词汇 co-occurrence 概率的算法。最大化每个窗口中的 lld  $J(\Phi) = -\log \Pr(u_k | \Phi(v_j))$ 。为了方便计算 lld，引入 Hierachical Softmax。

## 25.3 Hierachical Softmax

把每一个节点放到一个二叉树的树叶上，然后每个节点的 lld 为按照一个从根到叶子节点的路径（的乘积）

$$\Pr(u_k | \Phi(v_j)) = \prod_{l=1}^{\lceil \log |V| \rceil} \Pr(b_l | \Phi(v_j)) \quad (265)$$

中间每一层都是这样,一个节点的所有条件 lld 的计算代价为  $O(|V| \log |V|)$  还可以通过 Huffman 树来让常用的节点到根的长度更小.<sup>4</sup>

## 25.4 Parallelization

由于每个随机游走过程的 SGD 相对独立,可以并行化并使用 ASGD 来进行参数更新.

## 25.5 Variants

Streaming Learning: 在没有整个图的知识的情况下学习,此时应该不使用递减学习率(退火),可能也无法显式地建立树,如果能知道节点数的上限,则可以用那个最大值来建树.若具有对于节点出现频率的先验知识,则可以使用 Huffman 编码来建树.

Non-random Walks: 有些图是有特定的生成结构的,我们可以利用这些生成结构来指定随机游走的顺序.

# 26 DAGNN: Towards Deeper GNN

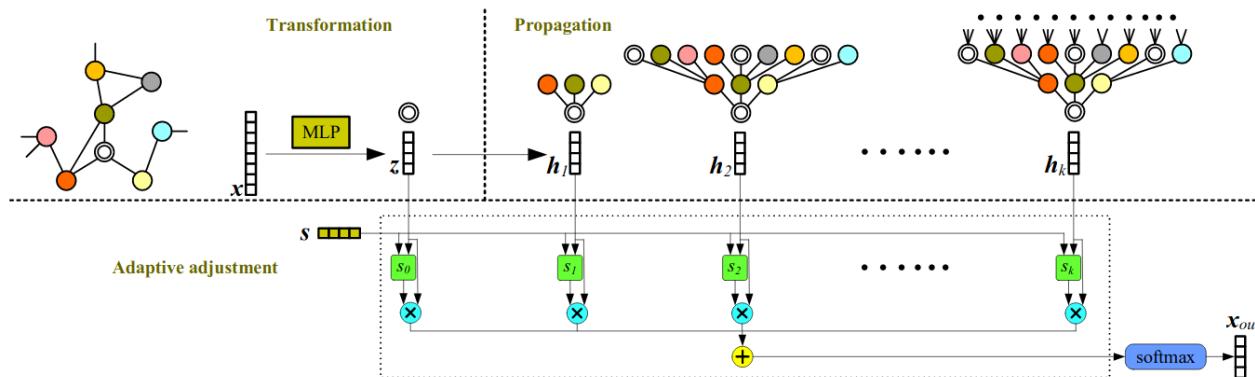


Figure 5: An illustration of the proposed Deep Adaptive Graph Neural Network (DAGNN). For clarity, we show the pipeline to generate the prediction for one node. Notation letters are consistent with Eq.(8) but bold lowercase versions are applied to denote representation vectors.  $s$  is the projection vector that computes retainment scores for representations generating from various receptive fields.  $s_0, s_1, s_2$ , and  $s_k$  represent the retainment scores of  $z, h_1, h_2$ , and  $h_k$ , respectively.

<sup>4</sup> A brief supplement from NLP: 中间节点每一项都是一个二分类器/Logistic Reg.:

$$p(d_j^w | \mathbf{x}_w, \theta_{j-1}^w) = \begin{cases} \sigma(\mathbf{x}_w^\top \theta_{j-1}^w), & d_j^w = 0 \\ 1 - \sigma(\mathbf{x}_w^\top \theta_{j-1}^w), & d_j^w = 1 \end{cases} \quad (266)$$

其中  $w$  是那个 word/节点,  $d_j^w$  是中间节点/二叉树指示编码, 写成一个式子为

$$p(d_j^w | \mathbf{x}_w, \theta_{j-1}^w) = [\sigma(\mathbf{x}_w^\top \theta_{j-1}^w)]^{1-d_j^w} \cdot [1 - \sigma(\mathbf{x}_w^\top \theta_{j-1}^w)]^{d_j^w} \quad (267)$$

lld 为

$$\begin{aligned} \mathcal{L}_w &= \log \prod_{j=2}^{l^w} \left\{ [\sigma(\mathbf{x}_w^\top \theta_{j-1}^w)]^{1-d_j^w} \cdot [1 - \sigma(\mathbf{x}_w^\top \theta_{j-1}^w)]^{d_j^w} \right\} \\ &= \sum_{j=2}^{l^w} \{(1 - d_j^w) \cdot \log [\sigma(\mathbf{x}_w^\top \theta_{j-1}^w)] + d_j^w \cdot \log [1 - \sigma(\mathbf{x}_w^\top \theta_{j-1}^w)]\} \end{aligned} \quad (268)$$

## 26.1 Smoothness Metrics

使用欧式距离为相似度度量

$$D(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{2} \left\| \frac{\mathbf{x}_i}{\|\mathbf{x}_i\|} - \frac{\mathbf{x}_j}{\|\mathbf{x}_j\|} \right\|, \quad (269)$$

点到图的平滑度

$$SMV_i = \frac{1}{n-1} \sum_{j \in V, j \neq i} D(x_i, x_j) \quad (270)$$

图的总平滑度度量

$$SMV_G = \frac{1}{n} \sum_{i \in V} SMV_i \quad (271)$$

GCN 随着层数增加, 特征的平滑度缓慢下降, 但准确度迅速下降 (数层). 这可能是由于 propag. 和特征变换的耦合导致的. 解耦了的 SGC 则在 75-100 层以后准确度/平滑度迅速下降 (over-smoothing 问题).

## 26.2 Convergence of Propagation

**Theorem 26.1** 给定图  $G$ ,  $\widehat{A}_{\oplus} = \tilde{D}^{-1} \tilde{A}$  and  $\widehat{A}_{\odot} = \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}}$ ,  $\Psi(x) = \frac{x}{\text{sum}(x)}$ ,  $\Phi(x) = \frac{x}{\|x\|}$ .

$$\lim_{k \rightarrow \infty} \widehat{A}_{\oplus}^k = \Pi_{\oplus}$$

, 其中  $\Pi_{\oplus}$  每行都是

$$\pi_{\oplus} = \Psi(e\bar{D})$$

**Theorem 26.2** 给定图  $G$ ,

$$\lim_{k \rightarrow \infty} \widehat{A}_{\odot}^k = \Pi_{\odot}$$

, 其中

$$\Pi_{\odot} = \Phi \left( \tilde{D}^{\frac{1}{2}} e^T \right) \left( \Phi \left( \tilde{D}^{\frac{1}{2}} e^T \right) \right)^T$$

这两个定理说明了, 如果使用无限层的 propagation, 会导致传递矩阵的收敛和退化, 进而导致不可分的 feat. repr./ over-smoothing. 这不可避免的是一个问题, 所以我们应该更加关心收敛速度.

## 26.3 DAGNN: Deep Adaptive GNN

DAGNN 的结构如下

$$\begin{aligned} Z &= \text{MLP}(X) && \in \mathbb{R}^{n \times c} \\ H_{\ell} &= \widehat{A}^{\ell} Z, \ell = 1, 2, \dots, k && \in \mathbb{R}^{n \times c} \\ H &= \text{stack}(Z, H_1, \dots, H_k) && \in \mathbb{R}^{n \times c} \\ S &= \sigma(Hs) && \in \mathbb{R}^{n \times (k+1) \times 1} \\ \tilde{S} &= \text{reshape}(S) && \in \mathbb{R}^{n \times 1 \times (k+1)} \\ X_{\text{out}} &= \text{softmax}(\text{squeeze}(\tilde{S}H)) && \in \mathbb{R}^{n \times (k+1) \times c} \end{aligned} \quad (272)$$

这里使用对称正规化的传播矩阵 (GCN-like),  $s \in \mathbb{R}^{c \times 1}$  是 (小型嵌入式 MLP 中的) 的可训练的投影向量 (计算出的  $\tilde{S}$  为赋予不同大小 receptive fields 的特征向量权重). DAGNN 没有 FC 层! 输出就直接为类别预测分数.

## 27 t-SNE(t-Distributed Stochastic Neighbor Embedding)

### 27.1 SNE

Target  $f : X \rightarrow Y \in R^{3or2}$ , 一个降维映射

将欧式距离变为条件概率  $p_{j|i}$ , 常用的概率如归一化的 Gaussian

$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)} \quad (273)$$

并且设置为自身相似度为 0:  $p_{i|i} = 0$

在低维度下的相似度也类似定义, 并固定方差为  $1/\sqrt{2}$

$$q_{j|i} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k \neq i} \exp\left(-\|y_i - y_k\|^2\right)} \quad (274)$$

则优化变化前后的两个分布的 KL 散度

$$C = \sum_i KL(P_i | Q_i) = \sum_{i,j} p_{j|i} \log \left\{ \frac{p_{j|i}}{q_{j|i}} \right\} \quad (275)$$

困惑度 (perplexity)

$$\text{Perp}(P_i) = 2^{H(P_i)} H(P_i) = - \sum_j p_{j|i} \log_2 p_{j|i} \quad (276)$$

用于选择方差. 使用二分搜索困惑度指标找到最优方差.

在优化开始阶段可以加入一些 Gaussian noise, 之后如同退火逐渐减少噪声幅度, 可以避免局部最优解. 无法避免 crowding 问题.

### 27.2 UNI-SNE

给低维空间一个均匀分布基准. 可通过退火逐渐减小这个基准

### 27.3 t-SNE

使用对称的联合  $\text{pdf}_{ij} = \frac{1}{2}(p_{i|j} + p_{j|i})$  来解决不对称性. 同时低维分布改为 t 分布

$$f(t) = \frac{\Gamma(\frac{v+1}{2})}{\sqrt{v\pi}\Gamma(\frac{v}{2})} \left(1 + \frac{t^2}{v}\right)^{-\frac{v+1}{2}} \quad (277)$$



$v = 1$  时为 Cauchy 分布

$$f(t) = \frac{1}{\pi(1+t^2)}$$

### Tricks

- 提前压缩: 开始初始化时点离得近些, 方便聚类中心移动. 可通过 L2 正则项的引入实现?
- 提前夸大: 开始优化阶段  $p_{ij}$  进行扩大, 避免太小导致优化太慢

### Cons

- 主要用于可视化, 难以用于特征提取.
- 倾向于保存局部特征. 对于内蕴维度 (intrinsic dim.) 较高的数据集不可能完整映射.
- 没有唯一解. 没有预估. 训练太慢 ( $O(n^2)$ ), 后续有基于树的改进.

## 27.4 Barnes-Hut-SNE

### 27.4.1 Approximating Input Similarities by Vantage-point Tree

使用一定数量的最近邻而非全部点, 来估计相似度.

$$p_{j|i} = \begin{cases} \frac{\exp(-d(\mathbf{x}_i, \mathbf{x}_j)^2 / 2\sigma_i^2)}{\sum_{k \in \mathcal{N}_i} \exp(-d(\mathbf{x}_i, \mathbf{x}_k)^2 / 2\sigma_i^2)}, & \text{if } j \in \mathcal{N}_i \\ 0, & \text{otherwise} \end{cases} \quad (278)$$
$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$$

最近邻集合的选取可以通过 VPT 来找到, 时间代价为  $O(uN \log N)$ ,  $u$  为 perplexity.

一个 VPT 中, 每个节点保存了一个数据对象和一个以其为中心的球. 所有非叶节点都有两个孩子, 左儿子保存了所有在球内部的数据对象, 右儿子则保存了所有在外的数据对象. VPT 通过一个一个遍历数据对象构建, 每次根据在外/内便利节点, 并且创建新节点, 其半径为父节点所有对象到他的距离中位数.

一次最近邻搜索可以用 VPT 上的 DFS 来实现, 计算所有节点到目标节点的距离, 维护已经找到的最近邻和到最远近邻的距离  $\tau$ .  $\tau$  决定了是否要继续探索: 若左节点里可能有比它更近的节点, 搜索左节点, 右边节点同理. 若目标节点在左节点的球中, 先搜索左节点, 右边同理.

### 27.4.2 Approximating t-SNE Gradients

有

$$\frac{\partial C}{\partial \mathbf{y}_i} = 4(F_{\text{attr}} - F_{\text{rep}}) = 4 \left( \sum_{j \neq i} p_{ij} q_{ij} Z(\mathbf{y}_i - \mathbf{y}_j) - \sum_{j \neq i} q_{ij}^2 Z(\mathbf{y}_i - \mathbf{y}_j) \right) \quad (279)$$

前半部分可以方便的算出, 后半部分可利用 Barnes-Hut 算法快速近似,  $O(N \log N)$ .

考虑两个点  $y_j, y_k$  很接近, 那么他们在  $y_i$  梯度中的贡献就很相近. BH 算法利用这一点, 在 embed. dist. 上建立一个 quad-tree 来估计总梯度.

Quad-Tree 是一个树，每个节点代表了一个矩形，非叶节点有四个子节点，代表了划分为个象限的四个矩形。叶节点包含最多一个 embedding 点。在每个节点，保存矩形的质心  $y_{cell}$ ，和总包含点数。一个  $N$  个点的 quad-tree 可以在  $O(N)$  时间构建。每个 cell 中对总梯度的贡献相似，所以

$$\sum_{j \in cell} q_{ij}^2 Z(\mathbf{y}_i - \mathbf{y}_j) \approx N_{cell} q_{i,cell}^2 Z(\mathbf{y}_i - \mathbf{y}_{cell}) \quad (280)$$

并且定义单元配分函数

$$q_{i,cell} Z = \left(1 + \|\mathbf{y}_i - \mathbf{y}_{cell}\|^2\right)^{-1} \quad (281)$$

定义 trade-off factor，衡量一个单元是否可以作为整体参与计算

$$\|\mathbf{y}_i - \mathbf{y}_{cell}\|^2 / r_{cell} < \theta \quad (282)$$

5

Dual-Tree Algorithms: 使用 cell-cell 距离来进一步减少计算。

## 28 Autoregressive Flows

Planar Flow

$$f(\mathbf{z}) = \mathbf{z} + \mathbf{u} h(\mathbf{w}^T \mathbf{z} + b) \quad (283)$$

行列式为

$$\left| \det \frac{\partial f}{\partial \mathbf{z}} \right| = \left| 1 + \mathbf{u}^T h'(\mathbf{w}^T \mathbf{z} + b) \mathbf{w} \right| \quad (284)$$

可以看做空间中的超平面，收缩或者扩张其附近的空间。

Radial Flow

$$f(\mathbf{z}) = \mathbf{z} + \beta h(\alpha, r)(\mathbf{z} - \mathbf{z}_0) \quad (285)$$

其中  $r = \|\mathbf{z} - \mathbf{z}_0\|_2$ ,  $h(\alpha, r) = \frac{1}{\alpha+r}$  类似的，可以看作超空间的一个球，收缩或者扩张其中的空间。由于这些都是比较 sparse 的变换，只影响空间的一小部分，需要很多层才能对高维空间有效。

### 28.1 Autoregressive Transformation

使用 per-dim 的 AF

$$\begin{aligned} y_1 &= \mu_1 + \sigma_1 z_1 \\ y_i &= \mu(\mathbf{y}_{1:i-1}) + \sigma(\mathbf{y}_{1:i-1}) z_i \end{aligned} \quad (286)$$

Jacobian 是下三角矩阵，行列式容易计算

$$|det J| = \left| \prod_i \sigma(\mathbf{y}_{1:i-1}) \right| \quad (287)$$

<sup>5</sup>In other words, more easy to comprehend:

$$r_{cell} > \|y_i - y_{cell}\|^2 / \theta$$

则不行

逆变换为

$$z_i = \frac{y_i - \mu(\mathbf{y}_{1:i-1})}{\sigma(\mathbf{y}_{1:i-1})} \quad (288)$$

由于无法并行计算所有维度, 必须顺序计算, 计算代价很高.

## 28.2 MAF: Masked Autoregressive Flow

MAF 用上文中的公式(286)进行变换, 这导致了他采样时极为缓慢. 在图像生成中尤为如此, 不过作为 VAE 的先验倒是可以接受 (如 1000 维).

## 28.3 IAF: Inverse Autoregressive Flow

IAF 使用(288)来进行重参数化 pdf. 使用之前的逆变换作为变换

$$y_i = z_i \sigma(\mathbf{z}_{1:i-1}) + \mu(\mathbf{z}_{1:i-1}) \quad (289)$$

此时所有的  $\sigma, \mu$  可以并行获得! i.e.

$$\mathbf{y} = \mathbf{z} \circ \sigma(\mathbf{z}) + \mu(\mathbf{z}) \quad (290)$$

假设  $\mathbf{z}_k = \mathbf{z}, \mathbf{z}_{k+1} = \mathbf{y}$ , 则有

$$\frac{\partial \mathbf{z}_k}{\partial \mathbf{z}_{k-1}} = \underbrace{\frac{\partial \mu_k}{\partial \mathbf{z}_{k-1}} + \frac{\partial \sigma_k}{\partial \mathbf{z}_{k-1}} \text{diag}(\mathbf{z}_{k-1})}_{\text{lower triangular with zeros on the diagonal}} + \text{diag}(\sigma_k) \underbrace{\frac{\partial \mathbf{z}_{k-1}}{\partial \mathbf{z}_{k-1}}}_{=\mathbf{I}} \quad (291)$$

以及行列式

$$\det\left(\frac{\partial \mathbf{z}_k}{\partial \mathbf{z}_{k-1}}\right) = \prod_{i=1}^d \sigma_{k,i} \quad (292)$$

最终 log-pdf 可以写作

$$\log q_K(\mathbf{z}_k) = \log q(\mathbf{z}) - \sum_{k=0}^K \sum_{i=1}^d \log \sigma_{k,i} \quad (293)$$

由于他的逆是 MAF, 所以从目标分布倒过来求 density 需要计算所有逆, 这就和 MAF 一样难于计算, 虽然仍是可能的.

## 28.4 Sylvester NF(UAI 18')

### 28.4.1 Idea

考虑单层 MLP 作为流

$$\mathbf{z}' = \mathbf{z} + \mathbf{A}h(\mathbf{B}\mathbf{z} + \mathbf{b}) \quad (294)$$

Jacobian 可以从 Sylvester 行列式恒等式推出. Sylvester 恒等式指出, 对于矩阵  $\mathbf{A} \in \mathbb{R}^{D \times M}, \mathbf{B} \in \mathbb{R}^{M \times D}$ , 有

$$\det(\mathbf{I}_D + \mathbf{AB}) = \det(\mathbf{I}_M + \mathbf{BA}) \quad (295)$$

据此, 上述 MLP 的 Jacobian 行列式为

$$\det\left(\frac{\partial \mathbf{z}'}{\partial \mathbf{z}}\right) = \det(\mathbf{I}_M + \text{diag}(h'(\mathbf{B}\mathbf{z} + \mathbf{b})) \mathbf{BA}) \quad (296)$$

### 28.4.2 Parametrization of A & B

一般来说,MLP 作为流不是可逆的, 提出以下特例

$$\mathbf{z}' = \mathbf{z} + \mathbf{Q}\mathbf{R}h\left(\tilde{\mathbf{R}}\mathbf{Q}^T\mathbf{z} + \mathbf{b}\right) = \phi(\mathbf{z}) \quad (297)$$

, 其中  $\mathbf{R}, \tilde{\mathbf{R}}$  是上三角矩阵,  $\mathbf{Q} = (\mathbf{q}_1 \dots \mathbf{q}_M)$  是一个正交基构成的矩阵/正交矩阵. 则 Jacobian 行列式变为

$$\begin{aligned} \det \mathbf{J} &= \det \left( \mathbf{I}_M + \text{diag} \left( h' \left( \tilde{\mathbf{R}}\mathbf{Q}^T\mathbf{z} + \mathbf{b} \right) \right) \tilde{\mathbf{R}}\mathbf{Q}^T\mathbf{Q}\mathbf{R} \right) \\ &= \det \left( \mathbf{I}_M + \text{diag} \left( h' \left( \tilde{\mathbf{R}}\mathbf{Q}^T\mathbf{z} + \mathbf{b} \right) \right) \tilde{\mathbf{R}}\mathbf{R} \right) \end{aligned} \quad (298)$$

可以在  $O(M)$  时间内计算.

给出以上流的可逆性条件

**Theorem 28.1** 以上流是可逆的, 若满足

$$r_{ii}\tilde{r}_{ii} > -\frac{1}{\|h'\|_\infty}$$

### 28.4.3 Preserving Orthogonality of Q

生成一个正交矩阵不总是可行的! 下面介绍两个显式可微地构建正交矩阵的方法, 和一个使用 permut-mat. 的方法.

**Orthogonal Sylvester Flows/O-SNF** 使用如下可微变换

$$\mathbf{Q}^{(k+1)} = \mathbf{Q}^{(k)} \left( \mathbf{I} + \frac{1}{2} (\mathbf{I} - \mathbf{Q}^{(k)\top} \mathbf{Q}^{(k)}) \right) \quad (299)$$

一个充分收敛性条件

$$\|\mathbf{Q}^{(0)\top} \mathbf{Q}^{(0)} - \mathbf{I}\|_2 < 1 \quad (300)$$

在本文实验中, 进行迭代直到

$$\|\mathbf{Q}^{(k)\top} \mathbf{Q}^{(k)} - \mathbf{I}\|_F \leq \epsilon \quad (301)$$

实验中大概会进行 30 次左右, 为了提高性能, 对所有流并行地计算这个正交化过程.

**Householder Sylvester Flows/H-SNF**

Householder reflection, with respect to  $\mathbf{v} \in \mathbb{R}^D$

$$H(\mathbf{z}) = \mathbf{z} - 2 \frac{\mathbf{v}\mathbf{v}^T}{\|\mathbf{v}\|^2} \mathbf{z} \quad (302)$$

<sup>6</sup> 具体使用的 Householder reflection 数量是一个超参数, 而且要求  $M = D$ , 因为它使用的是方阵.

**Triangular Sylvester Flows/T-SNF**

考虑为一个三角阵, 其中每个正交阵都在恒等矩阵和逆转  $z$  顺序的 permut-mat. 之间转换 (??), 这等同于在每个流之间交换  $\mathbf{R}, \tilde{\mathbf{R}}$  的上下三角性.

<sup>6</sup>In matrix sense,

$$H_{\mathbf{z}} = \mathbf{I} - \frac{2\mathbf{v}\mathbf{v}^T}{\|\mathbf{v}\|^2} \quad (303)$$

## 29 Contrastive Multi-view Representation Learning/MVRLG(ICML 20')

**Idea** 使用图上的扩散来生成增强图. 通过网络两个 view 得到不同的 repr., 让他们 contrast/作为正负样本. 来源于 CV 的对比学习 (使用 congruent/incongruent views).

可以考虑两种图增强: 初始 feature 上的增强 (如 masking/加入 Gaussian noise), 和图结构上的增强 (增减连接性, 降采样, 从最短距离或扩散矩阵生成全局 view). 前者往往会降低性能  $\Rightarrow$  使用全局图 + 降采样.

### 29.1 Augmentations

Diffusion:

$$\mathbf{S} = \sum_{k=0}^{\infty} \Theta_k \mathbf{T}^k \in \mathbb{R}^{n \times n} \quad (304)$$

, 假设  $\sum_k \theta_k = 1$ , 其中  $\mathbf{T}$  是广义转移矩阵. 在 PRR(Personal Page Rank) 和 heat kernel 算法中,

$$\mathbf{T} = \mathbf{A}\mathbf{D}^{-1}, \quad (305)$$

$$\theta_k = \alpha(1 - \alpha)^k (\text{Geometric!}), \text{ or } \theta_k = e^{-t} t^k / k! (\text{Poisson!}) \quad (306)$$

其中  $\alpha$  可以看作概率,  $t$  可以看作扩散时间. 可以得到闭形式的解:

$$\begin{aligned} \mathbf{S}^{\text{heat}} &= \exp(t\mathbf{A}\mathbf{D}^{-1} - t) \\ \mathbf{S}^{\text{PPR}} &= \alpha(\mathbf{I}_n - (1 - \alpha)\mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2})^{-1} \end{aligned} \quad (307)$$

关于降采样, 从一个 view 中采样, 并从另一个 view 中取相同的 node 和 edges.

### 29.2 Encoders

使用 GCN 作为基准 encoder, 对于两个 view 使用分别的编码器  $g_\theta, g_\omega$ , 使用邻接/扩散矩阵作为两个全等的结构视图. 其中邻接矩阵上的 GCN 使用对称正则化 ( $\tilde{\mathbf{A}} = \hat{\mathbf{D}}^{-1/2}\hat{\mathbf{A}}\hat{\mathbf{D}}^{-1/2}$ ,  $\hat{\mathbf{A}} = \mathbf{A} + \mathbf{I}_N$ ), 并且使用 PReLU 函数作为非线性. GCN 之后接着两层 MLP 得到图(节点)表示.

使用图池化提取全局特征 (readout): 通过 (JK-Net like)concat 所有 GCN 层的点特征的和, 并送到单层 MLP 中

$$\vec{h}_g = \sigma \left( \parallel_{l=1}^L \left[ \sum_{i=1}^n \vec{h}_i^{(l)} \right] \mathbf{W} \right) \in \mathbb{R}^{h_d} \quad (308)$$

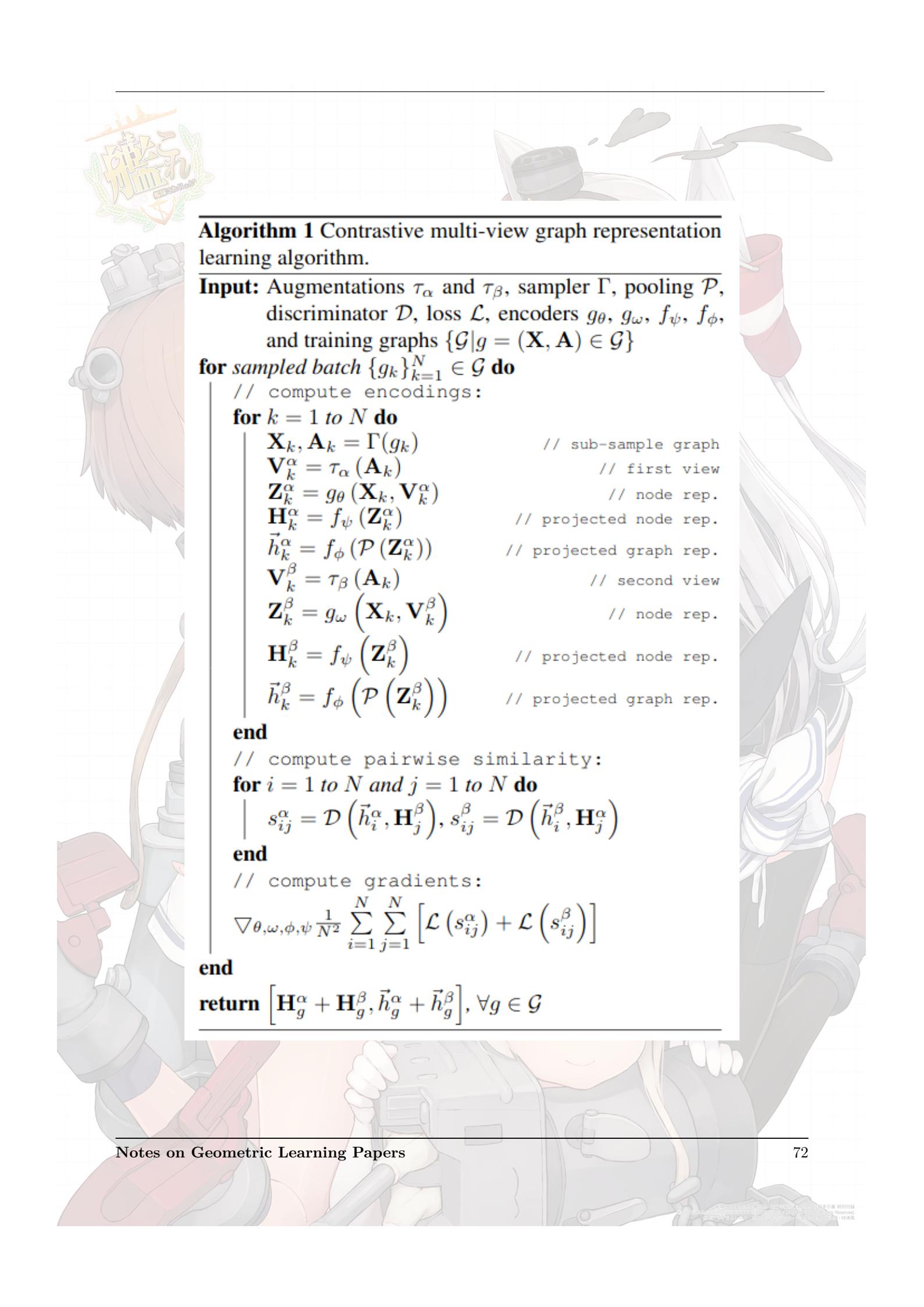
本实验中比更复杂的 readout 好 (如 DiffPool) 效果并不比这个好. 再送到共享的 proj. head, 一个双层 MLP 中.

### 29.3 Training

使用 Deep InfoMax, 并且最大化两个视图的 MI

$$\max_{\theta, \omega, \phi, \psi} \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \left[ \frac{1}{|g|} \sum_{i=1}^{|g|} \left[ \text{MI}(\vec{h}_i^\alpha, \vec{h}_g^\beta) + \text{MI}(\vec{h}_i^\beta, \vec{h}_g^\alpha) \right] \right] \quad (309)$$

---



### Algorithm 1 Contrastive multi-view graph representation learning algorithm.

**Input:** Augmentations  $\tau_\alpha$  and  $\tau_\beta$ , sampler  $\Gamma$ , pooling  $\mathcal{P}$ , discriminator  $\mathcal{D}$ , loss  $\mathcal{L}$ , encoders  $g_\theta, g_\omega, f_\psi, f_\phi$ , and training graphs  $\{\mathcal{G}|g = (\mathbf{X}, \mathbf{A}) \in \mathcal{G}\}$

**for** sampled batch  $\{g_k\}_{k=1}^N \in \mathcal{G}$  **do**

// compute encodings:

**for**  $k = 1$  to  $N$  **do**

$\mathbf{X}_k, \mathbf{A}_k = \Gamma(g_k)$  // sub-sample graph

$\mathbf{V}_k^\alpha = \tau_\alpha(\mathbf{A}_k)$  // first view

$\mathbf{Z}_k^\alpha = g_\theta(\mathbf{X}_k, \mathbf{V}_k^\alpha)$  // node rep.

$\mathbf{H}_k^\alpha = f_\psi(\mathbf{Z}_k^\alpha)$  // projected node rep.

$\vec{h}_k^\alpha = f_\phi(\mathcal{P}(\mathbf{Z}_k^\alpha))$  // projected graph rep.

$\mathbf{V}_k^\beta = \tau_\beta(\mathbf{A}_k)$  // second view

$\mathbf{Z}_k^\beta = g_\omega(\mathbf{X}_k, \mathbf{V}_k^\beta)$  // node rep.

$\mathbf{H}_k^\beta = f_\psi(\mathbf{Z}_k^\beta)$  // projected node rep.

$\vec{h}_k^\beta = f_\phi(\mathcal{P}(\mathbf{Z}_k^\beta))$  // projected graph rep.

**end**

// compute pairwise similarity:

**for**  $i = 1$  to  $N$  and  $j = 1$  to  $N$  **do**

$s_{ij}^\alpha = \mathcal{D}(\vec{h}_i^\alpha, \mathbf{H}_j^\beta), s_{ij}^\beta = \mathcal{D}(\vec{h}_i^\beta, \mathbf{H}_j^\alpha)$

**end**

// compute gradients:

$\nabla_{\theta, \omega, \phi, \psi} \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N [\mathcal{L}(s_{ij}^\alpha) + \mathcal{L}(s_{ij}^\beta)]$

**end**

**return**  $[\mathbf{H}_g^\alpha + \mathbf{H}_g^\beta, \vec{h}_g^\alpha + \vec{h}_g^\beta], \forall g \in \mathcal{G}$

MI 使用一个判别器来建模

$$\mathcal{D}(\cdot, \cdot) : \mathbb{R}^{d_h} \times \mathbb{R}^{d_h} \mapsto \mathbb{R}$$

最简单的, 可以使用内积作为相似度度量.

使用联合分布来做为正样本分布, 边缘分布的乘积作为负样本 (即取不同图的视图作为负 contrast)

## 30 GCC: Graph Contrastive Coding for GNN Pre-Training(KDD 20')

**Note** 本工作使用的是结构特征, 没有用节点特征. 用于更好的预测未见过的图: 完全的 transfer across domains.

### 30.1 GCC Pre-Training

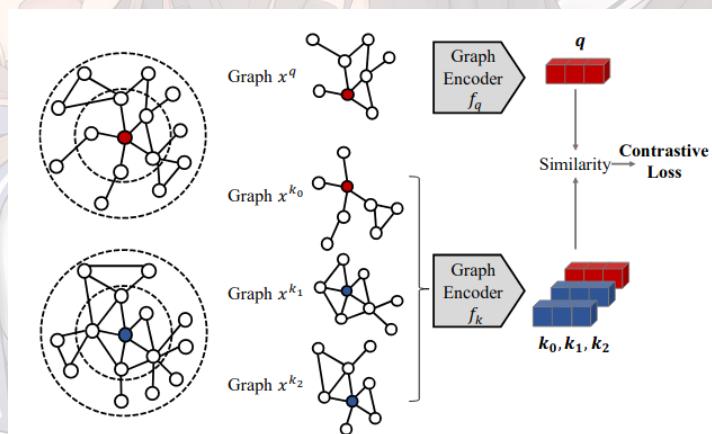


Figure 3: A running example of GCC pre-training.

**Idea** 使用 subgraph instance-discrimination + InfoNCE! 对于记录了一堆 encoded keys 的 dict $\{k_0, \dots, k_K\}$ , 和新编码的 query $q$ , contrast learning 找到一个匹配的 code $k_+$ , 可以计算 InfoNCE

$$\mathcal{L} = -\log \frac{\exp(q^\top k_+ / \tau)}{\sum_{i=0}^K \exp(q^\top k_i / \tau)} \quad (310)$$

使用子图作为对比学习中的 instance!

**Definition 30.1**  $r$ -ego network  $G_v$  是一个距离节点  $v$  最短距离  $\leq r$  的节点的诱导子图.  
(difference with  $r$ -hop?)

GCC 把不同  $r$ -ego net. 作为不相似实例.

在 CV 中, 两个经过不同 aug. 图 作为为了相似 inst. pair, GCC 中, 同样考虑同个  $r$ -ego net. 的不同数据增强作为相似实例对, 并且使用图采样作为数据增强. 具体上讲, GCC 使 步: 带重启

的随机游走, 子图诱导, 匿名化. 前两步的结果即 ISRW(Induced Subgraph Random Walk Sampling). 最后一步匿名化把节点重新标号. 最后, 两个经过如此图采样的方法被认为是相似实例对.

GCC 理论上使用任何 GNN 都可以 (作为 encoder), 本文使用 GIN(Graph Isomorphism Net.), 当前的 SOTA GNN. 由于大多数 GNN 使用节点 feature 作为输入, 使用 generalized positional embedding(\*GPE abbr.)

**Definition 30.2 GPE.** 对于每个子图, 其 GPE 为正则化 Laplacian 的主要 eig-vecs. 即, 在正则化 Laplacian 上进行 EVD:

$$I - D^{-1/2} A D^{-1/2} = U \Lambda U^\top \quad (311)$$

并取 top eig-vecs 作为 GPE. (Question: EVD 不唯一)

GPE 受到 NLP 的 Transformer 的启发. 此外还增加了 one-hot 编码的节点度, 以及 binary 的是否为中心节点的特征.

训练上, 由于维护一个字典代价很高, 所以使用一些别的方法来计算 contrast, 如 E2E(end-to-end) 和 MoCo(momentum constraint).

- E2E 使用 mini-batch, 并把 batch 内的所有 instance 作为 dictionary. Drawback: 词典大小受限与 batch 大小.
- MoCo 使用基于 momentum 更新的  $f_k$ :

$$\theta_k = \rho \theta_k + (1 - \rho) \theta_q$$

## 30.2 Finetuning GCC

- 对于 graph-level 下游任务, 使用图特征即可. 对于 node-level 下游任务, 使用 r-ego net. 的特征即可.
- 可以 freeze/full fine-tune.
- GCC 作为一个局部探索算法, 可以应用于大规模图和并行计算.

# 31 GIN: Graph Isomorphism Network(ICLR 19’)

## 31.1 Weisfeiler-Lehman Test

图同构还没有多项式时间算法. WL 算法是一个高效的近似算法, 它在每个节点上聚合邻接节点的 label, 然后 hash 为唯一的新 label. 若在某个层面上 label 不同, 则说明 graph 不同. WL 子树 kernel 把每一次聚合的 label 作为某个子树 (的数据).

### 31.2 Math Intuitions

**Lemma 31.1** 若一个 GNN 把两个非同构图映射为不同表示，则 WL 测试也把他们区分为非同构的。

这意味着所有给予聚合的 GNN 都最多和 WL 测试一样强 (在区分非同构图上)。而且，如果聚合操作和图 readout 函数都是单射，则和 WL 测试同样强。

**Theorem 31.2** Let  $\mathcal{A} : \mathcal{G} \rightarrow \mathbb{R}^d$  be a GNN. With a sufficient number of GNN layers,  $\mathcal{A}$  maps any graphs  $G_1$  and  $G_2$  that the Weisfeiler-Lehman test of isomorphism decides as non-isomorphic, to different embeddings if the following conditions hold: a)  $\mathcal{A}$  aggregates and updates node features iteratively with

$$h_v^{(k)} = \phi(h_v^{(k-1)}, f(\{h_u^{(k-1)} : u \in \mathcal{N}(v)\}))$$

where the functions  $f$ , which operates on multisets, and  $\phi$  are injective. b)  $\mathcal{A}$ 's graph-level readout, which operates on the multiset of node features  $\{h_v^{(k)}\}$ , is injective.

### 31.3 GIN

下面的引理说明了 sum-aggr 是单射，并且可以用于表示任意函数。

**Lemma 31.3** 假设  $\mathcal{X}$  是可数的，存在函数  $f : \mathcal{X} \mapsto \mathbb{R}^n$ ，使得  $h(X) = \sum_{x \in X} f(x)$  对于任何大小有界的 multi-set 唯一。进一步，任何 multiset 函数可以分解为  $g(X) = \phi(\sum_{x \in X} f(x))$ 。

然而，常用的 mean-aggr. 不是单射。

**Corollary 31.4** 假设  $\mathcal{X}$  是可数的，存在函数  $f : \mathcal{X} \mapsto \mathbb{R}^n$ ，使得存在任意多  $\epsilon$  的选择 (包括所有无理数)，使得  $h(c, X) = (1 + \epsilon)f(c) + \sum_{x \in X} f(x)$  对于任何大小有界的 multi-set 对  $(c, X)$  唯一。进一步，任何 bi-multiset 函数可以分解为  $g(c, X) = \varphi((1 + \epsilon) \cdot f(c) + \sum_{x \in X} f(x))$ 。

使用 MLP 近似函数  $f, \varphi$ ，而且直接表示他们的合成  $f^{(k+1)} \circ \varphi^{(k)}$ 。

$$h_v^{(k)} = \text{MLP}^{(k)} \left( (1 + \epsilon^{(k)}) \cdot h_v^{(k-1)} + \sum_{u \in \mathcal{N}(v)} h_u^{(k-1)} \right) \quad (312)$$

### 31.4 Graph Readout of GIN

使用每一层级上的 repr.(JK-Net like)。

$$h_G = \text{CONCAT} (\text{READOUT} (\{h_v^{(k)} \mid v \in G\}) \mid k = 0, 1, \dots, K) \quad (313)$$

根据已有定理，如果 READOUT 函数是求和同一层中的 feature，则确实推广了 WL。

---

**Algorithm 1** Training procedure of Local-instance and Global-semantic Learning (GraphLoG).

**Input:** Training set  $D = \{\mathcal{G}_j\}_{j=1}^{N_D}$ , the number of training iterations  $N_T$ , hierarchical prototypes' depth  $L_p$  and exponential decay rate  $\beta$ .  
**Output:** The pre-trained GNN.

```

Initialize hierarchical prototypes  $\{c_i^l\}_{i=1}^{M_l}$  ( $l = 1, 2, \dots, L_p$ )
for  $t = 1$  to  $N_T$  do
     $B_G \leftarrow \text{RandomSample}(D)$                                 # Get a mini-batch of graphs
     $B'_G \leftarrow \text{AttrMasking}(B_G)$                             # Get the correlated graphs
     $h_{\mathcal{V}_j}, h_{\mathcal{V}'_j}, h_{\mathcal{G}_j}, h_{\mathcal{G}'_j} \leftarrow \text{Eqs. (4, 5)}$  ( $j = 1, 2, \dots, N$ ) # Extract patch and graph embeddings
     $\mathcal{L}_{\text{local}}, \mathcal{L}_{\text{global}} \leftarrow \text{Eqs. (8, 13)}$           # Compute losses
     $\theta_{\text{GNN}} \leftarrow -\nabla_{\theta_{\text{GNN}}} (\mathcal{L}_{\text{local}} + \mathcal{L}_{\text{global}})$  # Update GNN's parameters
     $\theta_T \leftarrow -\nabla_{\theta_T} (\mathcal{L}_{\text{local}} + \mathcal{L}_{\text{global}})$           # Update discriminator's parameters
     $\{c_i^l\}_{i=1}^{M_l} \leftarrow \text{Eqs. (11, 12)}$  ( $l = 1, 2, \dots, L_p$ )      # Maintain hierarchical prototypes
end for
```

---

## 32 GraphLoG: Self-Supervised Representation Learning with Local & Global Structure

### 32.1 Preliminaries

GNN and READOUT GNN-layer

$$h_v^{(l)} = \text{COMBINE}^{(l)} \left( h_v^{(l-1)}, \text{AGGREGATE}^{(l)} \left( \{(h_u^{(l-1)}, h_u^{(l-1)}, X_{uv}) : u \in \mathcal{N}(v)\} \right) \right) \quad (314)$$

Graph Readout:

$$h_G = \text{READOUT}(\{h_v \mid v \in \mathcal{V}\}) \quad (315)$$

Mutual Info. Est. InfoNCE

$$\mathcal{L}_{\text{NCE}}(q, z_+, \{z_i\}_{i=1}^K) = -\log \frac{\exp(T(q, z_+))}{\exp(T(q, z_+)) + \sum_{i=1}^K \exp(T(q, z_i))} \quad (316)$$

其中  $T$  是某个参数化的判别函数.

**RPCL: Rival Penalized Competitive Learning** RPCL 在每个新样本上不仅推近 winning cluster(最近聚类), 也推开 rival cluster(次近聚类). 可以不预先指定聚类数地进行 clustering.

### 32.2 Local-Inst. Stru. Learning

使用 mask 来获得相近 (correlated) 图. 对于每个 mini-batch 获得相近图 (通过 mask). 得到他们的 node(patch)/graph repr.

$$\begin{aligned} h_{\mathcal{V}_j} &= \{h_v \mid v \in \mathcal{V}_j\} = \text{GNN}(X_{\mathcal{V}_j}, X_{\mathcal{E}_j}), \quad h_{\mathcal{V}'_j} = \{h_v \mid v \in \mathcal{V}'_j\} = \text{GNN}(X_{\mathcal{V}'_j}, X_{\mathcal{E}'_j}) \\ h_{\mathcal{G}_j} &= \text{READOUT}(h_{\mathcal{V}_j}), \quad h_{\mathcal{G}'_j} = \text{READOUT}(h_{\mathcal{V}'_j}) \end{aligned} \quad (317)$$

最小化 InfoNCE<sup>7</sup>

$$\begin{aligned}\mathcal{L}_{\text{patch}} &= \frac{1}{\sum_{j=1}^N |\mathcal{V}'_j|} \sum_{j=1}^N \sum_{v' \in \mathcal{V}'_j} \sum_{v \in \mathcal{V}_j} \mathbb{1}_{v \leftrightarrow v'} \cdot \mathcal{L}_{\text{NCE}}(h_{v'}, h_v, \{h_{\tilde{v}} \mid \tilde{v} \in \mathcal{V}_j, \tilde{v} \neq v\}) \\ \mathcal{L}_{\text{graph}} &= \frac{1}{N} \sum_{j=1}^N \mathcal{L}_{\text{NCE}}(h_{\mathcal{G}'_j}, h_{\mathcal{G}_j}, \{h_{\mathcal{G}_k} \mid 1 \leq k \leq N, k \neq j\}) \\ \mathcal{L}_{\text{local}} &= \mathcal{L}_{\text{patch}} + \mathcal{L}_{\text{graph}}\end{aligned}\quad (318)$$

**Note** 同个 batch 中的其他图 repr. 作为 negative instance. 同个图中其他 node repr. 作为 negative instance.

### 32.3 Global-Semantic Repr. Learning

Graph(如分子图) 常常有 hierarchical structure  $\Rightarrow$  *hierarchical prototypes*, 建模图嵌入的分布. 他们是一些树的集合, 每个节点对应了一个 prototype, 并且指向唯一的父节点 (除非是根节点). 正式地, 这些节点是  $\{c_i^l\}_{i=1}^{M_l}$ , ( $l = 1, 2, \dots, L_p$ ), 除了树叶节点, 每个节点都有一些子节点集合  $C(c_i^l)$ .

#### 32.3.1 Init. of HP(Hierarchical Prototypes)

先用  $\mathcal{L}_{\text{local}}$  预训练一个 epoch, 利用这个 GNN 得到所有训练集中图的表示  $\{h_{\mathcal{G}_i}\}_{i=1}^{N_D}$ , 作为 HP 的叶子节点, 并且使用 RPCL 得到底层的节点

$$\left\{c_i^{L_p}\right\}_{i=1}^{M_{L_p}} = \text{RPCL}\left(\{h_{\mathcal{G}_i}\}_{i=1}^{N_D}\right) \quad (319)$$

之后 RPCL 被迭代的应用来得到整个 HP 树.

#### 32.3.2 Maintainance of HP

训练过程中, 图嵌入在动态改变. 使用一下策略来更新 HP:

- 每有一个 batch 训练完, 得到图 embedding, 分成  $M_{L_p}$  个组 (属于 HP 最底层聚类), 计算每一组的平均 embedding. 并且更新节点上的 embedding, 使用 (momentum-like) 指数移动平均更新模型

$$c_i^{L_p} \leftarrow \beta c_i^{L_p} + (1 - \beta) \hat{c}_i^{L_p}, \quad 1 \leq i \leq M_{L_p} \quad (320)$$

此处  $\beta$  是衰减常数, 之后上层节点 embedding 更新为子节点的平均.

2.

为了捕捉 global-semantic structure, 让相关图属于同一聚类! 具体上讲, 对于每一个图  $\mathcal{G}_j$ , 使用 cosine sim., 在每一层上寻找最接近的 embedding

$$s(\mathcal{G}_j) = \{s_1(\mathcal{G}_j), s_2(\mathcal{G}_j), \dots, s_{L_p}(\mathcal{G}_j)\}$$

这里要让相关图的 embedding 和这些 repr. 相近, 使用 InfoNCE

$$\mathcal{L}_{\text{global}} = \frac{1}{N \cdot L_p} \sum_{j=1}^N \sum_{l=1}^{L_p} \mathcal{L}_{\text{NCE}}(h_{\mathcal{G}'_j}, s_l(\mathcal{G}_j), \{c_i^l \mid 1 \leq i \leq M_l, c_i^l \neq s_l(\mathcal{G}_j)\}) \quad (321)$$

<sup>7</sup>如同其他 contrast learning

即, 对于每一层, 正样本是那个 (原图) 最相近的 embedding, 负样本是 (原图) 其他样本 repr.  
模型在每一个 iteration 上最小化全 loss

$$\min_{\text{GNN}, T} \mathcal{L}_{\text{local}} + \mathcal{L}_{\text{global}} \quad (322)$$

### 32.4 Sup-GraphLoG: A Supervised Baseline

使用 label 来简单监督学习, 作为 baseline. 对于任何一个图, 找到在 HP 底层匹配的 (随机的) 一个匹配的 prototype, 把搜索路径作为正样本, 再随机选择另一条不正确的路径作为负样本. 然后 mini-batch 上的全局 loss 为

$$\mathcal{L}_{\text{global}}^{\text{sup}} = \frac{1}{N \cdot L_p} \sum_{j=1}^N \sum_{l=1}^{L_p} \mathcal{L}_{\text{NCE}}(h_{\mathcal{G}_j}, s_l(\mathcal{G}_j), s_l^n(\mathcal{G}_j)) \quad (323)$$

## 33 Orthogonal Weights in DNNs

### 33.1 Formulation & Good Properties

优化参数, 使得权重是 (伪) 正交的

$$\begin{aligned} \theta^* = \arg \min_{\theta} \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \in D} [\mathcal{L}(\mathbf{y}, f(\mathbf{x}; \theta))] \\ \text{s.t. } \mathbf{W}^l \in \mathcal{O}_l^{n_l \times d_l}, l = 1, 2, \dots, L \end{aligned} \quad (324)$$

此处伪正交矩阵在一个实 Stiefel 流形  $\mathcal{O}_l^{n_l \times d_l} = \left\{ \mathbf{W}^l \in \mathbb{R}^{n_l \times d_l} : \mathbf{W}^l (\mathbf{W}^l)^T = \mathbf{I} \right\}$  上.

$\Rightarrow$  OMDSM(Optim. on Multiple Dependent Stiefel Manifolds) Problem

好处

- $s = \mathbf{Wx}$ ,  $\mathbf{W} \in \mathbb{R}^{n \times d}$ , 若输入  $\mathbf{x}$  是白化了的, 则输出  $s$  也是 (0 均值且分量不相关); 如果  $n = d$ , 则  $\|s\| = \|\mathbf{x}\|$ ; 梯度相等  $\left\| \frac{\partial \mathcal{L}}{\partial \mathbf{x}} \right\| = \left\| \frac{\partial \mathcal{L}}{\partial s} \right\|$
- 自动地, 这正则化了权重. 减少了自由度 (Stiefel Manifold 的维度减少了):  
 $\dim \mathcal{O}^{n \times d} = nd - n(n+1)/2$

### 33.2 OWN: Orthogonal Weight Normalization

使用 Riemann(流形) 优化算法 (like in RNNs<sup>8</sup>) 导致了收敛的不稳定性/性能差. 显式地使用参数矩阵的重参数化

$$\phi : \mathbb{R}^{n_l \times d_l} \rightarrow \mathbb{R}^{n_l \times d_l} \quad (325)$$

并且  $\phi(\mathbf{V}) = \mathbf{W}$ , s.t.  $\mathbf{W}\mathbf{W}^T = \mathbf{I}$ .

<sup>8</sup>What's that actually?



使用一个线性变换来作为函数

$$\phi(\mathbf{V}) = \mathbf{P}\mathbf{V}_C \quad (326)$$

$$\mathbf{V}_C = \mathbf{V} - \mathbf{c}\mathbf{1}_d^T \quad (327)$$

$$\mathbf{c} = \frac{1}{d}\mathbf{V}\mathbf{1}_d \quad (328)$$

$$\text{which means } \mathbf{V}_C = \mathbf{V} - \frac{1}{d}\mathbf{V}\mathbf{E}_d = \mathbf{V}(\mathbf{I} - \frac{1}{d}\mathbf{E}_d) \quad (329)$$

使用以下变换来得到权重, 使得 Jacobian 的奇异值接近于 1

$$\begin{aligned} & \min_{\mathbf{P}} \text{tr} \left( (\mathbf{W} - \mathbf{V}_C)(\mathbf{W} - \mathbf{V}_C)^T \right) \\ & \text{s.t. } \mathbf{W} = \mathbf{P}\mathbf{V}_C \text{ and } \mathbf{W}\mathbf{W}^T = \mathbf{I} \end{aligned} \quad (330)$$

进而, 这个优化问题有 closed form sol.

$$\mathbf{P}^* = \mathbf{D}\Lambda^{-1/2}\mathbf{D}^T \quad (331)$$

其中  $\mathbf{V}_C\mathbf{V}_C^T = \Sigma = \mathbf{D}\Lambda\mathbf{D}^T$  是协方差矩阵的 EVD(SVD). 最后的形式为

$$\mathbf{W} = \phi(\mathbf{V}) = \mathbf{D}\Lambda^{-1/2}\mathbf{D}^T (\mathbf{V} - \mathbf{c}\mathbf{1}_d^T) \quad (332)$$

类似但没有最小化(330)的选择  $\mathbf{P}_{var} = \Lambda^{-1/2}\mathbf{D}^T$

### 33.2.1 Backpropagation

计算 Jacobian

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \Lambda} &= -\frac{1}{2}\mathbf{D}^T \frac{\partial \mathcal{L}}{\partial \mathbf{W}} \mathbf{W}^T \mathbf{D}\Lambda^{-1} \\ \frac{\partial \mathcal{L}}{\partial \mathbf{D}} &= \mathbf{D}\Lambda^{\frac{1}{2}}\mathbf{D}^T \mathbf{W} \frac{\partial \mathcal{L}}{\partial \mathbf{W}} \mathbf{D}\Lambda^{-\frac{1}{2}} + \frac{\partial \mathcal{L}}{\partial \mathbf{W}} \mathbf{W}^T \mathbf{D} \\ \frac{\partial \mathcal{L}}{\partial \mathbf{S}} &= \mathbf{D} \left( \left( \mathbf{K}^T \odot \left( \mathbf{D}^T \frac{\partial \mathcal{L}}{\partial \mathbf{D}} \right) \right) + \left( \frac{\partial \mathcal{L}}{\partial \Lambda} \right)_{\text{diag}} \right) \mathbf{D}^T \\ \frac{\partial \mathcal{L}}{\partial \mathbf{c}} &= -\mathbf{1}_d^T \frac{\partial \mathcal{L}}{\partial \mathbf{W}} \mathbf{D}\Lambda^{-\frac{1}{2}}\mathbf{D}^T - 2 \cdot \mathbf{1}_d^T (\mathbf{V} - \mathbf{c}\mathbf{1}_d^T)^T \left( \frac{\partial \mathcal{L}}{\partial \Sigma} \right)_s \\ \frac{\partial \mathcal{L}}{\partial \mathbf{V}} &= \mathbf{D}\Lambda^{-\frac{1}{2}}\mathbf{D}^T \frac{\partial \mathcal{L}}{\partial \mathbf{W}} + 2 \left( \frac{\partial \mathcal{L}}{\partial \Sigma} \right)_s (\mathbf{V} - \mathbf{c}\mathbf{1}_d^T) + \frac{1}{d} \frac{\partial \mathcal{L}^T}{\partial \mathbf{c}} \mathbf{1}_d^T \end{aligned} \quad (333)$$

此处

$$\mathbf{K}_{ij} = \frac{1}{\sigma_i - \sigma_j} [i \neq j] \quad (334)$$

是一个对角线为 0 的矩阵.

### 33.2.2 As Convolution

卷积层的参数  $\mathbf{W}^C \in \mathbb{R}^{n \times d \times F_h \times F_w}$ , 以及上一层的输入特征  $\mathbf{X} \in \mathbb{R}^{d \times h \times w}$ , 那么 activation 可以计算为

$$s_{k,\delta} = \sum_{i=1}^d \sum_{\tau \in \Omega} w_{k,i,\tau} h_{i,\delta+\tau} = \langle \mathbf{w}_k, \mathbf{h}_\delta \rangle \quad (335)$$

此处使用  $\mathbf{W} \in \mathbb{R}^{n \times p}$ ,  $p = dF_hF_w$  作为总的 filters.

### 33.2.3 Group Based Orthogonalization: Divided Filters

权重按行分为多组, 每组大小一致 (为一个小数  $N_G < d$ , 如 64/128), 使得 EVD 的计算代价很小. 大大降低计算难度.

## 34 OrthDNNs: Orthogonal DNNs(TPAMI 19')

在数据分布  $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$  上, ML 的目标是最优化期望风险

$$R(f) = \mathbb{E}_{z \sim P}[\mathcal{L}(f(\mathbf{x}), y)] \quad (336)$$

然而真实分布未知, 所以用样本期望 (训练集  $S_m$ ) 代替

$$R_m(f) = \frac{1}{m} \sum_{i=1}^m \mathcal{L}(f(\mathbf{x}_i), y_i) \quad (337)$$

统计学习的一大目标是计算泛化误差/gen. gap

$$\text{GE}(f_{S_m}) = |R(f_{S_m}) - R_m(f_{S_m})| \quad (338)$$

这里考虑由 DNN 建模的分类-表示模型 (Class. Repr. Learning)

$$\begin{aligned} R(f, T) &= \mathbb{E}_{z \sim P}[\mathcal{L}(f(T\mathbf{x}), y)] \\ R_m(f, T) &= \frac{1}{m} \sum_{i=1}^m \mathcal{L}(f(T\mathbf{x}_i), y_i) \end{aligned} \quad (339)$$

### 34.1 GE Analysis in a Robustness and Isomeric Mapping Perspective

**Definition 34.1** ( $K, \epsilon(\cdot)$ -robustness) 对于  $K \in \mathbb{N}, \epsilon : \mathcal{Z}^m \mapsto \mathbb{R}$  一个算法是  $K, \epsilon(\cdot)$ -robust 的, 如果  $\mathcal{Z}$  可以分为  $K$  个分离的集合, 记为  $\mathcal{C} = \{C_k\}_{k=1}^K$ , 且

$$\begin{aligned} \forall s_i = (\mathbf{x}_i, y_i) \in C_k, \forall z = (\mathbf{x}, y) \in C_k \\ \implies |\mathcal{L}(f(\mathbf{x}_i), y_i) - \mathcal{L}(f(\mathbf{x}), y)| \leq \epsilon(S_m) \end{aligned} \quad (340)$$

对于任何健壮的算法, 有以下定理<sup>9</sup>

**Theorem 34.2** 如果一个算法是  $(K, \epsilon(\cdot)$ -robust, 并且一个损失函数  $\mathcal{L}$  是有界的, 则

$$\Pr \left( \text{GE}(f_{S_m}) \leq \epsilon(S_m) + M \sqrt{\frac{2K \log(2) + 2 \log(1/\nu)}{m}} \right) \geq 1 - \nu \quad (341)$$

**Definition 34.3** (*Covering Number*) 给出一个度量空间  $(\mathcal{M}, d)$ , 集合  $\hat{S}$  是另一个集合  $S$  的  $\gamma$ -cover, 若  $\forall s \in S, \exists \hat{s} \in \hat{S}, s.t. d(\hat{s}, s) \leq \gamma$ . 集合  $S$  的  $\gamma$ -covering number 是

$$\mathcal{N}_\gamma(S, \rho) = \min \{|\hat{S}| : \hat{S} \text{ is a } \gamma \text{-covering of } S\} \quad (342)$$

<sup>9</sup>Huan Xu and Shie Mannor. Robustness and generalization. Machine Learning, 86(3):391–423, 2012.1,2,4,6,8

**Definition 34.4** ( $\delta$ -isometry) 映射  $T : \mathcal{P} \mapsto \mathcal{Q}$ (其中  $\mathcal{P}, \mathcal{Q}$  度量空间) 是  $\delta$ -保距的, 若

$$\forall \mathbf{x}, \mathbf{x}' \in \mathcal{P}, |\rho_Q(T\mathbf{x}, T\mathbf{x}') - \rho_P(\mathbf{x}, \mathbf{x}')| \leq \delta \quad (343)$$

**Theorem 34.5** 对与任何 CRL 问题的算法, 若  $\mathcal{L} \circ f$ (关于  $T\mathbf{x}$ ) 的 Lipschitz Constant 有上界  $A$ , 且  $T$  是  $\delta$ -保距的, 且  $\mathcal{X}$  是紧的, 且有 covering number  $\mathcal{N}_{\gamma/2}(\mathcal{X}, \rho)$ , 则这个算法是  $(|\mathcal{Y}| \mathcal{N}_{\gamma/2}(\mathcal{X}, \rho), A(\gamma + \delta))$ -robust.

## 34.2 GE Analysis of DNN

<sup>10</sup>

## 34.3 OrthDNN by SVB(Singular Value Bound)

类似于之前的一篇, 为在 Stiefel 流形上的约束优化. 可以通过流形上的 SGD(切空间投影法) 或者 Frank-Wolfe 算法(流形投影法). 太慢! 使用 SVB 来进行优化.

- 使用估计的带偏移的梯度投影方向. 每隔一定数量迭代拉回到流形上.
- 考虑到 DNN 的优化问题含有巨大数量的局部极小/临界点. 在目标流形附近探索可能会得到更好的解(避免 local minima/critical points)
- BN 会改变权重矩阵的谱, 使得严格正交化的努力白费.

Algorithm Sketch(SVB)

1. 使用普通的 SGD 来更新参数.
2. 每隔一些 epochs, 对于每一层, 使用 SVD 来 clamp 奇异值到  $((1 + \epsilon)^{-1}, 1 + \epsilon)$ :

$$\mathbf{W} = \mathbf{U}\Sigma\mathbf{V}^T \quad (344)$$

$$\mathbf{W} \leftarrow \mathbf{U}\Sigma_{\epsilon-clamped}\mathbf{V}^T \quad (345)$$

还可以使用罚函数法来作为无约束优化问题优化 (SoftRegu)

$$\min_{\Theta=\{\mathbf{W}_l, b_l\}_{l=1}^L} \mathcal{L}(\{\mathbf{x}_i, y_i\}_{i=1}^m; \Theta) + \lambda \sum_{l=1}^L \|\mathbf{W}_l^\top \mathbf{W}_l - \mathbf{I}\|_F^2 \quad (346)$$

需要假设  $n_l \geq n_{l-1}$ , 为了放松这一假设, 使用自然 1-范数/谱范数 (SRIP) 而不是 Frobenius 范数

$$\min_{\Theta=\{\mathbf{W}_l, b_l\}_{l=1}^L} \mathcal{L}(\{\mathbf{x}_i, y_i\}_{i=1}^m; \Theta) + \kappa \sum_i^L \sigma_{\max}(\mathbf{W}_l^\top \mathbf{W}_l - \mathbf{I}) \quad (347)$$

<sup>10</sup> TO READ AND MAKE NOTES

### 34.3.1 BN Compatibility

BN 实际上干了

$$\text{BN}(\mathbf{h}) = \Upsilon \Phi(\mathbf{h} - \boldsymbol{\mu}) + \boldsymbol{\beta} \quad (348)$$

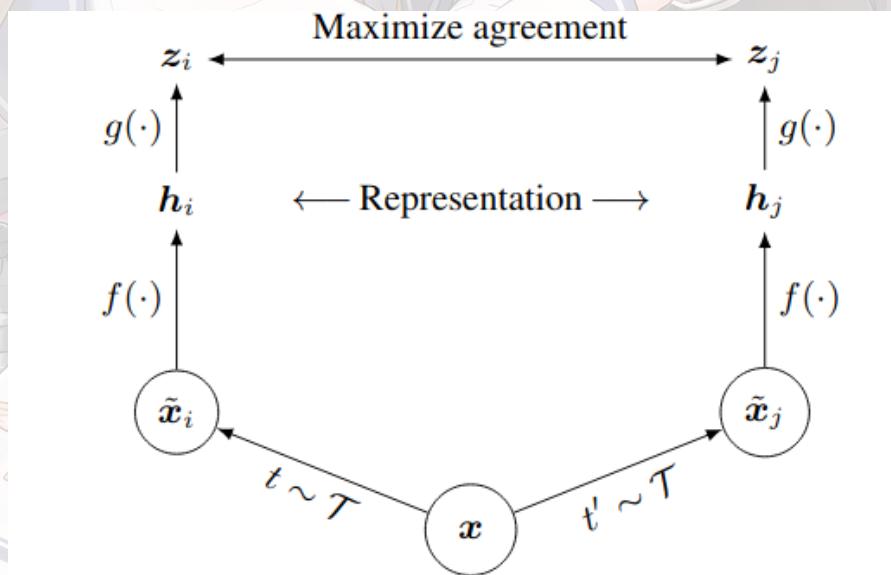
其中  $\boldsymbol{\mu} \in \mathbb{R}^n$  是这一层 n 个神经元的 (batch 的) 均值, 对角阵  $\Phi = \text{Diag}(1/\phi_i)$  是神经元输出的标准差倒数矩阵 (加上小值来增加数值稳定性). 并且  $\Upsilon, \boldsymbol{\beta}$  是可训练的对角矩阵和 bias. 最终的均值和标准差由 running average 给出训练样本的总体估计.

但是乘上一个对角矩阵会导致原先的权重矩阵 sig-val 不再相等! 所以让可训练的对角阵和偏差在每层都相等  $\Rightarrow$  对于精确形式的 OrhDNN, 使用 DBN(Degenerate Batch Norm), 其中  $\bar{\phi} = \frac{1}{n} \sum_n \phi_i$  对于 approx. OrthDNN, 使用 BBN(Bounded BN), 控制 (通过 clamp) $\{v_i/\phi_i\}$  在均值  $\alpha = \frac{1}{n} \sum_i \frac{v_i}{\phi_i}, v_i/\phi_i \in [\alpha(1+\epsilon)^{-1}, \alpha(1+\epsilon)]$  内.

### 34.3.2 On CNN: OrthDNN as Convolution

对于 CNN 的某一层, 要求的参数形状为  $n_l \times n_{l-1} \times n_h \times n_w$ , 本作使用生成  $n_l \times n_{l-1} n_h n_w$  形状的参数矩阵来得到这些 filter.

## 35 SimCLR: A Simple Framework for Contrastive Learning of Visual Representations



### 35.1 Ideas & Basics

1. 使用随机数据增强来得到两个相关视图:

- 随机剪裁/裁剪 + 缩放至原大小



- 随机色彩 distortion
- 随机高斯模糊
- 没啥用的: cutoff, Sobel filters etc.

2. Encoder 使用经典的 ResNet 结构.
3. 一个小的 MLP 投影头 (proj. head) 用于投影到 contrastive loss 用于计算的空间.
4. 在最后的特征空间上最小化 InfoNCE(yet another *contrastive loss*), 最大化相关图的 MI

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j) / \tau)}{\sum_{k=1}^{2N} \mathbf{1}_{[k \neq i]} \exp(\text{sim}(z_i, z_k) / \tau)} \quad (349)$$

其中相似度度量使用归一化点积 (i.e. cosine similarity)  $\text{sim}(\mathbf{u}, \mathbf{v}) = \mathbf{u}^\top \mathbf{v} / \|\mathbf{u}\| \|\mathbf{v}\|$  其中负样本来自同一 minibatch,  $\tau$  是温度系数.

本方法没有利用类似 MoCo 的 memory bank, 而是使用较大的 batchsize(256-8192).

SGD+Momentum 的优化器似乎和较大的 batchsize 上不稳定, 所以使用 LARS 优化器.

使用 Global BN: 在各个设备上本来 BN 的均值和方差是分别计算的, 这里在 BN 里使用聚合了所有设备的 mean/variance. 其他方法包括在设备间 shuffle samples, 或者使用 layer norm<sup>11</sup>而不是 BN.

## 36 BYOL: Build Your Own Latent

### 36.1 Ideas & Method

使用两个网络来学习: online/target 网络, 每个网络都类似得由三个阶段组成:encoder  $f_\theta$ , proj. head  $g_\theta$ , predictor  $q_\theta$ (discriminator). online/target 网络使用相同的结构, 但是为不同的参数, 并且 target 网络的参数  $\xi$  是 online 参数  $\theta$  的 moving average(under decay const.  $\tau$ )

$$\xi \leftarrow \tau \xi + (1 - \tau) \theta \quad (352)$$

<sup>11</sup>在 MLP 中, 归一化每一层的权重为期望 0 标准差 1, 先计算均值和方差

$$\mu^l = \frac{1}{H} \sum_{i=1}^H w_i^l \quad \sigma^l = \sqrt{\frac{1}{H} \sum_{i=1}^H (w_i^l - \mu^l)^2} \quad (350)$$

在计算 FP 前进行权重归一化

$$\hat{\mathbf{w}}^l = \frac{\mathbf{w}^l - \mu^l}{\sqrt{(\sigma^l)^2 + \epsilon}} \quad (351)$$

LN 在权重够多的情况下和 BN 效果类似. 似乎在 CNN 上不能使用???

**Remark** 对于权重的更弱的限制 (相比 OrthDNN 和 OWN 中近似正交化权重), 注意随机正交矩阵必然是归一化了的, 分量还是不相关的.

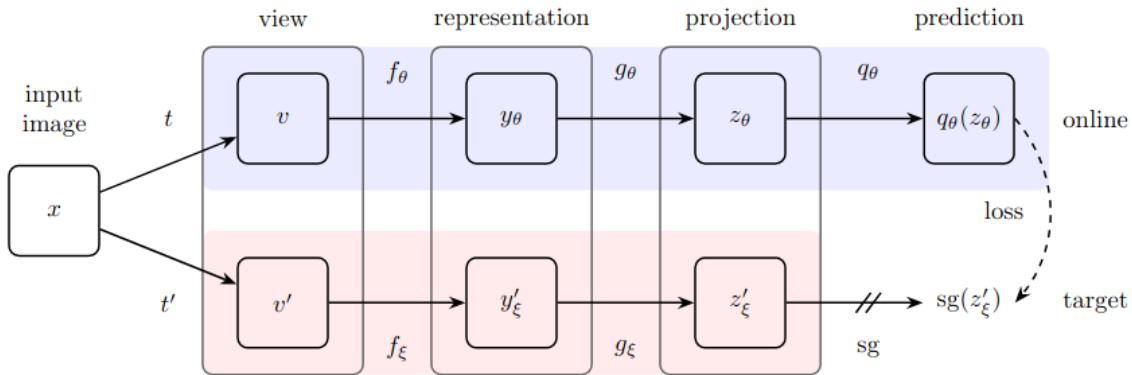


Figure 2: BYOL's architecture. BYOL minimizes a similarity loss between  $q_\theta(z_\theta)$  and  $\text{sg}(z'_\xi)$ , where  $\theta$  are the trained weights,  $\xi$  are an exponential moving average of  $\theta$  and  $\text{sg}$  means stop-gradient. At the end of training, everything but  $f_\theta$  is discarded, and  $y_\theta$  is used as the image representation.

假设通过两个图像变换之后的 view 是  $v, v'$ , 然后我们使用要区分两者的 repr., 所以最小化负的 cosine-similarity<sup>12</sup>

$$\mathcal{L}_{\theta,\xi} \triangleq \|\overline{q}_\theta(z_\theta) - \bar{z}'_\xi\|_2^2 = 2 - 2 \cdot \frac{\langle q_\theta(z_\theta), z'_\xi \rangle}{\|q_\theta(z_\theta)\|_2 \cdot \|z'_\xi\|_2} \quad (353)$$

同样的交换两个视图送到网络的顺序, 并且得到另一个 loss, 加和来得到对称化 loss

$$\mathcal{L}_{\theta,\xi}^{\text{BYOL}} = \mathcal{L}_{\theta,\xi} + \tilde{\mathcal{L}}_{\theta,\xi} \quad (354)$$

参数的更新

$$\begin{aligned} \theta &\leftarrow \text{optimizer } (\theta, \nabla_\theta \mathcal{L}_{\theta,\xi}^{\text{BYOL}}, \eta), \\ \xi &\leftarrow \tau \xi + (1 - \tau) \theta \end{aligned} \quad (355)$$

BYOL 没有使用显式的方法防止 mode collapse(negative samples etc.). 但是根据假设, 他们证明了, 至少在 predictor 是 optimal 的时候 ( $q_\theta = q^*$ ), 鞍点是不稳定的.

## 37 MoCo: Momentum Contrast for Unsupervised Visual Representation Learning

### 37.1 Ideas & Method

InfoNCE on memory bank

$$\mathcal{L}_q = -\log \frac{\exp(q \cdot k + / \tau)}{\sum_{i=0}^K \exp(q \cdot k_i / \tau)} \quad (356)$$

Momentum encoder: 按照 moving average 更新 k-encoder

$$\theta_k \leftarrow m\theta_k + (1 - m)\theta_q \quad (357)$$

Tricks: Shuffling BNs

<sup>12</sup>equivalent to MSE of normalized feature

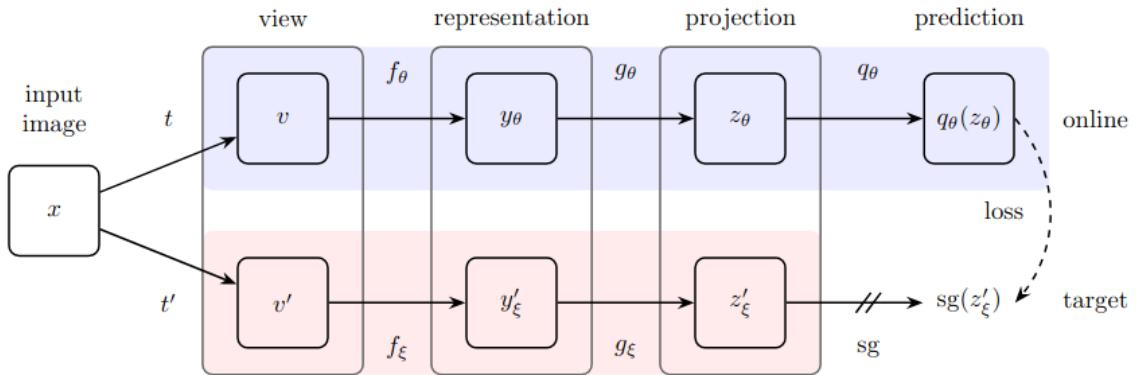


Figure 2: BYOL’s architecture. BYOL minimizes a similarity loss between  $q_\theta(z_\theta)$  and  $\text{sg}(z'_\xi)$ , where  $\theta$  are the trained weights,  $\xi$  are an exponential moving average of  $\theta$  and  $\text{sg}$  means stop-gradient. At the end of training, everything but  $f_\theta$  is discarded, and  $y_\theta$  is used as the image representation.

## 38 SimSiam: Exploring Simple Siamese Representation Learning

### 38.1 Intuitions

BYOL 既没有使用 negative samples, 但是使用了 momentum encoder+ 不对称的 proj. head+stop-grad 策略来避免 collapse, 并且不需要巨大的 batchsize! 所以本方法可以考虑为 BYOL-momentum encoder, 使用两个共享权重的网络 (like SwAV without online clustering/SimCLR without negative pairs), 即使如此, 不会导致 mode collapse!

**Remark** stop-grad 是关键的, 没有 stop-grad 则会直接收敛到 trivial 解 (mode collapse)

### 38.2 Method

使用共享权重的 backbone(like ResNet)+proj. head(MLP), 最小化不同视图最后特征的负 cosine-sim.

$$\mathcal{D}(p_1, z_2) = -\frac{p_1}{\|p_1\|_2} \cdot \frac{z_2}{\|z_2\|_2} \quad (358)$$

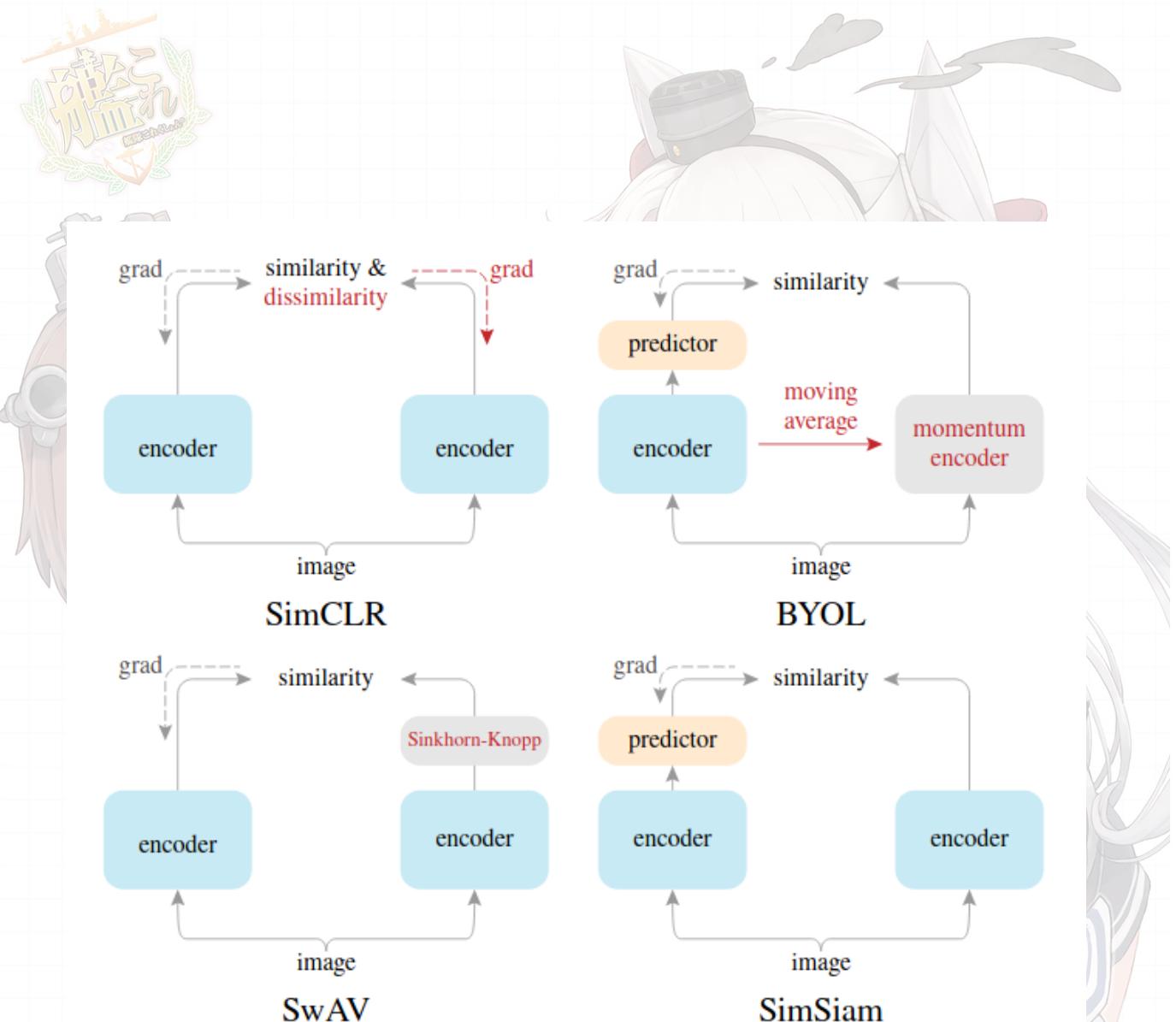
并且使用对称化 loss

$$\mathcal{L} = \frac{1}{2}\mathcal{D}(p_1, z_2) + \frac{1}{2}\mathcal{D}(p_2, z_1) \quad (359)$$

并且不对一边的网络进行参数更新.

#### Settings

- Optimizer: SGD for pre-training.  $lr = lr_{base} \times \text{BatchSize}/256$ , base lr 为 0.05, 使用 cosine decay(annealing) schedule, L2 正则化 1e-4, 动量 0.9.
- Batch Size 512, Synchronized BN across devices.



**Figure 3. Comparison on Siamese architectures.** The encoder includes all layers that can be shared between both branches. The dash lines indicate the gradient propagation flow. In BYOL, SwAV, and SimSiam, the lack of a dash line implies stop-gradient, and their symmetrization is not illustrated for simplicity. The components in red are those missing in SimSiam.



- Proj. Head: 3FCN of 2048-d.
- Pred. Head: 2FCN of 2048-512-2048-d.

## 39 Graph Layouts by t-SNE

### 39.1 Backgrounds

#### Dimension Reduction

投影高维数据到低维, 可分为 distance-preserving 方法, 最小化 (aggregated normalized) stress

$$\sigma = \sum_{i,j} \left( \frac{d(\mathbf{x}_i, \mathbf{x}_j) - \|\mathbf{y}_i - \mathbf{y}_j\|}{d(\mathbf{x}_i, \mathbf{x}_j)} \right)^2 \quad (360)$$

以及 neighborhood-preserving 方法, 最大化 kNN 的重叠. distance-preserving 方法在特别高维时失效很严重 (由于高维空间的各种奇性).

### 39.2 Method

#### 39.3 tsNET

## 40 细粒度图像数据分类 by Xiangteng He

困难: 标注成本巨大, 依赖人工先验, 忽略辨识速度 (like RPN, 2-stages), 忽略语义关联

### 40.1 细粒度图像数据分类

基于显著性图

### 40.2 RL-based 图像部件/对象识别

### 40.3 多层注意力区域辨识

基于 IoU 的奖励函数强化学习

### 40.4 多模态

- 引入文本信息, 音频, 视频
- 数据集构建 PKU FG-Xmedia
- ResNet-based 统一处理各种模态, 视频抽帧, 音频使用 mel 谱图, 文本使用 word embedding

Future: 大规模细粒度图像分类, 细粒度视觉推理, 图像到跨媒体迁移

## 41 Dirac Operator for Extrinsic Shape Analysis

为什么需要微分算子? 他们提供了一组流形上的基函数 (Hilbert 空间的基), 以及对应特征值. 流形上的函数可以在这些基上展开, 并且提供了傅立叶变换和卷积. 在紧流形上, 一个算子有离散的特征基, 若他是自伴和椭圆算子 (矩阵为对称的和正定的). 然而很多微分算子并不满足这些条件:

- Laplace-Beltrami Op., i.e.  $\Delta$ , 一种曲面光滑性的度量.
- Hessian 矩阵, 和 modified Dirichlet 能量.  $\sum_i E_D(N^i \phi)$
- 各项异性 Laplacian  $\Delta_A = \operatorname{div} A \nabla$

对于曲面  $M$ ,  $L$  算子可以看做 Dirichlet 能量的 Hessian. 我们提出 Dirac Op.

### 41.1 Math

使用四元数来表示曲面 (一个嵌入  $f : M \hookrightarrow \operatorname{Im} \mathbb{H}$ ). 假设这个嵌入是共形 (conformal) 的. 使用  $\kappa_1, \kappa_2$  代表两个主曲率, 有高斯曲率  $K = \kappa_1 \kappa_2$  和平均曲率  $H := \frac{1}{2}(\kappa_1 + \kappa_2)$ .  $|df|^2$  代表嵌入  $f$  诱导的体积元.

对于实值函数,  $L$  算子是最低阶非平凡自伴椭圆算子. 对于复变函数, 则存在 1-order 微分算子: Dirac 算子. 经典例子是 Cauchy-Riemann(Poincare) 算子  $\bar{\partial}$ , 以及微分形式上的 Hodge-Dirac 算子  $\star d + d\star$ .

我们使用的是 extrinsic D-Op.

$$D\psi := -\frac{df \wedge d\psi}{|df|^2} \quad (361)$$

其中  $|df|^2$  的除法为应用 2-形式的 Hodge star 算子.

平方, 有

$$D^2\psi = \Delta\psi + \frac{dN \wedge d\psi}{|df|^2} \quad (362)$$

相对微分算子

$$D_{f_1, f_2}\psi := -\frac{df_2 \wedge d\psi}{|df_1|^2} \quad (363)$$

单参数内插

$$L(\tau) := (1 - \tau)\Delta + \tau D_N \quad (364)$$

算子  $D_N$  只是半正定的. 但是增加一些  $L$  算子的部分就能成为强正定的. 这些算子和曲率也有一定关系,  $\Delta = \Delta_{\mathbb{R}^2} e^{-2u}$ ,  $u$  是对数共形尺度因子, 且满足 Yamabe 方程  $\Delta u = K$ . 由于要最小化  $\langle\langle \Delta\varphi, \varphi \rangle\rangle$ ,  $\Delta\varphi$  在有很大高斯曲率的地方会很小.

MDE(modified Dirichlet energy) 也可写成

$$E_{MDE}(\varphi) = \langle\langle \Delta\varphi, \varphi \rangle\rangle + \int_M \varphi^2 (\kappa_1^2 + \kappa_2^2) dA \quad (365)$$

这意味着它的 Hessian 是  $\Delta + U$ , 后者是 Willmore 势能  $U = \kappa_1^2 + \kappa_2^2$ , 这鼓励特征函数在曲率很大地方有小的方差.

## 41.2 离散化

对以一个 2-流形的三角 mesh  $K = (V, E, F)$ , 不同于连续的情况, 使用法向量  $N$  代替  $f$ . 对于任何 mesh 上的三角  $ijk \in F, \psi : V \mapsto \mathbb{H}$  有

$$(D_N \psi)_{ijk} := -\frac{1}{2A_{ijk}} \sum_{pqr \in C(ijk)} (N_r - N_q) \psi_p \quad (366)$$

其中  $A_{ijk}$  是三角面积,  $C(ijk)$  是轮换. 这个算子可以写成张量  $D \in \mathbb{H}^{|F| \times |V|}$

$$D_{ijk,p} = -(N_r - N_q) / 2A_{ijk} \quad (367)$$

## 41.3 实值表示

用  $4 \times 4$  矩阵代替四元数运算. 为了找到典型表示, 需要找到乘积常数  $q$ , 使得某个特征向量范数接近 1. 近似的有

$$q := \sum_{i=1}^{|V|} A_i \phi_i^{-1} \quad (368)$$

其中  $A_i$  是对偶区域的面积.

## 41.4 有界区域: 边界条件

边界条件可以是 Dirichlet/Neumann 条件. 这里提出无限势阱边界. 增加一个势能项  $U$ , 并且在有界区域之外迅速趋向无穷, 使得优化  $\langle \{(\Delta + U)\psi, \psi\} \rangle$  让波函数迅速在区域外收敛到 0.

$$U(p) := \frac{c}{1 + (e^{-(d(p,q)-\beta)})^\gamma} \quad (369)$$

$q$  是质心,  $c$  是大常数.

为了使用这个罚函数, 在算子中加入对角矩阵  $U_{ii} = A_i U(p_i)$

# 42 Mesh-Based Simulation with GNNs

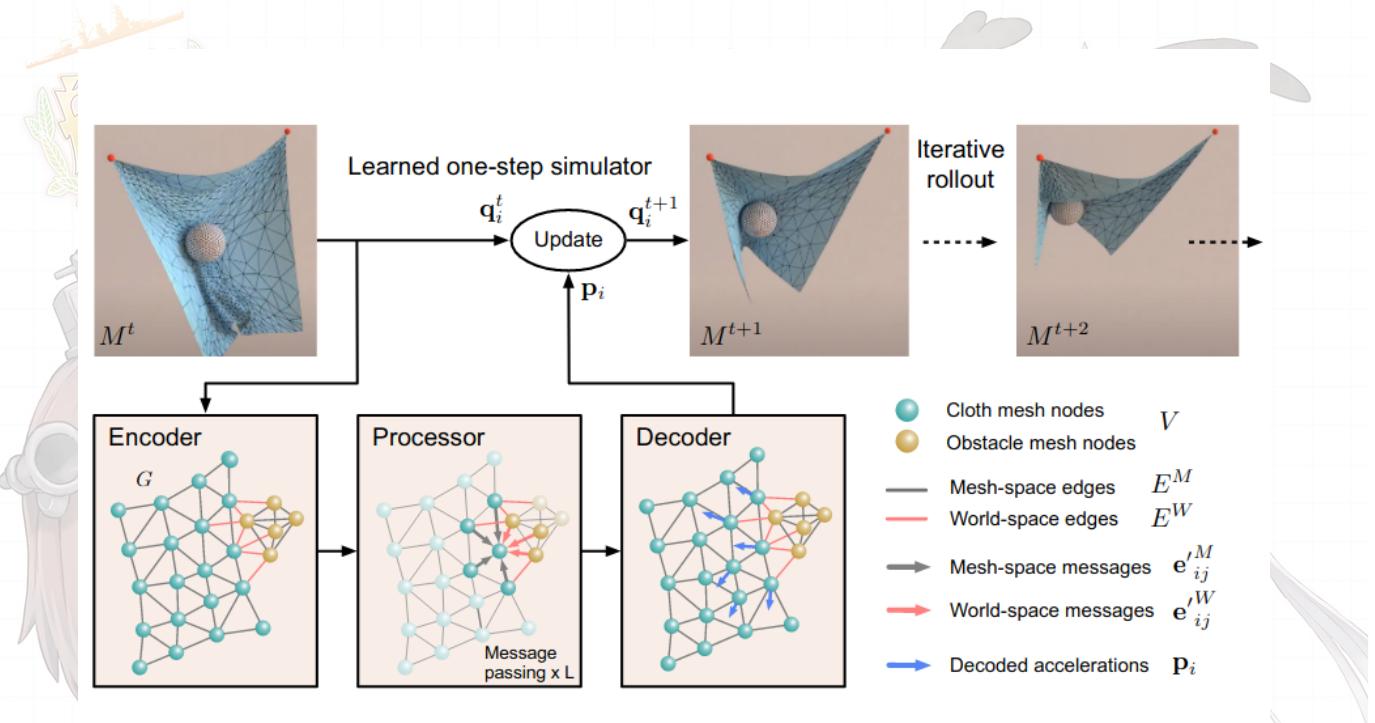
使用 World/Mesh 两种空间的输入. 预测物理体系的时间序列. 使用  $M^t = (V, E^M)$  描述  $t$  时间的系统状态. 每个结点都有 mesh-space 坐标  $\mathbf{u}_i$  和动量/速度坐标  $\mathbf{q}_i$ . 对于欧拉系统, 建模的是固定的速度场, 对于 Lagrange 系统, mesh 是动态移动的, 并且包含了绝对世界坐标  $\mathbf{x}_i$ .

## 42.1 结构

Encoder: 把  $M^t$  编码成多重图  $G = (V, E^M, E^W)$ , 并且增加了世界坐标下的边  $E^W$ . 这里使用简单的 rNN 采样, 选取半径为最短的 mesh 边长  $r_W$ . 并且 mesh 图边权为相对坐标  $\mathbf{u}_{ij} = \mathbf{u}_i - \mathbf{u}_j$ , 同理世界坐标图中也为相对坐标.

Processor: GNN. 其中

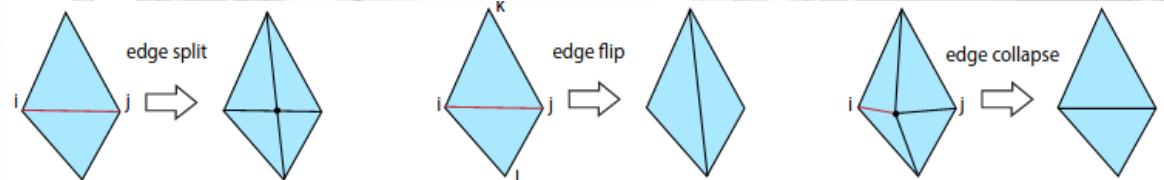
$$\mathbf{e}'_{ij}^X = f^X(\mathbf{e}_{ij}^X, \mathbf{v}_i, \mathbf{v}_j) \quad (370)$$



Decoder: 使用在结点 feature 上的 MLP. 得到输出向量  $p_i$

Updater: 对于一阶系统,  $\mathbf{q}_i^{t+1} = \mathbf{p}_i + \mathbf{q}_i^t$ , 二阶系统:  $\mathbf{q}_i^{t+1} = \mathbf{p}_i + 2\mathbf{q}_i^t - \mathbf{q}_i^{t-1}$ . 输出向量的额外维度也用于预测附加属性(压力/张力 etc.)

## 4.2.2 Adaptive Remeshing



在需要的地方增加 mesh 数量.(有限元那味儿) 使用 sizing filed 方法. 定义 sizing field 张量

$\mathbf{S}(\mathbf{u}) \in \mathbb{R}^{2 \times 2}$ , 同时一条边是合理的 iff.  $\mathbf{u}_{ij}^T \mathbf{S}_i \mathbf{u}_{ij} \leq 1$ , 否则需要拆分这条边.

使用动态的 remeshing. 使用如上文的 GNN 的结构来学习 sizing 张量场. 每一步, 我们都计算动量信息和 resizing 张量场, 再用 remesher 计算下一步的 mesh.

给定了 sizing 张量场  $\mathbf{S}_i$ , 对于每条边,

- 需要分割, 若  $\mathbf{u}_{ij}^T \mathbf{S}_{ij} \mathbf{u}_{ij} > 1$ ,  $\mathbf{S}_{ij} = \frac{1}{2} (\mathbf{S}_i + \mathbf{S}_j)$
- 需要 collapse, 若 collapse 不会产生新的无效边
- 需要 flip, 若满足各项异性 Delaunay 条件

$$(\mathbf{u}_{jk} \times \mathbf{u}_{ik}) \mathbf{u}_{il}^T \mathbf{S}_A \mathbf{u}_{jl} < \mathbf{u}_{jk}^T \mathbf{S}_A \mathbf{u}_{ik} (\mathbf{u}_{il} \times \mathbf{u}_{jl}), \quad \mathbf{S}_A = \frac{1}{4} (\mathbf{S}_i + \mathbf{S}_j + \mathbf{S}_k + \mathbf{S}_l) \quad (371)$$

顺序上, 先尽可能 split 所有边, 再尽可能 collapse 所有边, 再尽可能 flip 所有边  
若没有能估计 sizing field 的数据, 我们从 mesh 序列中估计之. 我们要找到能诱导出下一时刻 mesh 的 sizing field. 假设 remesher 几乎是最优的, 所有结果边都有效, 那么最大化度规  $\mathbf{S}$  下的边长

$$\mathbf{S}_i = \operatorname{argmax} \sum_{j \in \mathcal{N}_i} \mathbf{u}_{ij}^T \mathbf{S}_i \mathbf{u}_{ij}, \quad \text{s.t. } \forall j \in \mathcal{N}_i : \mathbf{u}_{ij}^T \mathbf{S}_i \mathbf{u}_{ij} \leq 1 \quad (372)$$

这是简单的凸优化 (找到最小面积, 0-中心的包含  $\mathbf{u}_{ij}$  的椭圆).

## 43 SENet

Squeeze-and-Excitation Module.

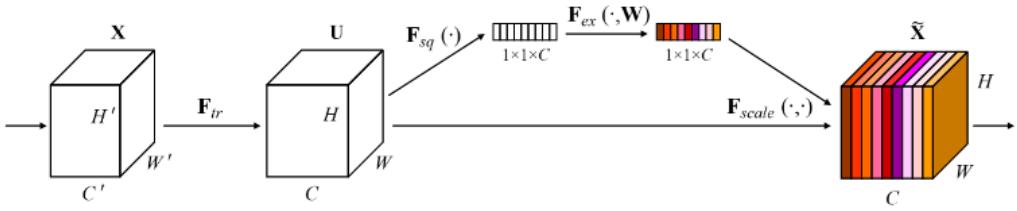


Fig. 1. A Squeeze-and-Excitation block.

计算 squeeze 变换 (单纯的计算每个 channel 的均值)

$$z_c = \mathbf{F}_{sq}(\mathbf{u}_c) = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W u_c(i, j) \quad (373)$$

计算 excitation,  $\sigma$  是 sigmoid,  $\delta$  是 ReLU

$$\mathbf{s} = \mathbf{F}_{ex}(\mathbf{z}, \mathbf{W}) = \sigma(g(\mathbf{z}, \mathbf{W})) = \sigma(\mathbf{W}_2 \delta(\mathbf{W}_1 \mathbf{z})) \quad (374)$$

其中  $\mathbf{W}_1 \in \mathbb{R}^{\frac{C}{r} \times C}$ ,  $\mathbf{W}_2 \in \mathbb{R}^{C \times \frac{C}{r}}$  维度形成某种 bottleneck, 降低模型复杂度.

将 excitation 乘到 channel 上

$$\tilde{\mathbf{x}}_c = \mathbf{F}_{scale}(\mathbf{u}_c, s_c) = s_c \mathbf{u}_c \quad (375)$$

13

<sup>13</sup> in matrix-sense,

$$\mathbf{z} = \frac{1}{HW} \mathbf{u}_{ij} \mathbf{1}^{ij} \in \mathbb{R}^{C \times 1} \quad (376)$$

$$\mathbf{s} = \mathbf{F}_{ex}(\mathbf{z}) \in \mathbb{R}^{C \times 1} \quad (377)$$

$$\tilde{\mathbf{X}} = \mathbf{s}^T \odot \mathbf{U} \in \mathbb{R}^{N (= H \times W) \times C} \quad (378)$$

## 44 Continuous-Time Spiking Neural Network

模拟真实神经元. alternative to Hebb 神经元. 传统的 SNN 模拟手段涉及时间驱动 (time-driven) 和事件驱动 (event-driven), 前者涉及时间格点上的离散化, 可能带来误差. 采用 event-driven 来实现连续时间.

### 44.1 Neurons

$$\begin{aligned} S &= S_p + P_r P_w - T_l, \text{ for } S < S_{th} \\ S &= S_p + P_r P_w + T_r, \text{ for } S \geq S_{th} \end{aligned} \quad (379)$$

- $S_p$ , 之前的状态
- $P_r$ : presynaptic weight, 受到的刺激
- $P_w$ : postsynaptic weight, 神经元连接强度
- $T_l$ : 电位 leakage  $Ld\Delta t$
- $T_r$ :  $T_r = \frac{(S_p-1)^2 \Delta t}{1-(S_p-1)\Delta t}$
- $t_f$ : firing time, obeys firing equation  $t_f = \frac{1}{S-1}$

可以有 excitatory/inhibitory 的输入区别.

### 44.2 Network Topology

一般来说, en 比 in 多, 后者占网络的 15-25%, 这样可以增进网络的稳定性. 没有 in-in 突触, 这回导致不稳定性, 因为这会让 in 减少, 进而导致不受控制的激发.

### 44.3 突触塑性规则

”后突触规则”: 是相对于同一突触的突触总体模式/时间点控制了后突触效用.

- 对数衰减. 所有后突触权重 postsynaptic weight 对数衰减.

$$P_w = P_{w,\min} + (P_w - P_{w,\min}) e^{-\frac{\Delta t}{\tau}}$$

- 同突触增强. 当一个 spiking event 在一个突触上发生, PW 增加, 和同神经元 + 同突触的上一个刺激成函数 (在一个时间窗口内).
- 异突触增强. 当一个 spiking event 在一个突触上发生, PW 增加, 和同神经元 + 其他突触的上一个刺激成函数 (在一个时间窗口内).

同/异突触增强使用以下方程控制.

$$\Delta P_w = \eta (P_{w,\max} - P_w) \quad (380)$$

子后, 还可以采取其他策略, 如 STDP, Synaptic Scaling etc.

## 45 Towards Deep Learning Models Resistant to Adversarial Attacks

从优化的角度来看, 这是个 robust-optimization 的问题. 给出 empirical loss  $\mathbb{E}_{(x,y) \sim \mathcal{D}}[J(x, y, \theta)]$ , 我们可以适用 robust optimization 模型

$$\begin{aligned} & \min_{\theta} \rho(\theta) \\ & \text{where } \rho(\theta) = \mathbb{E}_{(x,y) \sim \mathcal{D}} [\max_{\delta \in \mathcal{S}} J(\theta, x + \delta, y)] \end{aligned} \tag{381}$$

其中 perturbation set 可以是  $l_\infty$ -球. 当然也可以基于别的视觉相似度.

### 45.1 Inner Maximize Prob. : 如何提出好的对抗样本

关于对抗样本和攻击, 存在两个问题:

1. 如何获得强的对抗样本
2. 如何训练一个没有/难以获得对抗样本的网络.

Fast Gradient Sign Method(FGSM) 是一个如下获得  $l_\infty$ -有界对抗样本的

$$x + \varepsilon \operatorname{sgn}(\nabla_x J(\theta, x, y)) \tag{382}$$

一个更强大的方法可以进行多步:FGSM<sup>k</sup>, 基本上就是多步 Projected GD

$$x^{t+1} = \Pi_{x+\mathcal{S}} (x^t + \alpha \operatorname{sgn}(\nabla_x J(\theta, x, y))) \tag{383}$$

实验证明,  $l_\infty$ -球中包含了很多有着相似 loss 值的 local maxima. 具体上讲,

1. PGD 导致的 loss 上升具有同种趋势 (曲线形状)
2. PGD 导致的最终 loss 非常密集
3. 计算了这些最终点的距离, 都和平均距离相差不大, 且趋向于成 90°.

这说明了 PGD 是泛用的一阶对抗方法.

猜想: PGD 对抗样本不会比其他一阶样本的 loss 小. 那么, 就是用 PGD 来最大化内部最大化问题.

### 45.2 最小化问题

SGD, 使用 perturbed points 来进行最小化的输入. Danskin 定理说对于连续可微函数<sup>14</sup>, 此时的梯度确实是 saddle/minmax optimization 的一个下降梯度. 虽然 ReLU/max-pooling 并不是连续可微的, 但是大概用也行.

<sup>14</sup>待补充

### 45.3 网络能力 & 对抗健壮性

对抗健壮性和网络能力相关. high-level 的讲, 更强的分类器趋向于有好的健壮性, 因为对抗样本的存在让分类边界更复杂.

一些观察

1. 网络容量提高了对抗性能. 训练超长时间提高了对抗性能.
2. one-step attack 也能提高健壮性
3. 弱模型可能再 saddle opt. 中不会收敛
4. 更高的容量和强对抗降低 transferred attacks 的有效性.

## 46 CAS: Channel-wise Activation Suppressing Module for Adversarial Robustness

考虑 channel-wise 的激活强度. 他们发现对抗训练可以显著降低对抗样本的 channel-wise act. magnitude, 以至于和自然样本相当近似. 考虑 channel-wise 激活频率, 可以发现对抗样本的激活频率非常均匀 (按照自然样本输入的激活频率排序), 对抗训练可以对齐这两个趋势, CAS 可以直接大大减少 activation 频率.

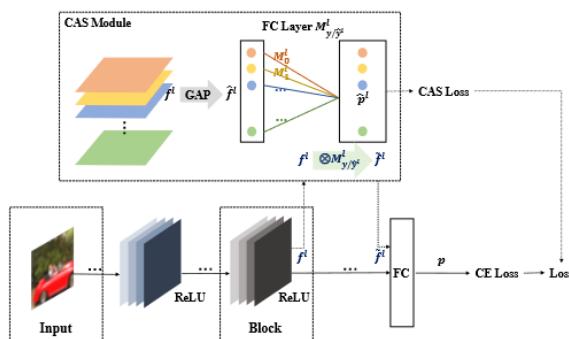


Figure 3: Framework of our proposed Channel-wise Activation Suppressing (CAS).

假设第 1 层的网络输出是  $f^l$ , 考虑 in-channel average

$$\hat{f}_k^l = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W f_k^l(i, j) \quad (384)$$

这被送进一个辅助 MLP(FC) classifier

$$\hat{p}^l = \text{softmax}(\hat{f}^l M^l) \in \mathbb{R}^C \quad (385)$$

其中每个类别的参数可以写成  $M^l = [M_1^l, M_2^l, \dots, M_C^l] \in \mathbb{R}^{K \times C}$  用 GT 的 label 分量去按通道乘 (在 test set 上直接用 classifier 的结果来计算)act. map

$$\tilde{f}^l = \begin{cases} f^l \otimes M_y^l, & (\text{training phase}) \\ f^l \otimes M_{\hat{y}^l}^l, & (\text{test phase}) \end{cases} \quad (386)$$

模型训练基于两个 loss, 其中每个 CAS loss 为

$$\mathcal{L}_{\text{CAS}}(\hat{p}^l(x', \theta, M), y) = - \sum_{c=1}^C \mathbb{1}\{c = y\} \cdot \log \hat{p}_c^l(x') \quad (387)$$

注意这就是在辅助分类器上的 CE loss. 同时还有主 Loss

$$\mathcal{L}(x', y; \theta, M) = \mathcal{L}_{\text{CE}}(p(x', \theta), y) + \frac{\beta}{S} \cdot \sum_{s=1}^S \mathcal{L}_{\text{CAS}}^s(\hat{p}^s(x', \theta, M), y) \quad (388)$$

使用标准的 AT 技术. 也可使用 TRADES/MART.

## 47 Resisting Adversarial Attacks by $k$ -Winners-Takes-All

不使用梯度信息. 试图使用  $C^0$  不连续函数, 即  $k$ -Winners-Takes-All(kWTA) 激活函数 (代替 ReLU). kWTA 激活函数的优势是, 具有稠密的不连续点, 使得难以利用小的 perturbation 可控地改变结果, 因为这会经过不连续的部分, 导致梯度信息失效. 可以证明, 即使是无穷小位移也会导致神经元激活模式彻底改变!

\* k-WTA 可以看作是 ReLU 和 max-pool 的结合.

### 47.1 Related Work: Obfuscated Gradients

一种类型是利用随机性: 在输入前进行随机变换/增加随机化的计算层. 然而梯度仍然可以使用 MC 估计来得到, 仍然不是很 robust. 另一种是 shattered gradient, 它们让梯度对于攻击者不存在或者不正确, 通过故意离散化输入/人工地提高梯度数值不稳定性. 这些方法仍然 vulnerable, 可以通过 backward pass differentiable approx.(BPDA) 来估计梯度. 假设  $f = f_1 \circ f_2 \circ \dots \circ f_n$ , 其中  $f_i(x)$  是不连续的函数组分, 可以使用 smooth delegate function  $g$  来近似这些不连续函数! 对于 k-WTA, 他们没找到任何平滑的近似函数.

### 47.2 $k$ -Winners-Takes-All

$k$ -Winners-Takes-AllModule:

$$\phi_k(\mathbf{y})_j = \begin{cases} y_j, & y_j \in \{k \text{ largest elements of } \mathbf{y}\} \\ 0, & \text{Otherwise} \end{cases} \quad (389)$$

接着显然要决定  $k$  的大小. 使用自适应的大小: 定义  $\gamma \in (0, 1)$  作为比例系数, 称为稀疏度比例. 并且对所有层都固定. k-WTA 可以在每个 channel 上进行, 但是计算最有效/最简单的是直接在整个 tensor 上进行.

渐进时间复杂度  $O(N)$ <sup>15</sup>

### 47.3 Training

小的  $\gamma$  是 preferable 的, 但是越小训练越难以收敛. 所以使用一种迭代 finetune 的方法. 使用线性退火降低稀疏系数, 每次降低后训练两个 epoch.

### 47.4 Theory Understand of the Discontinuity

定义 activation pattern

$$\mathcal{A}(\mathbf{x}) := \{i \in [l] \mid x_i \text{ is one of the } k \text{ largest values in } \mathbf{x}\} \subseteq [l] \quad (390)$$

可以证明, 极小的输入变化也会导致  $\mathcal{A}(\mathbf{x})$  的变化. 整个网络是一个分段 linear 的映射.

注意  $\phi_k(W\mathbf{x} + \mathbf{b})$  的 act. pat. 对于正定线性变换不变, 应该衡量垂直于输入  $\mathbf{x}$  的分量的距离  $d(\mathbf{x}, \mathbf{x}')$ , 且  $\mathbf{x}' = c \cdot (\mathbf{x} + d(\mathbf{x}, \mathbf{x}') \mathbf{x}_\perp)$ <sup>16</sup>, 那么若权值矩阵  $W_{ij} \sim \mathcal{N}(0, 1/l)$ ,  $\mathbf{b} = 0$  初始化, 有以下定理

**Theorem 47.1** (Dense discontinuities). Given any input  $\mathbf{x} \in \mathbb{R}^m$  and some  $\beta$ , and  $\forall \mathbf{x}' \in \mathbb{R}^m$  such that  $\frac{d^2(\mathbf{x}, \mathbf{x}')}{\|\mathbf{x}\|_2^2} \geq \beta$ , if the following condition

$$l \geq \Omega\left(\left(\frac{m}{\gamma} \cdot \frac{1}{\beta}\right) \cdot \log\left(\frac{m}{\gamma} \cdot \frac{1}{\beta}\right)\right)$$

is satisfied, then with a probability at least  $1 - 2^{-m}$ , we have  $\mathcal{A}(W\mathbf{x} + \mathbf{b}) \neq \mathcal{A}(W\mathbf{x}' + \mathbf{b})$ .

此处  $l$  是层宽.

为什么 k-WTA 可以被训练? 即使那么不连续.

**Theorem 47.2** Consider  $N$  data points  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N \in \mathbb{R}^m$ . Suppose  $\forall i \neq j$ ,  $\frac{\mathbf{x}_i}{\|\mathbf{x}_i\|_2} \neq \frac{\mathbf{x}_j}{\|\mathbf{x}_j\|_2}$ . If  $N$  is sufficiently large, then with a high probability, we have  $\forall i \neq j$ ,  $\mathcal{A}(W\mathbf{x}_i + \mathbf{b}) \cap \mathcal{A}(W\mathbf{x}_j + \mathbf{b}) = \emptyset$ .

所以网络对于不同的输入可以独立的训练, 实际上这些 act. pat. 往往是弱相关的.

实际上, 这些性质在 ReLU 中也有体现

### 47.5 Related Works

White-box attack model: PDG, Deep-fool, C&W Attack, Momentum Iterative Method(MIM). Black-box(BB) attacks (Papernot et al., 2017). 基于 transfer, 涉及训练一个对抗性的网络来生成对抗样本.

Adversarial Training: AT, TRADES, free adversarial training(FAT).

<sup>15</sup> Problem: How to impl. in CUDA?

<sup>16</sup>i.e.,  $d(\mathbf{x}, \mathbf{x}') = (\mathbf{x} - \mathbf{x}')/c$

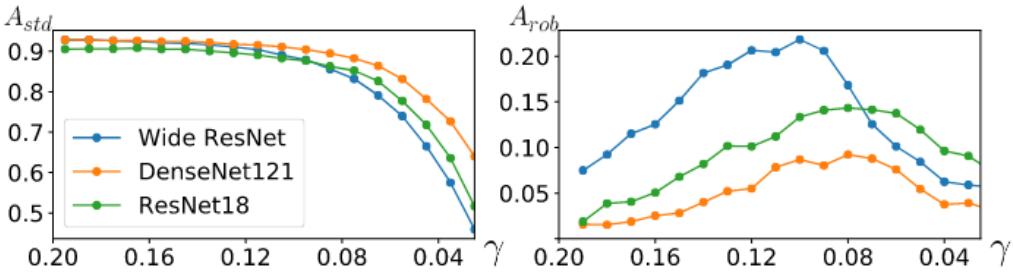


Figure 4: Robustness changing w.r.t.  $\gamma$  on CIFAR. When  $\gamma$  decreases, the standard test accuracy (left) starts to drop after a certain point. The robust accuracy (right) first increases then decreases.

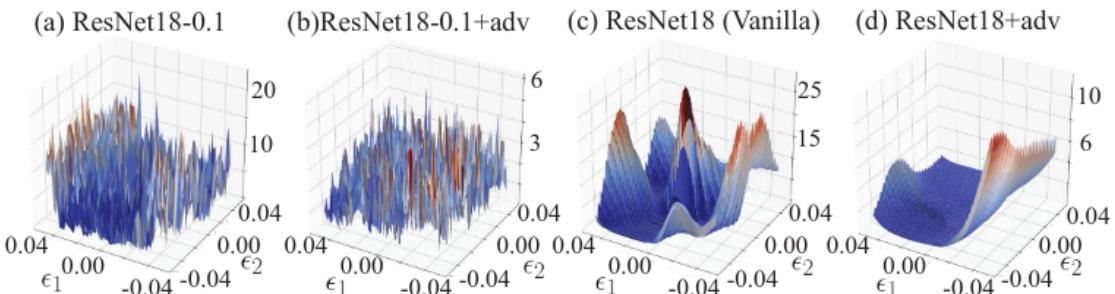


Figure 5: Gradient-based attack's loss landscapes in  $k$ -WTA (a, b) and conventional ReLU models (c, d). (a,b)  $k$ -WTA Models have much more non-convex and non-smooth landscapes. Also, the model optimized by adversarial training (b) has a lower absolute value of loss.

## 48 Normalized Loss Functions for Deep Learning with Noisy Labels

理论证明,  $l_1$  范数作为 loss 是 label-noise robust 的, 但是一些常用 loss 比如 cross-ent. 不是. GCE(Zhang & Sabuncu, 2018) 是 MAE 和 CE 的一种混合来提供健壮性. Symmetric CE(SCE) 结合了 RCE 和 CE, 只有 RCE 项是 robust 的. 本工作证明, 任何 loss 可以是 noise-robust 的, 只要加上正则化. 还发现许多 loss 都 underfit. 分类 loss 为 Active loss, 只显式最大化正样本概率和 Passive loss, 同时还显式地最小化负标签概率. 提出 APLs(Active Passive Losses)

Reverse Cross Entropy (RCE) loss (Wang et al., 2019c) 定义为:  $RCE = -\sum_{k=1}^K \mathbf{p}(k | x) \log \mathbf{q}(k | x)$ , with  $\mathbf{q}(k \neq y | x) = 0$  截断到小的阈值,  $\log(\mathbf{q}(k \neq y | x)) = A$  (eg.  $A = -4$ ). RCE robust to label noise, 可以和 CE 结合得到 SCE for robust classification and boosted learning (Wang et al., 2019c).

Focal Loss.

我们知道, 如果  $\sum_j^K \mathcal{L}(f(x), j) = C, \forall x \in \mathcal{X}, \forall f$  那么这个损失函数在温和的假设下是 noise robust 的. 那么我们手动地正则化 loss

$$\mathcal{L}_{\text{norm}} = \frac{\mathcal{L}(f(\mathbf{x}), y)}{\sum_{j=1}^K \mathcal{L}(f(\mathbf{x}), j)} \quad (391)$$

那么分别正则化 CE, MAE, RCE, FL:



1. NCE:

$$NCE = \frac{-\sum_{k=1}^K \mathbf{q}(k | \mathbf{x}) \log \mathbf{p}(k | \mathbf{x})}{-\sum_{j=1}^K \sum_{k=1}^K \mathbf{q}(y=j | \mathbf{x}) \log \mathbf{p}(k | \mathbf{x})} = \log_{\prod_k^K \mathbf{p}(k | \mathbf{x})} \mathbf{p}(y | \mathbf{x}) \quad (392)$$

2. NMAE:

$$NMAE = \frac{\sum_{k=1}^K |\mathbf{p}(k | \mathbf{x}) - \mathbf{q}(k | \mathbf{x})|}{\sum_{j=1}^K \sum_{k=1}^K |\mathbf{p}(k | \mathbf{x}) - \mathbf{q}(y=j | \mathbf{x})|} = \frac{1}{K-1} (1 - \mathbf{p}(y | \mathbf{x})) = \frac{1}{2(K-1)} \cdot MAE \quad (393)$$

显然这只是 MAE 乘上了一个系数

3. NRCE:

$$NRCE = \frac{-\sum_{k=1}^K \mathbf{p}(k | \mathbf{x}) \log \mathbf{q}(k | \mathbf{x})}{-\sum_{j=1}^K \sum_{k=1}^K \mathbf{p}(k | \mathbf{x}) \log \mathbf{q}(y=j | \mathbf{x})} = \frac{1}{K-1} (1 - \mathbf{p}(y | \mathbf{x})) = \frac{1}{A(K-1)} \cdot RCE \quad (394)$$

类似 MAE, 这也只是加上了一个系数

4. NFL:

$$NFL = \frac{-\sum_{k=1}^K q(k | \mathbf{x})(1 - \mathbf{p}(k | \mathbf{x}))^\gamma \log \mathbf{p}(k | \mathbf{x})}{-\sum_{j=1}^K \sum_{k=1}^K \mathbf{q}(y=j | \mathbf{x})(1 - \mathbf{p}(k | \mathbf{x}))^\gamma \log \mathbf{p}(k | \mathbf{x})} = \log_{\prod_k^K (1 - \mathbf{p}(k | \mathbf{x}))^\gamma \mathbf{p}(k | \mathbf{x})} (1 - \mathbf{p}(y | \mathbf{x}))^\gamma \mathbf{p}(y | \mathbf{x}) \quad (395)$$

注意这四个 loss 只是 proof-of-concept, 完全可以用其他 loss.

**Lemma 48.1** In a multi-class classification problem, any normalized loss function  $\mathcal{L}_{norm}$  is noise tolerant under symmetric (or uniform) label noise, if noise rate  $\eta < \frac{K-1}{K}$ .

**Lemma 48.2** In a multi-class classification problem, given  $R(f^*) = 0$  and  $0 \leq \mathcal{L}_{norm}(f(x), k) \leq \frac{1}{K-1}, \forall k$ , any normalized loss function  $\mathcal{L}_{norm}$  is noise tolerant under asymmetric (or class-conditional) label noise, if noise rate  $\eta_{jk} < 1 - \eta_y$ .

对于48.2,  $R(f^*) = 0$  并不一定能达到, 由于类别并不一定可分. 但是经验上效果也很好.

### 48.1 Robustness Alone is not Sufficient

**Remark** 然而 robustness 并不是代表实际好用. 注意, 同 minimizer 的 robustness 显著弱于无偏估计性.

注意到实验上, 正则化的 loss 虽然不会随训练时间增加 generalization gap, 但是性能很差, 差于为正则化的版本, 暗示了 underfitting 的问题.

定义 APL

$$\mathcal{L}_{APL} = \alpha \cdot \mathcal{L}_{Active} + \beta \cdot \mathcal{L}_{Passive} \quad (396)$$

其中两个 loss 都应该是 noise robust 的. 可以证明若如此, 则 APL 也是 noise robust 的.  
为什么 robust loss underfit? 考虑 NCE, 由于最大化分子 (分布的熵)

$$Q = - \sum_{k \neq \nu} \log(p_k) \quad (397)$$

也会导致 loss 变小, 所以容易让学习器学到无意义的预测. 这可能妨碍了收敛并且 underfit.  
为什么 APL 可以解决 underfit? 根据定义, passive loss 显式地让  $Q$  减少, 来控制上述问题.  
类似地, 对 GCE 使用正则化

$$NGCE = (1 - p(y | x)^\rho) / \left( K - \sum_{k=1}^K p(k | x)^\rho \right) \quad (398)$$

## 49 A Survey of Label-noise Representation Learning(LNRL): Past, Present and Future

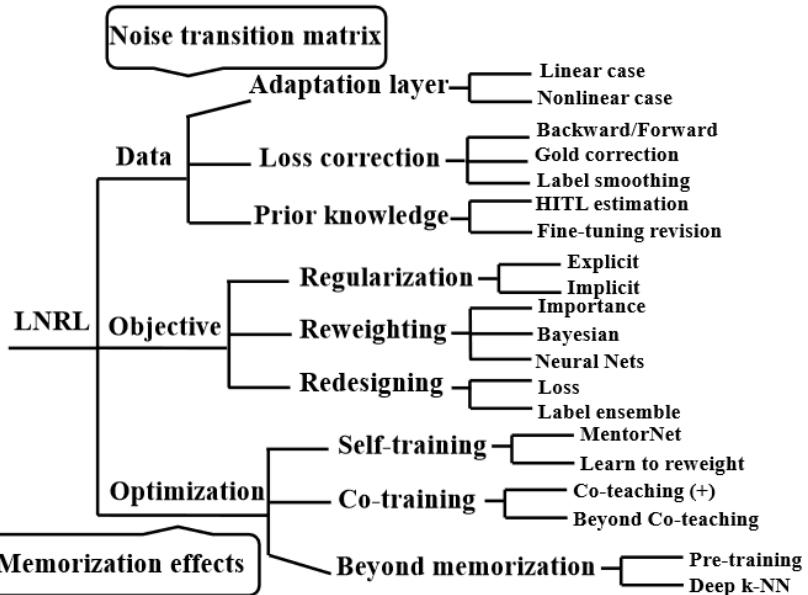


Fig. 2. A taxonomy of LNRL based on the focus of each method. For each technique branch, we list a few representative works here.

**Notations** 包含了损坏数据集上的训练集和 authentic 的测试数据集  $\mathcal{D} = \{\mathcal{D}^{\text{tr}}, \mathcal{D}^{\text{te}}\}$   
相关问题

- 半监督学习 (SSL)
- 正标签学习 (PUL, Positive-unlabeled Learning): 只有 positive-label 和无标签数据
- 互补学习 (CL, Comple. Learning): 带有不属于某类的标签
- 无标签学习 (UUL, Unlabeled-unl. Learning): 有两个不同 prior 的数据类, 没有标签.

LN 下的风险  $R_{\ell,D}(f_\theta) := \mathbb{E}_{(x,y) \sim D} [\ell(f_\theta(x), y)]$  经验风险  $\widehat{R}_{\tilde{\ell},\bar{D}}(f_\theta) := 1/N \sum_{i=1}^N \tilde{\ell}(f_\theta(x_i), \bar{y}_i)$  其中  $\tilde{\ell}$  是 LN-robust 的损失函数.

### 理论 (玄学) 分析

1. 对于数据, 发现噪声转移模式, 将干净/噪声的类别后验联系. 估计噪声转移矩阵  $T$
2. 对于 loss func., 找到 LN-robust 的损失函数. 训练一个 robust 的 classifier.
3. 对于优化策略, 需要关注

## 49.1 Perspective of Data

存在实例无关的噪声分布, e.g.  $p(\bar{Y} | Y, X)$ , 或者实例相关分布  $p(\bar{Y} | Y)$ . 若噪声是实例相关, 学习  $T(X)$  可能是 ill-posed 的, 噪声转移矩阵可能是不可区分的, 所以只考虑实例无关噪声.

此时噪声转移矩阵  $T$  近似建模了 label corrupt. 的过程. 说  $x^i$  是 anchor point, 若  $p(Y = e_i | x^i) = 1$ , 此时可以得到转移矩阵

$$\begin{aligned} p(\bar{Y} = e_j | x^i) &= \sum_{k=1}^C p(\bar{Y} = e_j | Y = e_k, x^i) p(Y = e_k | x^i) \\ &= p(\bar{Y} = e_j | Y = e_i, x^i) p(Y = e_i | x^i) \\ &= p(\bar{Y} = e_j | Y = e_i, x^i) = T_{ij} \end{aligned} \tag{399}$$

anchor points 肯能难以找到, 使用简单近似  $x^i = \arg \max_x p(\bar{Y} = i | x)$

转移矩阵很重要, 可以用  $T^{-1}$  来转换 noisy/clean 类别后验

$$p(Y | x) = T^{-1} p(\bar{Y} | x) \tag{400}$$

## 49.2 Perspective of Statistics/Learning Theory

$R^* = R_D(f^*)$  是 clean data 上的最优分类器的 risk,  $\hat{f} = \arg \min_{f \in \mathcal{H}} \widehat{R}_{\tilde{\ell},\bar{D}}(f)$  是找到的最好分类器,  $L_\rho$  是  $\tilde{\ell}$  的 Lipschitz 常数. 假设权值矩阵的 F 范数分别不超过  $M_1, \dots, M_d$ , 且  $\|x\| \leq B, \forall x$ , 存在不减函数  $\xi_\ell$ ,  $\xi_\ell(0) = 0$ , s.t. 以  $1 - \delta$  的概率, 若  $\ell$  是 classification-calibrated<sup>17</sup>

$$\begin{aligned} R_D(\hat{f}) - R^* &\leq \xi_\ell \left( \min_{f \in \mathcal{H}} R_{\ell,D}(f) - \min_f R_{\ell,D}(f) \right. \\ &\quad \left. + 4L_\rho \mathcal{R}(\mathcal{H}) + 2\sqrt{\log(1/\delta)/2N} \right) \\ &\leq \xi_\ell \left( \min_{f \in \mathcal{H}} R_{\ell,D}(f) - \min_f R_{\ell,D}(f) + 4L_\rho C + 2\sqrt{\log(1/\delta)/2N} \right) \end{aligned} \tag{401}$$

其中 where  $C = B(\sqrt{2d \log 2} + 1) \prod_{i=1}^d M_i / \sqrt{N}$  and  $R_D(f) = \mathbb{E}_{(x,y) \sim D} \left[ \mathbb{1}_{\{\text{sign}(\hat{f}(x)) \neq y\}} \right]$  denotes the risk of  $\hat{f}$  w.r.t. the 0-1 loss. 注意对于 DNN, Rademacher 复杂度  $\mathcal{R}(\mathcal{H})$  被上界  $C$  限制.

**Remark** 这个结论说明了给出足够复杂的假设空间和足够多的数据 (i.e. N), 可以让有噪声的学到的  $\hat{f}$  趋近于统计最优分类器.<sup>18</sup>

<sup>17?</sup>

<sup>18</sup> 然而存在不减函数  $\xi_\ell$ ,  $\xi_\ell(0) = 0$  这一外部包装显然让这个 conclusion 不那么 convincing.

### 49.3 Perspective of Opt. Policy

考虑经典的 early-stopping 策略<sup>19</sup>. 假设初始权值矩阵按照标准高斯分布  $\mathcal{N}(0, 1)$  初始化, 并且经过学习率为  $\eta$  的 SGD 更新  $\tau$  步, 则有若  $\varepsilon_0 \leq \delta\lambda(C)^2/K^2$  且  $\rho \leq \delta/8$  则经过  $I \propto \|C\|^2/\lambda(C)$  步, 有大概率成立: 模型  $W^I$  对于关于  $\text{cluste}\{c_k\}_{k=1}^K$  所有  $\varepsilon_0$  邻域中的点, 都会预测相同的 label. 并且若  $y(x) = \arg \min_l |f_{W^I}(x) - \alpha_l|$ , where  $\{\alpha_l\}_{l=1}^K \in [-1, 1]$  是每一类的 labels. 那么对于所有训练样本, 有训练后的矩阵关系

$$\|W^\tau - W^0\|_F \lesssim \left( \sqrt{K} + \tau \varepsilon_0 K^2 / \|C\|^2 \right) \quad (402)$$

其中  $0 \leq \tau \leq I$  and  $A \lesssim B$  denotes  $A \leq \beta B$  for some  $\beta$

**Remark** 这说明了在 early stopping 下权值不会离初始值太远. 直觉上, DNN 随着 epoch 增加会 overfit 数据, early stopping 肯定能改善情况. 可能 robust 的 weight 离初始权值不远.

### 49.4 Taxonomy

1. Data: 给定噪声转移矩阵  $T$ (的估计), 给出变换后的损失  $\ell \xrightarrow{T} \tilde{\ell}$
2. Objective: 可以通过增强项来得到新的目标函数. 比如增加辅助正则项. 隐式正则化 (e.g. soft/hard bootstrapping, virtual adversarial training(VAT)). 重新给目标函数分权重.
  - $\tilde{\ell} = \ell + r$
  - $\bar{\ell} = \sum_i w_i \ell_i$
  - $\tilde{\ell}$  has a special format  $\ell'$  independent of  $\ell$

### 49.5 Data: Noise Trans. Mat.

**Definition 49.1** (*Noise transition matrix*) Suppose that the observed noisy label  $\bar{y}$  is drawn independently from a corrupted distribution  $p(X, \bar{Y})$ , where features are intact. Meanwhile, there exists a corruption process, transition from the latent clean label  $y$  to the observed noisy label  $\bar{y}$ . Such a corruption process can be approximately modeled via a noise transition matrix  $T$ , where  $T_{ij} = p(\bar{y} = e_j | y = e_i)$

两种经典的噪声转移矩阵, 对称 flipping/配对 flipping

$$\begin{bmatrix} 1 - \tau & \frac{\tau}{n-1} & \cdots & \frac{\tau}{n-1} \\ \frac{\tau}{n-1} & 1 - \tau & & \frac{\tau}{n-1} \\ \vdots & \ddots & \ddots & \vdots \\ \frac{\tau}{n-1} & \frac{\tau}{n-1} & \cdots & 1 - \tau \end{bmatrix} \quad \begin{bmatrix} 1 - \tau & \tau & 0 & 0 \\ 0 & 1 - \tau & \tau & 0 \\ \vdots & & \ddots & \vdots \\ 0 & & & \tau \\ \tau & 0 & \cdots & 1 - \tau \end{bmatrix} \quad \begin{array}{l} (\text{a) Sym-flipping}) \\ (\text{b) Pair-flipping.}) \end{array}$$

其中  $\tau$  是噪声比例. 在 LNL(label-noise learning), 常常使用这些方法来模拟 LN. 当然现实中很有可能是不规则的 noise(如淘宝衣物数据集 Clothing 1M).

<sup>19</sup>M. Li, M. Soltanolkotabi, and S. Oymak, “Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks,” in AISTATS, 2020

## 49.6 Data: Adapt. Layer

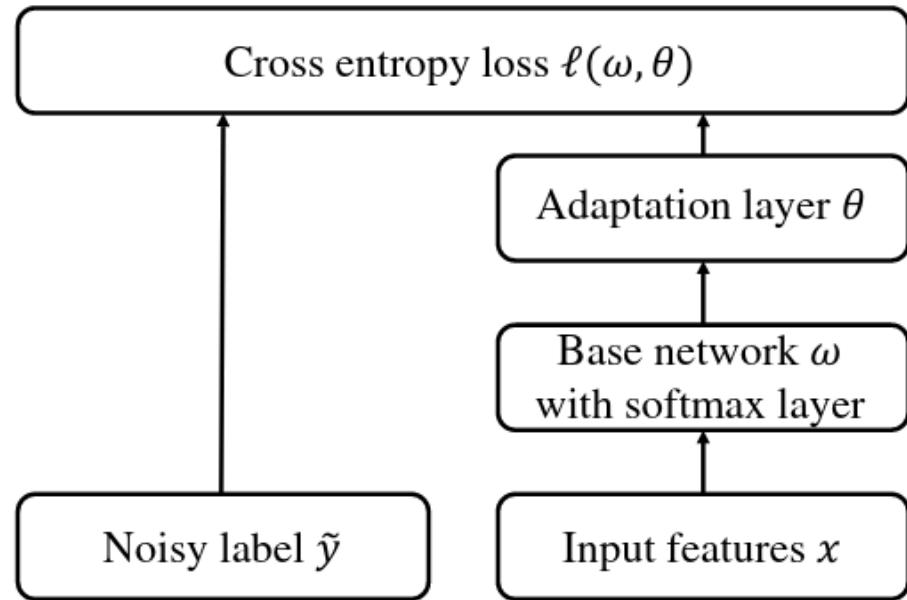


Fig. 4. A general case of adaptation layer.

使用 end-to-end 模型估计转移矩阵  $T$ .

线性模块:(Sukhbattar, 2015) 提出了在网络输出之后增加一个由  $T$  参数化的 adapt. layer. 单独用 CE 优化两个不同的模块并不能达到 optimal 的  $T$ . 他们又增加了一个  $T$  的正则化项, trace norm/ridge regression.

非线性模块: (Goldberger et al., 2017) 使用了 base model param. by  $\omega$ , 以及噪声模型 param. by  $\theta$ . 既然 base model 的输出是 hidden 的, 那么他们用 EM 算法来估计隐藏输出 (E-step), 以及当前的参数 (M-step). EM 也会导致局部最优和可伸缩性的问题. 他们还提出了 c-model: 使用 latent true label 和输入特征来训练; s-model: 使用 latent true label only 训练 (noisy label).

## 49.7 Loss Correction: Backward/Forward Correction

(Patrini et al., 2017) 前向后向矫正. 后向通过在 loss 上乘上  $T^{-1}$ , 前向通过将网络输出乘上  $T$ .

**Theorem 49.2** (后向校正) 假设转移矩阵是非奇异的. 后向矫正定义为

$$\ell^{\leftarrow}(f(x), \bar{y}) = [T^{-1} \ell_{y|f(x)}]_{\bar{y}} \quad (403)$$

其中

$$\ell_{y|f(x)} = (\ell(f(x), 1), \dots, \ell(f(x), k)) \quad (404)$$

那么矫正后的 loss 是 unbiased, 即

$$\mathbb{E}_{\bar{y}|x} \ell^{\leftarrow}(f(x), \bar{y}) = \mathbb{E}_{y|x} \ell(f(x), y), \forall x \quad (405)$$

**Theorem 49.3** (前向矫正) 假设转移矩阵是非奇异的. 前向矫正定义为

$$\ell^{\rightarrow}(f(x), \bar{y}) = [\ell_{y|T^\top} f(x)]_{\bar{y}} \quad (406)$$

那么最小化前向矫正 loss 和最小化原 loss(*under clean data*) 同解

$$\arg \min_f \mathbb{E}_{x, \bar{y}} \ell^{\rightarrow}(f(x), \bar{y}) = \arg \min_f \mathbb{E}_{x, y} \ell(f(x), y) \quad (407)$$

注意这比前一个定理的无偏性弱.

一般来说,  $T$  是未知的, 需要被估计. 于是 Patrini et al 提出一个两阶段训练. 首先使用 noisy data 训练网络, 再获得一个  $T$  的估计, 再重新训练网络, 使用  $T$  校正的 loss.

## 49.8 Loss Correction: Gold Correction

(Hendrycks et al., 2018) 提出了 Gold 校正来处理严重噪声. 关键思路时, 假设一部分数据是可信的且可用的, 比如有一些专家来得出的 trusted set  $D$ . 他们使用  $D$  来估计  $T$ , 再用前向矫正来训练 DNN, 这就是 GLC. 具体来讲

$$\hat{T}_{ij} = \frac{1}{A_i} \sum_{x \in A_i} \hat{p}(\bar{Y} = e_j | Y = e_i, x) \quad (408)$$

其中  $A_i$  是 label 为  $i$  的  $D$  的子集.

## 49.9 Loss Correction: Label Smoothing

使用 label-smoothing 技术. 这也是一种 loss correction 方法.

$$\ell^{\text{SM}}(f_\theta(X), Y) = M \cdot \ell(f_\theta(X), Y) \quad (409)$$

其中  $M$  是 smearing matrix.

- $M = I$ , 标准训练
- $M = (1 - \alpha)I + \frac{\alpha E}{L}$ , label smoothing.
- $\frac{1}{1-\alpha} \cdot (I - \frac{\alpha \cdot J}{L})$ , 对称噪声下的后向校正. 其中  $M \sim T^{-1}$ .

## 49.10 Prior Knowledge: Human-in-the-Loop Estimation

(Han et al., 2018) 提出了分离  $T$  的结构和具体数值, 通过 human-assisted approach "masking". 那么怎么把这些先验结构放到模型中呢, 通过一个生成模型.

$t \sim p(t)$  通过生成模型  $t_o \sim p(t_o)$  来获得. 似乎是基于 flow 的  $p(t_o) = p(t) \frac{dt}{dt_o} \Big|_{t_o=f(t)}$ . 最终通过 evidence lower bound opt. (ELBO) 来优化.<sup>20</sup>

<sup>20</sup>(来自原文)

$$\ln P(\tilde{y} | x) \geq \mathbb{E}_{Q(s)} [\underbrace{\ln \sum_y P(\tilde{y} | y, s) P(y | x)}_{\text{previous model}} - \ln \left( Q(s_o) / \underbrace{P(s_o)}_{\text{structure prior}} \right) \Big|_{s_o=f(s)}], \quad (410)$$

## 49.11 Prior Knowledge: Fine-tuning Revision

(Xia et al., 2019) 介绍了 transition-revision 方法: Reweighting T-Revision. 他们首先使用数据里的 anchor points 的相近点 (高混淆概率) 来初始化 T. 之后把这个初始化的 T 作为先验, 然后使用松弛变量来进行 fine-tune.

具体上讲, 先使用 noisy data 学习 (没有 adaptation layer)  $\hat{p}(\bar{Y} | X = x)$ . 再初始化  $\hat{T}$ , 其中 anchor points 是  $\hat{p}(\bar{Y} = e_i | X = x)$  最高的点.<sup>21</sup>

再第二个阶段, 神经网络使用 noisy adaptation layer  $\hat{T}^\top$  并且最小化 weighted loss. 得到并加上一个松弛变量  $\Delta T$ . 重复两阶段直到收敛 (在验证集上达到最小误差).<sup>22</sup>

**Remark** 某种形式上和 EM 类似.

## 49.12 Regularization: Explicit Regularization

(Azadi et al., 2016) 提出了一种正则化项  $\Omega_{\text{aux}}(w) = \|Fw\|_g$  其中  $\|\cdot\|_g$  是 group norm,  $F^\top = [X_1, \dots, X_n]$ ,  $X_i = \text{Diag}(\mathbf{x}_i)$ , 鼓励稀疏性. 这会鼓励一小部分 clean data 来 control model. (Berthelot et al., 2019) 提出了 MixMatch 来进行 SSL, 并且达到了 SOTA. 其中的一个关键部分是 Minimum Ent. Reg.(MER), 也是一种显式正则化. MER 提出于 (Grandvalet & Bengio, 2005), 关键 idea 是把 CE 加入一个正则项, 鼓励在 unlabeled data 上给出 high-confidence 的输出, 具体地, 最小化在 unlabeled 数据上的熵.

类似于 MER, psedo-label 方法 (D.-H. Lee, 2013)(i.e. label guessing) 进行隐式的 ent. 最小化. 具体上讲, 首先计算模型 (通过各种 augmentation) 预测的类型分布, 再通过 temperature sharpening func. 来最小化 label dist. 的熵.

(Miyato et al., 2018) 提出了一个 virtual adversarial loss, 是一种新的衡量条件标签分布 (i.e.,  $p(y|x)$ ) 的局部平滑度的方法. 具体地, 他们要求输出分布在输入点附近各向同性地平滑, 通过把模型在其最各向异性的方向上选择性地平滑. 使得模型对输入不敏感. 为了实现, 设计了 virtual adversarial direction.

定义 LDS(local dist. smooth. )

$$\text{LDS}(x^*, \theta) := D \left[ p \left( y | x^*, \hat{\theta} \right), p \left( y | x^* + r_{\text{vadv}}, \hat{\theta} \right) \right] \quad (414)$$

使用 thresholding binarization 中的算子 tempered sigmoid

$$f(s) = \frac{1}{1 + \exp \left( -\frac{s-\alpha}{\beta} \right)}, \quad \text{where } \alpha \in (0, 1), \beta \ll 1 \quad (411)$$

<sup>21</sup>注意如果  $P(Y = i | X = x) = 1$ ,

$$P(\bar{Y} = j | X = x) = \sum_{k=1}^C T_{kj} P(Y = k | X = x) = T_{ij} \quad (412)$$

<sup>22</sup>关于 loss, 使用的权重为  $\hat{P}(Y = y | X = x)/\hat{P}(\bar{Y} = y | X = x)$  最终的 loss 为

$$\bar{R}_{n,w}(T, f) = \frac{1}{n} \sum_{i=1}^n \frac{g_{\bar{Y}_i}(X_i)}{(T^\top g)_{\bar{Y}_i}(X_i)} \ell(f(X_i), \bar{Y}_i) \quad (413)$$

其中  $f(X) = \arg \max_{j \in \{1, \dots, C\}} g_j(X)$

<sup>2324</sup> 其中  $r_{\text{vadv}} := \arg \max_{\|r\|_0 < \epsilon} D \left[ p \left( y \mid x^*, \hat{\theta} \right), p \left( y \mid x^* + r \right) \right]$  是 virtual adversarial direction, 然后有

$$R_{\text{vadv}} (D_l, D_{ul}, \theta) := \frac{1}{N_l + N_{ul}} \sum_{x^* \in D_l, D_{ul}} \text{LDS}(x^*, \theta) \quad (419)$$

其中  $D_l, D_{ul}$  分别是有无标签的数据, final loss:

$$\ell(D_l, \theta) + \alpha R_{\text{vadv}} (D_l, D_{ul}, \theta) \quad (420)$$

### 49.13 Objective Regularization: Implicit

最近出现了很多隐式正则化方法.

(Reed et al., 2015) 的 Bootstrapping. 学习器和自己 bootstrap, 使用 label 和模型目前的 prediction 的凸组合来生成训练目标. 直觉上, 随着 learner 学习, predictions 也变得可信. 进而避免对 noise 的直接建模. 具体地, 有 soft/hard 两种 bootstrapping. 对于 soft bootstrapping, 使用预测的类概率  $q$  来得到回归目标.

$$\ell_{\text{soft}} (q, t) = \sum_{k=1}^L [\beta t_k + (1 - \beta) q_k] \log (q_k) \quad (421)$$

这等价于 softmax regression + MER.

对于 hard bootstrapping, 使用  $q$  的 MAP 估计.

$$\ell_{\text{hard}} (q, t) = \sum_{k=1}^L [\beta t_k + (1 - \beta) z_k] \log (q_k) \quad (422)$$

其中  $z_k = \mathbf{1}[k = \arg \max_{i=1, \dots, L} q_i]$  为了能够优化 hard 版本, 需要使用 EM-like 算法. E-step 中计算凸组合的 targets, M-step 根据 targets 进行优化参数.

<sup>23</sup>注意, 原文中是 SSL, 对于有标签的数据可以使用 ground truth  $q(y \mid x_*)$  来进行计算. 对于 LNRL 显然 clean data 不是 available 的, 直接使用模型的输出.

<sup>24</sup>Fast Approx. of  $r_{\text{vadv}}$  定义  $D(r, x^*, \theta) := D \left[ p \left( y \mid x^*, \hat{\theta} \right), p \left( y \mid x^* + r, \theta \right) \right]$  由于在  $r = 0$  时,  $\nabla_r D(r, x, \hat{\theta}) \Big|_{r=0} = 0$ , 那么有二阶估计

$$D(r, x, \hat{\theta}) \approx \frac{1}{2} r^T H(x, \hat{\theta}) r \quad (415)$$

那么  $r_{\text{vadv}}$  可以是 Hessian 的第一个 dominant eigenvector 具有长度  $\epsilon$

$$\begin{aligned} r_{\text{vadv}} &\approx \arg \max_r \left\{ r^T H(x, \hat{\theta}) r; \|r\|_2 \leq \epsilon \right\} \\ &= \overline{\epsilon u(x, \hat{\theta})}, \end{aligned} \quad (416)$$

接着有  $O(I^3)$  的找到如此 eigenvect. 的算法, 通过幂迭代和有限差分法.

随机采样一个 unit vector  $d$ , 迭代计算 mat-vec prod.  $d \leftarrow \bar{H}d$ .

为了避免直接计算  $H$ , 使用有限差分法来估计这个乘积

$$\begin{aligned} \bar{H}d &\approx \frac{\nabla_r D(r, x, \hat{\theta}) \Big|_{r=\xi d} - \nabla_r D(r, x, \hat{\theta}) \Big|_{r=0}}{\xi} \\ &= \frac{\nabla_r D(r, x, \hat{\theta}) \Big|_{r=\xi d}}{\xi} \end{aligned} \quad (417)$$

i.e.,

$$d \leftarrow \overline{\nabla_r D(r, x, \hat{\theta}) \Big|_{r=\xi d}} \quad (418)$$

他们发现一步迭代就能达到类似 FGSM 里的估计精度.

(Zhang et al., 2018) 的 Mixup, 启发于 vicinal risk minimization(O. Chapelle, J. Weston, L. Bottou, and V. Vapnik, 2000).

(Han et al., 2020) 的 SIGUA(data-agnostic). 注意到随着网络容量的提升, 网络能逐渐地 overfit noisy data. 所以他们提出了 Stochastic Integrated Gradient Underweighted Ascent(SIGUA) 的一种训练策略, 在一个 mini-batch 中, 线照常使用 SGD, 再在 bad-data 上使用 (lr 递减的) 梯度递增. 在训练哲学上, SIGUA 让网络忘记不想要的记忆, 来更好的加强想要的记忆.

#### 49.14 Objective Reweighting: Importance Reweighting

(Liu and Tao, 2015) 使用 importance reweighting 来 LNL. 将 noisy data 作为 source domain, clean data 作为目标 domain. Idea 是重写经验风险 w.r.t. clean data

$$\begin{aligned}
 R(f) &= \mathbb{E}_{(X,Y) \sim D} [\ell(f(X), Y)] \\
 &= \int_x \sum_i p_D(X = x, Y = i) \ell(f(x), i) dx \\
 &= \int_x \sum_i p_{\bar{D}}(X = x, \bar{Y} = i) \frac{p_D(X=x, \bar{Y}=i)}{p_D(X=x, \bar{Y}=i)} \ell(f(x), i) dx \\
 &= \int_x \sum_i p_{\bar{D}}(X = x, \bar{Y} = i) \frac{p_D(\bar{Y}=i|X=x)}{p_D(\bar{Y}=i|X=x)} \ell(f(x), i) dx \\
 &= \mathbb{E}_{(X,\bar{Y}) \sim \bar{D}} [\beta(X, \bar{Y}) \ell(f(X), \bar{Y})],
 \end{aligned} \tag{423}$$

倒数第二个方程成立, 因为 noise 和 instance 独立. 最后的  $\beta(X, \bar{Y}) = p_D(\bar{Y} = i | X = x) / p_{\bar{D}}(\bar{Y} = i | X = x)$  就是 IW. 这可以通过转移矩阵 T 或者使用小数据集的 clean data (like GLC) 来学到.

#### 49.15 Objective Reweighting: Bayesian Methods

(Wang et al., 2017) 的 reweighted prob. models(RPM) 来应对 label noise. Idea 在于, 降低 bad labels 的权重且增加 clean labels 的权重. 具体地,

- 定义概率模型  $p_{\beta}(\beta) = \prod_{n=1}^N \ell(y_n | \beta)$
- 给出 latent weight 的先验分布  $p_w(w), w = (w_1, \dots, w_N)$

$$p(y, \beta, w) = 1/z \cdot p_{\beta}(\beta) p_w(w) \prod_{n=1}^N \ell(y_n | \beta)^{w_n} \tag{424}$$

- 推理  $\beta, w$ , 通过后验分布  $p(\beta, w | y)$ . 先验分布  $p_w(w)$  可以是 Beta 分布, scaled Dirichlet 分布, Gamma 分布. 不同的选择 trade off 小概率的 cases.

(Arazo et al., 2019) 使用了两组分 beta mixture model(BMM), 视为 clean-noisy 的混合, 使用了一个 bootstrapping loss. 具体地, 使用 dyn-w. 的 bootstrapping loss. 数学上, 定义 loss 上的 pdf

$$p(\ell) = \sum_{k=1}^K \lambda_k p(\ell | k) \tag{425}$$

并且  $p(\ell | k)$  可以使用 Beta 分布建模.

$$p(\ell | \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \ell^{\alpha-1} (1 - \ell)^{\beta-1} \tag{426}$$

上述问题可以使用 EM 算法来解决.

更具体地, 引入  $\gamma_k(\ell) = p(k | \ell)$ , E-step 中固定  $\lambda, \alpha, \beta$ , 计算  $\gamma$ . M-step 中固定  $\gamma$ , 使用带权动量估计  $\alpha, \beta$ , 动态权重则使用简单的方法来得到  $\lambda_k = \frac{1}{N} \sum_{i=1}^N \gamma_k(\ell_i)$ . 基于这个 BMM 模型, 他们还提出了动态 hard/soft bootstrapping loss, 其中每个 sample 的 weight 动态的设置为  $p(k = 1 | \ell_i)$ (sample 为 clean 的概率).

### 49.16 Objective Reweighting: NNs

(Shu et al., 2019) 使用 Meta-Weight-Net(MW-Net) 来学习显示的 weighting function. w. func. 是一个单层 MLP, 从 loss 到 weight. 数学上

$$w^*(\theta) = \arg \min_w \ell^{\text{tr}}(w; \theta) = 1/N \sum_{i=1}^N \mathcal{V}(t_i^{\text{tr}}(w); \theta) \ell_i^{\text{tr}}(w) \quad (427)$$

这里, 可以使用元学习来优化 MW-Net: 给出一些 clean, balanced 元数据  $\left\{x_i^{(\text{meta})}, y_i^{(\text{meta})}\right\}_{i=1}^M$ , 最小化 meta-loss

$$\theta^* = \arg \min_{\theta} \ell^{\text{meta}}(w^*(\theta)) = 1/M \sum_{i=1}^M \ell_i^{\text{meta}}(w^*(\theta)). \quad (428)$$

使用 SGD 迭代的分别更新  $w$  和  $\theta$

### 49.17 Objective Redesigning

这些方法常常是 scenario-specific 的.

### 49.18 Objective Redesigning: Loss Redesign

最近有许多新的对抗 label noise 的 loss. 基于诸如梯度裁剪/课程学习的原则.

(Zhang et al., 2018) 提出了推广的 CE, 使用 MAE(mean absolute error/l1-norm)+CCE(categorical CE). 数学上, 他们使用 Box-Cox 变换作为  $\ell_q$  loss func.

$$\ell_q(f(x), e_j) = (1 - f_j(x)^q) / q \quad (429)$$

注意  $q \rightarrow 0$  为 CCE,  $q \rightarrow 1$  为 MAE. 并且引入了截断  $\ell_q$  损失的估计

$$\ell_{\text{trunc}}(f(x), e_j) = \begin{cases} \ell_q(k) & \text{if } f_j(x) \leq k, \\ \ell_q(f(x), e_j) & \text{otherwise} \end{cases} \quad (430)$$

并且  $\ell_q(k) = 1 - k^q/q$

(Charoenphakdee et al., 2019) 分析了对称 loss. Idea 是设计 loss 并不一定是对称的, 同时给不对称的区域的惩罚, i.e.,  $\ell(z) + \ell(-z) = C$ . 所以他们给出了一个 barrier hinge loss

$$\ell(z) = \max(-b(r+z) + r, \max(b(z-r), r-z)) \quad (431)$$

不对称, 并且在对称区外给出惩罚.

(Thulasidasan et al., 2019) 提出了放弃一些 confusing data 的方法. 他们的网络 DAC(deep abstaining classi.) 有一个额外输出  $p_{k+1}$ , 是放弃概率. 具体的 loss 是

$$\ell(x_j) = -\tilde{p}_{k+1} \sum_{i=1}^k t_i \log(p_i/\tilde{p}_{k+1}) - \alpha \log \tilde{p}_{k+1} \quad (432)$$

其中  $\tilde{p}_{k+1} = 1 - p_{k+1}$  他们使用了一个动态的调整  $\alpha$  的算法. DAC 可以作为 data cleaner 使用在不论有无结构的噪声上.

(Aditya et al., 2020) 使用梯度裁剪来设计 loss. 直觉上, 这防止了过于自信的递降. 基于梯度裁剪, 他们设计了部分 Huberized loss

$$\tilde{\ell}_\theta(x, y) = \begin{cases} -\tau p_\theta(x, y) + \log \tau + 1 & \text{if } p_\theta(x, y) \leq \frac{1}{\tau} \\ -\log p_\theta(x, y) & \text{otherwise} \end{cases} \quad (433)$$

(Lyu et al. 2020) 提出 curriculum loss. 是 0-1 loss 更紧的上界. 此外, CL 可以动态地选择 samples. 对于任何 base loss 是 0-1 loss 的上界  $\ell(u) \geq \mathbf{1}(u < 0)$ ,  $u \in \mathbb{R}$ , CL 定义为

$$Q(\mathbf{u}) = \max \left( \min_{\mathbf{v} \in \{0,1\}^n} f_1(\mathbf{v}), \min_{\mathbf{v} \in \{0,1\}^n} f_2(\mathbf{v}) \right) \quad (434)$$

其中

$$f_1(\mathbf{v}) = \sum_{i=1}^n v_i \ell(u_i) \text{ and } f_2(\mathbf{v}) = n - \sum_{i=1}^n v_i + \sum_{i=1}^n \mathbf{1}(u_i < 0). \quad (435)$$

为了在 DL 中使用 CL, 进一步他们提出了 noise pruned CL.

### 49.19 Label Ensemble

(Laine & Aila, 2017) 引入了 SSL 中的 self-ensembling, 包括  $\pi$ -model 和时序聚合, 也可用于 noise 净化. self-ensembling 的 idea 是, 使用网络的输出对位置标签形成共识. 具体上,  $\pi$ -model 是让不同的 dropout 模式下对同一个输入得到相容的输出. 除了  $\pi$ -model, 多个 epochs 的时序聚合也被考虑.

$\pi$ -model 的 loss 是

$$\ell = -1/B \sum_i \log z_i [y_i] + w(t)/C|B| \sum_i \|z_i - \tilde{z}_i\|^2 \quad (436)$$

其中第一项是 batch 的 CE, 后一项处理无标签数据. 其中  $z_i, \tilde{z}_i$  分别是不同 dropout 下的输出. 第二项也被时间相关函数  $w(t)$  加权.

这个方法的时序集成使用了几次之前的网络求值来进行一个集成预测, 具体上, 在  $\pi$ -model 中的  $\tilde{z}$  是前几次的输出的 moving average

$$Z_i \leftarrow \alpha Z_i + (1 - \alpha) z_i \quad (437)$$

$$\tilde{z}_i = \frac{Z_i}{1 - \alpha^t} \quad (438)$$

**Remark** 既然 kWTA 是更高级的 Dropout, 那么...

(Nguyen et al., 2020) 提出了 self-ensemble label filtering(SELF) 的方法来在训练中逐渐地去除错误 labels. 使用网络在不同训练迭代时的不同输出来得到一个预测的共识, 用于去除 bad labels. 假设  $L_i = \{(y, x) \mid \hat{y}_x = y; \forall (y, x) \in L_0\}$  是 label y 和其最大似然预测,  $L_0$  是最开始的样本集合.

根据他们的策略, 模型首先和 Mean Teacher 进行集成, 后者是一个 exponential running average of model(snapshots). 其次, 使用在多个 epochs 上的预测结果  $\bar{z}_j = \alpha \bar{z}_{j-1} + (1 - \alpha) \hat{z}_j$ , 代表了 moving average of predicted labels.

**Remark** 这和之前那个时序集成挺类似的. 区别是, 不是用不同 dropout 来进行集成而是和 mean teacher 集成.

(Ma et al., 2018) 研究了训练样本上的深度表示子空间的维度. 提出了基于维度的训练策略, 在训练中检测子空间维度. Key idea 是使用 local intrinsic dim.(LID), 用于区分 noisy/clean labels.

数学上

$$\widehat{\text{LID}} = - \left( 1/k \sum_{i=1}^k \log r_i(x)/r_{\max}(x) \right)^{-1} \quad (439)$$

<sup>25</sup> 其中  $r_i(x)$  是 x 到 i-th neighbor 的距离. 根据他们的观察, 在 clean label 上训练的时候 LID 一直下降, 但是在 noisy label 上训练的时候先下降后上升. 可以据此观察训练动态.

## 49.20 Optimization Policies

一个例子是 Early-stopping. 根据记忆效应避免了 overfitting. 存在更好的方法: small-loss trick. MentorNet, Learning to Reweight. Co-teaching, Co-teaching+. cross-validation, automated learning, GMMs.

<sup>25</sup> Extract from *Dimensionality-Driven Learning with Noisy Labels*

原始定义是

$$\text{LID}_F(r) \triangleq \lim_{\epsilon \rightarrow 0} \frac{\ln(F((1+\epsilon)r)/F(r))}{\ln(1+\epsilon)} = \frac{r F'(r)}{F(r)}, \quad (440)$$

以及在点 x 的 LID

$$\text{LID}_F = \lim_{r \rightarrow 0} \text{LID}_F(r). \quad (441)$$

值得注意的是, 对于 DNN 来说, 应该使用如下 in-batch 的 LID, 使用倒数第二层得到网络输出

$$\widehat{\text{LID}}(x, X_B) = - \left( \frac{1}{k} \sum_{i=1}^k \log \frac{r_i(g(x), g(X_B))}{r_{\max}(g(x), g(X_B))} \right)^{-1} \quad (442)$$

据此, batch size 需要比较大.

据此, 提出 dimensionality-driven training, 具体地, 使用 LID 动态矫正 labels

$$y^* = \alpha_i y + (1 - \alpha_i) \hat{y} \quad (443)$$

其中

$$\alpha_i = \exp \left( -\frac{i}{T} \frac{\widehat{\text{LID}}_i}{\min_{j=0}^{i-1} \widehat{\text{LID}}_j} \right) \quad (444)$$

校准因子是递增的, 代表随着训练进行, noisy label 变得越来越不可信. 可以直接使用 augmented 的标签的熵作为 loss

$$\mathcal{L} = -\frac{1}{N} \sum_{n=1}^N \sum_{y_n^*} y_n^* \log P(y_n^* | x_n) \quad (445)$$

技术上, 设置 w 个 epochs 的初始化窗口. Turning point 是, LID 高于前 w 个 epochs 的 mean 两个标准差时. 此时 rollback. CSR(Arpit et al., 2017) 为新的衡量假设空间复杂性的 measure.

### 49.20.1 Memorizaiton Eff.

记忆效应指出, NNs 趋向于先记住容易学习的模式 (clean), 再逐渐 overfit 难以学习的模式 (noisy). 这启发了 small-loss trick. 具体上, 这是说把 small-loss 的样本作为 clean samples, 并且只在这些样本上进行模型参数更新. 数学上, 等价于建立一个 masked loss

$$\tilde{\ell} = \text{sort}(\ell, 1 - \tau) \quad (446)$$

其中  $\tau$  时 noise rate.

### 49.20.2 Self-training

(Jiang et al., 2018) 提出了 MentorNet, 监督称为 StudentNet 的基网络. 数学上, MentorNet  $g_m$  估计一个 predefined curriculum

$$\arg \min_{\theta} \sum_{(x_i, y_i) \in \mathcal{D}} g_m(z_i; \theta) \ell_i + G(g_m(z_i; \theta); \lambda_1, \lambda_2), \quad (447)$$

第一项是课程加权的 loss, 第二项是正则化惩罚项. 可以得到闭形式解

$$g_m(z_i; \theta) = \begin{cases} \mathbf{1}(\ell_i \leq \lambda_1) & \text{if } \lambda_2 = 0 \\ \min(\max(0, 1 - \ell_i - \lambda_1/\lambda_2), 1) & \text{if } \lambda_2 \neq 0 \end{cases} \quad (448)$$

直觉上, 当  $\lambda_2 = 0$  时, M-Net 只会使用 losses 小于  $\lambda_1$  的 sample. 否则 MentorNet 不会提供大于  $\lambda_1 + \lambda_2$  的样本. 同时这也使 MNet 能够自行发现 curriculum.

(Ren et al., 2018) 使用 meta-learning 来给不同的 samples 不同的权重, 基于他们的梯度方向. 基本上, 小 loss 数据分配更多权重, 并且他们相信在小的验证集上最小化 loss 的权重是最好的权重, 具体上, 在每个 iteration 之前在验证集上决定样本权重. 数学上, 最小化带权损失

$$\theta^*(w) = \arg \min_{\theta} \sum_{i=1}^N w_i \ell_i(\theta) \quad (449)$$

他们在验证机上进行参数选择

$$w^* = \arg \min_w 1/M \sum_{i=1}^M \ell_i^v(\theta^*(w)) \quad (450)$$

具体上有三步

1. 将 noisy data 进行前向/后向传播, 得到 training net. 的新的参数  $\theta$  和梯度  $\nabla \theta$
2. 梯度  $\nabla \theta$  影响 validation net.
3. training net. 使用 meta-leanrning 来更新  $w$ , 通过二次求导/二阶导.

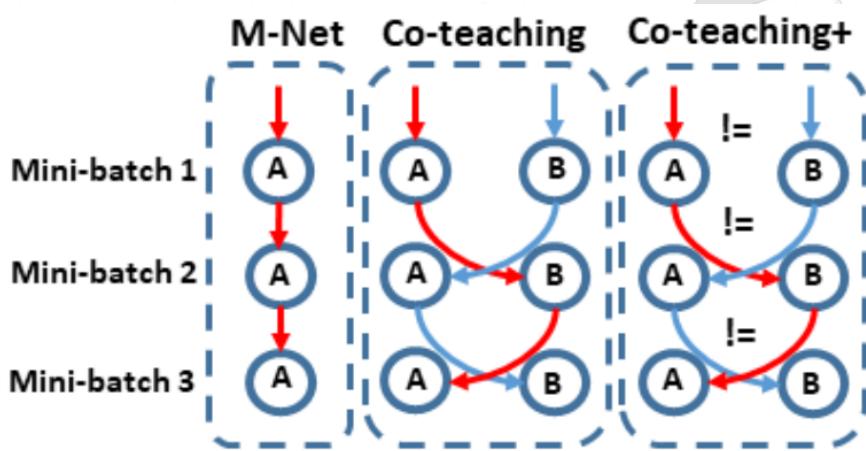


Fig. 6. Self-training (i.e., MentorNet, abbreviated as M-Net) vs. Co-training (i.e., Co-teaching and Co-teaching+).

#### 49.20.3 Co-training

(Han et al., 2018) 提出了 co-teaching, 让两个 DNN 在每个 batch 上互相 teach. 具体地, 每个网络前向传播所有数据, 然后选出 clean data, 再根据 peer net. 选出的 clean data 来进行参数更新. 在 MentorNet 中误差会不断累积, 但是在 co-teaching 中, 两个模型能够学到不同类型的 noise, 降低误差的累积, 最终达到一个共识. 注意集成学习中关键的一点是让两个学习器有差异.

(Yu et al., 2019) 提出了 co-teaching+, “通过不同意来更新”, 来保持 coteaching 中两个网络的差异性. 包含 disaggrement-upd. + cross-upd. steps. 在 disaggr.-upd. 中两个网络分别预测, 然后只保留预测不同的数据 (对称差?). cross-upd. 和 Co-teaching 一样. 两个网络对于 big-loss 有相同的 drop-rate.

(Chen et al., 2019) 使用 noisy dataset 上的随机分割作为 cross-validation. 意味着认为大多数样本是 correct labeled 的. 他们提出了 Iterative Noisy Cross-Validation(INCV), 来选择一个噪声很小的数据子集. 然后他们在这个子集上使用 Co-teaching 来训练, 并且逐步去除 big-loss 样本.

(Yao et al., 2020) 使用 AutoML 来探索记忆效应. 注意 Co-teaching(+) 两个网络的 drop-rate 相同且都是人工选择的. 他们使用 Domain-specific search, 并且用一种牛顿法来估计最优的 drop-rate, 数学上可建模为 bi-level opt.

$$\begin{aligned} R^* &= \arg \min_{R(\cdot) \in \mathcal{F}} \mathcal{L}_{\text{val}}(f(w^*; R), \mathcal{D}_{\text{val}}) \\ \text{s.t. } w^* &= \arg \min_w \mathcal{L}_{\text{tr}}(f(w; R), \mathcal{D}_{\text{tr}}) \end{aligned} \quad (451)$$

(Li et al., 2020) 启发与 MixMatch, 提出了 DivideMix, 使用 SSL 技术. 它使用了 GMM 来动态地分割训练数据为 labeled clean data + unlabeled noisy data. 在 SSL 部分, 他们使用了 Co-training 的一个变体 Co-refinement 在有标签数据上, 并且在无标签数据上使用 Co-guessing. 具体上讲, Co-refinement 把 ground-truth 和网络预测结合, 得到 refined label; Co-guessing 把两个网络的预测输出平均. 使用 MixMatch 中的方法来合并数据.

#### 49.20.4 Beyond Memorization

(Hendrycks et al., 2019) 展示了, pre-training 技术可以提高模型健壮性和不确定性, 包括对抗健壮性和标签噪声.

一般来说, pre-training 设计在更大的数据集上进行训练, 再使用 fine-tuning 来再小数据集上进行训练. Pre-traininig 展示了多种健壮性, 包括标签噪声, 对抗样本, 分布外检测和校准.

(Bahri et al., 2020) 提出了 Deep k-NN 方法, 作为 base DNN 的中间层来滤除噪声. 具体上由两步组成, 首先具有结构  $\mathcal{A}$  的模型  $\mathcal{M}$  通过 kNN 被训练来得到噪声数据  $\mathcal{D}_{noise}$ , 通过去除和其邻居数据 label 不同的样本. 接着在  $\mathcal{D}_{filtered} \cup \mathcal{D}_{clean}$  上重新训练结构  $\mathcal{A}$ .

### 49.21 Future

#### 49.21.1 Build Up New Datasets

从合成数据集 (MNIST, CIFAR) 到真实噪声数据集 (Clothing1M). 到更新更大的数据集 (“web-label noise”, Jiang et al., 2020). 需要 NLP 的噪声数据集.

#### 49.21.2 Instance-Dependent LNRL

显然 CCN(instance independent noise) 是一个经典的假设, 但是实际场景中这只是一个近似. 我们应该考虑 IDN 模型. 直觉上这是显然的, 因为模糊或者低质量的样本显然有更高概率被 mislabel. 但是和输入 feature 相关的 noise 显然大大提高了体系维度, 增加了复杂度. 并且如果没有额外的假设/信息, IDN 是不可区分的 (Cheng et al., 2020).

(Menon et al., 2018) 提出了一个边界相容的 IDN, 考虑在分类边界附近的样本具有更大的噪声. 但是这个方法仅限于二元分类, 并且无法估计噪声函数.

(Cheng et al., 2020) 研究了一种特殊的 IDN, 其中噪声函数是有上界的. 同样他们的方法也限于二分类, 并且只在小数据集上进行了测试.

(Berthon et al., 2020) 考虑了使用每个 instance 的置信度: confidence-scored IDN. 基于此, 他们提出了 instance-level 的前向矫正.

#### 49.21.3 Adversarial LNRL

(Wang et al., 2020) 提出了新的 defense algorithm MART, 将错误/正确分类的样本分别微分. 具体上讲, 使用了 Misclassification-Aware-Regularization

$$1/n \sum_{i=1}^n \mathbf{1}(h_\theta(x_i) \neq h_\theta(\hat{x}'_i)) \cdot \mathbf{1}(h_\theta(x_i) \neq y_i) \quad (452)$$

其中  $\hat{x}'_i$  是某种对抗样本<sup>26</sup>

<sup>26</sup>Namely,  $\hat{x}'_i = \arg \max_{x'_i \in \mathcal{B}_\epsilon(x_i)} \mathbb{1}(h_\theta(x'_i) \neq y_i)$

具体上, 对于错分类样本

$$\mathcal{R}^-(h_\theta, x_i) := \mathbb{1}(h_\theta(\hat{x}'_i) \neq y_i) + \mathbb{1}(h_\theta(x_i) \neq h_\theta(\hat{x}'_i)) \quad (453)$$

同时 (Zhang et al., 2020) 也考虑了同样的问题: 对抗训练中的错分类问题. 提出了 friendly adversarial training(FAT), 在训练 DNN 时使用错分类对抗样本最小化 loss, 使用正确分类对抗样本最大化 loss. 使用了 early-stopped PGD, 一旦对抗样本被错分类就停止 PGD.

#### 49.21.4 Beyond Labels: Noisy Data

除了 noisy labels, 下面介绍 noisy data 的领域.

- Feature:** 对抗样本是一种特殊的 feature noise. 可以表示为  $p(\bar{X} | Y)$ . 对抗训练显然是一种主要办法. 注意随机扰动是另一种噪声. (Zhang et al., 2019) 提出了一种 robust ResNet, 启发于动力系统. 使用显示 Euler 法的 step factor 来研究问题, 他们证明了, 小 step factor 有利于泛化健壮性.
- Preference:** ranking 中的 preference 噪声问题, 常见于各种推荐系统/在线排序/评分系统. (Han et al., 2018) 提出了 POPAL 模型, 使用集成了一个 denoising vector 的 Plackett-Luce 模型. 基于 Kendall-tau 距离, 这个向量用确定的概率校正 k-ary noisy preferences. 然而 POPAL 不能处理 k-ary noisy preferences 的动态长度, 于是 (Pan et al., 2018) 提出了 COUPLE, 使用 stagewise training. 使用了 online Bayesian inference 来优化两个模型.

对于正确分类样本

$$\mathcal{R}^+(h_{\theta}, \mathbf{x}_i) := \max_{\mathbf{x}'_i \in \mathcal{B}_{\epsilon}(\mathbf{x}_i)} \mathbb{1}(h_{\theta}(\mathbf{x}'_i) \neq y_i) = \mathbb{1}(h_{\theta}(\hat{\mathbf{x}}'_i) \neq y_i) \quad (454)$$

总的对抗经验风险

$$\begin{aligned} \min_{\theta} \mathcal{R}_{\text{misc}}(h_{\theta}) &:= \frac{1}{n} \left( \sum_{i \in \mathcal{S}_{h_{\theta}}^+} \mathcal{R}^+(h_{\theta}, \mathbf{x}_i) + \sum_{i \in \mathcal{S}_{h_{\theta}}^-} \mathcal{R}^-(h_{\theta}, \mathbf{x}_i) \right) \\ &= \frac{1}{n} \sum_{i=1}^n \{ \mathbb{1}(h_{\theta}(\hat{\mathbf{x}}'_i) \neq y_i) + \mathbb{1}(h_{\theta}(\mathbf{x}_i) \neq h_{\theta}(\hat{\mathbf{x}}'_i)) \cdot \mathbb{1}(h_{\theta}(\mathbf{x}_i) \neq y_i) \} \end{aligned} \quad (455)$$

0-1 loss 并不能直接很好的优化, 使用 surrogates, 并且使用 PGD 来进行内部极值寻找, 总目标函数为

$$\mathcal{L}^{\text{MART}}(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(\mathbf{x}_i, y_i, \theta), \quad (456)$$

其中

$$\ell(\mathbf{x}_i, y_i, \theta) := \text{BCE}(\mathbf{p}(\hat{\mathbf{x}}'_i, \theta), y_i) + \lambda \cdot \text{KL}(\mathbf{p}(\mathbf{x}_i, \theta) \| \mathbf{p}(\hat{\mathbf{x}}'_i, \theta)) \cdot (1 - \mathbf{p}_{y_i}(\mathbf{x}_i, \theta)). \quad (457)$$

其中 BCE 是 boosted CE

$$\text{BCE}(\mathbf{p}(\hat{\mathbf{x}}'_i, \theta), y_i) = -\log(\mathbf{p}_{y_i}(\hat{\mathbf{x}}'_i, \theta)) - \log\left(1 - \max_{k \neq y_i} \mathbf{p}_k(\hat{\mathbf{x}}'_i, \theta)\right) \quad (458)$$

对于 SSL 场景可以如下定义

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{sup}}(\theta) + \gamma \cdot \mathcal{L}_{\text{unsup}}(\theta) \quad (459)$$

使用 UAT++ 的无监督对抗 loss

$$\ell_{\text{sup}}^{\text{UAT}++}(\mathbf{x}, y; \theta) = \ell_{\text{unsup}}^{\text{UAT}++}(\mathbf{x}, y; \theta) = \max_{\mathbf{x}' \in \mathcal{B}_{\epsilon}} \text{CE}(\mathbf{p}(\mathbf{x}', \theta), y) + \lambda \cdot \max_{\mathbf{x}' \in \mathcal{B}_{\epsilon}} \text{KL}(\mathbf{p}(\mathbf{x}, \theta) \| \mathbf{p}(\mathbf{x}', \theta)) \quad (460)$$

同时一个工作也提出了 RST loss

$$\ell_{\text{sup}}^{\text{RST}}(\mathbf{x}, y; \theta) = \ell_{\text{unsup}}^{\text{RST}}(\mathbf{x}, y; \theta) = \text{CE}(\mathbf{p}(\mathbf{x}, \theta), y) + \lambda \cdot \max_{\mathbf{x}' \in \mathcal{B}_{\epsilon}} \text{KL}(\mathbf{p}(\mathbf{x}, \theta) \| \mathbf{p}(\mathbf{x}', \theta)) \quad (461)$$

MART 的 SSL loss

$$\mathcal{L}_{\text{semi}}^{\text{MART}}(\theta) = \sum_{i \in \mathcal{S}_{\text{sup}}} \ell_{\text{sup}}^{\text{MART}}(\mathbf{x}_i, y_i; \theta) + \gamma \cdot \sum_{i \in \mathcal{S}_{\text{unsup}}} \ell_{\text{unsup}}^{\text{MART}}(\mathbf{x}_i, y_i; \theta) \quad (462)$$

3. **Domain**: Domain Adaptation(DA) 是一个 ML 的基本问题. 传统 DA 认为 source domain 的有标签数据是 clean 的, 但不一定为真. 当 source domain 的 label 是 noisy 的, 称为 wild domain adaptation(WDA). (Liu et al., 2019) 提出了 Butterfly 框架, 同时维护四个 DNNs, 可以迭代地获得 DIR(domain-specific repr.), 和 TSR(target-specific repr.). (Yu et al., 2020) 提出了 Denoising Conditional Invariant Component(DCIC) 框架, 可证明保证能提取不变表示, 并且无偏地估计 target domain 的数据分布.
4. **Similarity**: 基于相似度的学习是 weak-supervised learning 的问题, 可用的是无标签数据和相似数据 (挺像 Graph ULL).(Wu et al., 2020) 使用一个噪声转移模型来建模相似性噪声.
5. **Graph**: 图结构是否对噪声 (node/edge feature/label, noisy links) 健壮? (Wang et al., 2020) 提出了健壮和无监督的框架 Cross-Graph, 可以处理图中的结构损坏. (Hu et al., 2019) 是 pre-training GNNs, 可能能够提高健壮性.
6. **Demonstration**: 模仿学习 (Imitation Learning, IL) 的目标时从高质量演示中学到好的策略. 当演示的质量不齐时, 显然导致了 diverse-quality demos. 可以使用置信度/ranking scores/小部分高质量 demos 来估计 noise, 来轻松的学习. 但是没有已有 expert data 时不一定能 work. (Tangkaratt et al., 2020) 提出了一个模型, 使用概率图模型来 VILD 建模 demo 质量. 具体上, 他们通过奖励函数 (代表了某种专家决策) 来估计质量, 使用了变分法来处理大的 s-a 空间, 使用 importance sampling.

## 50 Multimodal Research in Vision and Language: A Review of Current and Emerging Trends

VisLang 任务包括分类, 归纳, (retrieval), 导航. 有很多困难的人物: Vision-Language Navigation 作为对环境的详尽理解. Visual Captioning 是生成视觉信息的语言描述.

### 50.1 Tasks

#### 50.1.1 归纳任务

**Visual Question Answering, VQA** 视觉输入 + 问题 + 回答 triplet  $\langle \mathcal{I}_i, \mathcal{Q}_i, \mathcal{A}_i \rangle$  目标是学习视觉和问题输入到回答的映射  $\hat{\mathcal{A}}_i = f(\mathcal{V}_i, \mathcal{Q}_i)$  模型的输出可以属于一些答案的集合 MCQ, 也可以是没有限制的 free form. 我们认为 free form 是归纳任务, 而 MCQ VQA 是分类任务.

**Visual Captioning (VC)** VC 是给视觉输入一个合理的描述. 一般从输入学习一个 semantic repr., 再获得一个好的输出序列.

**Visual Commonsense Reasoning (VCR)** 对图像推理一个 commonsense understanding. 需要机器回答 VQA, 并且给出一个相关的理由. 相比 VQA 多了  $\{\langle \mathcal{A}_i, \mathcal{R}_i \rangle\}$  输出常常建模为, 给出多选题来回答. 所以 VCR 可以分为 question answering 和 answer justification.

NLVR(Natural Language for Visual Reasoning), 在于决定关于输入的一个 statement 是 true/false. 常常有更长的文本长度.

	Task	Interpretability Justification	Textual Output	Generative	One-to- Many	Modality Transition
Generation	VQA		✓	✓	✓	
	VC		✓	✓	✓	✓
	VCR	✓	✓	✓		
Classification	VG			✓	✓	✓
	MAC		✓			
Retrieval	NLVR		✓			
	VR				✓	✓
Others	MMT		✓	✓		
	VLN					✓

**Figure 1:** A summary of VisLang tasks based on various underlying key characteristics.

**Visual Generation (VG).** 通过文字生成视觉输出.

#### 50.1.2 分类任务

**Multimodal Affective Computing (MAC)** 情感计算涉及对情感现象的归因.

#### 50.1.3 Retrieval 任务

**Visual Retrieval** 根据文字描述, 在一个视觉 pools 中找到一个最 match 的视觉数据. 常见于各种搜索引擎.

#### 50.1.4 其他任务

**Vision-Language Navigation (VLN)** 根据语言指令, 一个 agent 根据视觉输入行动. 常看做 seq2seq 的 transcoding, 类似 VQA. 但是 VLN 常常有相对长的多的序列.

**Multimodal Machine Translation (MMT)** 包含了两部分: 翻译和描述生成. 涉及翻译一个描述到另一个语言, 给出多模态信息 (视频/音频).

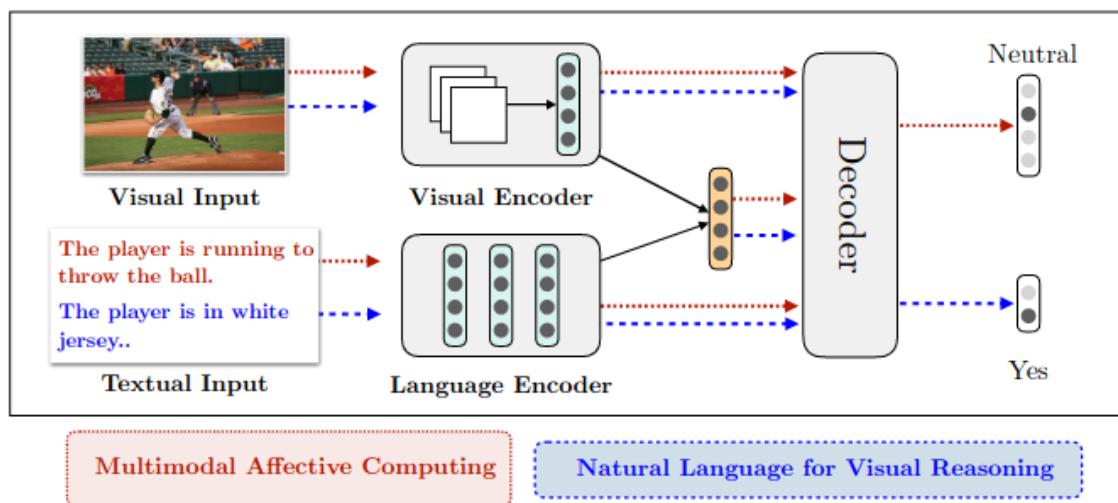
### 50.2 任务相关 Trends in VisLang

#### 50.2.1 VC

**Image Captioning, IC**

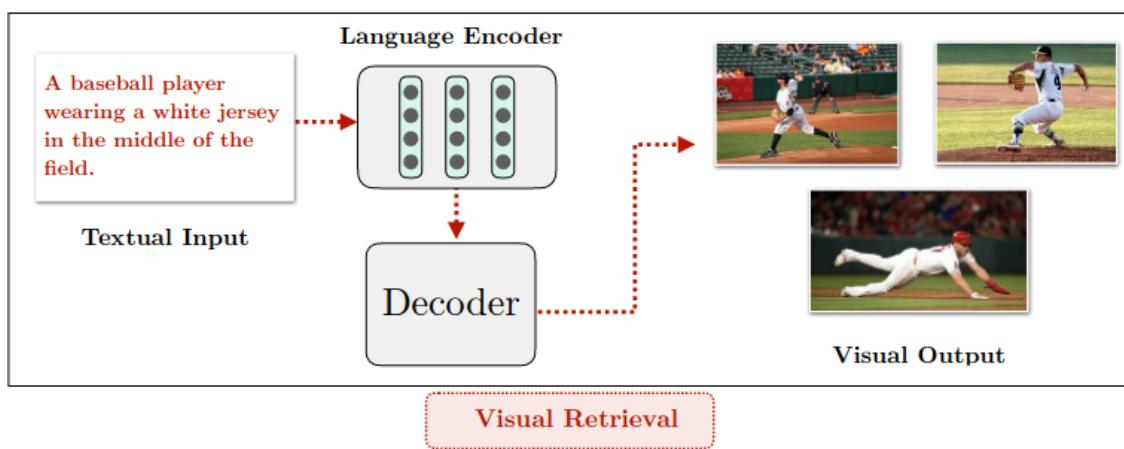


**Figure 3: Overview of Classification Tasks.**



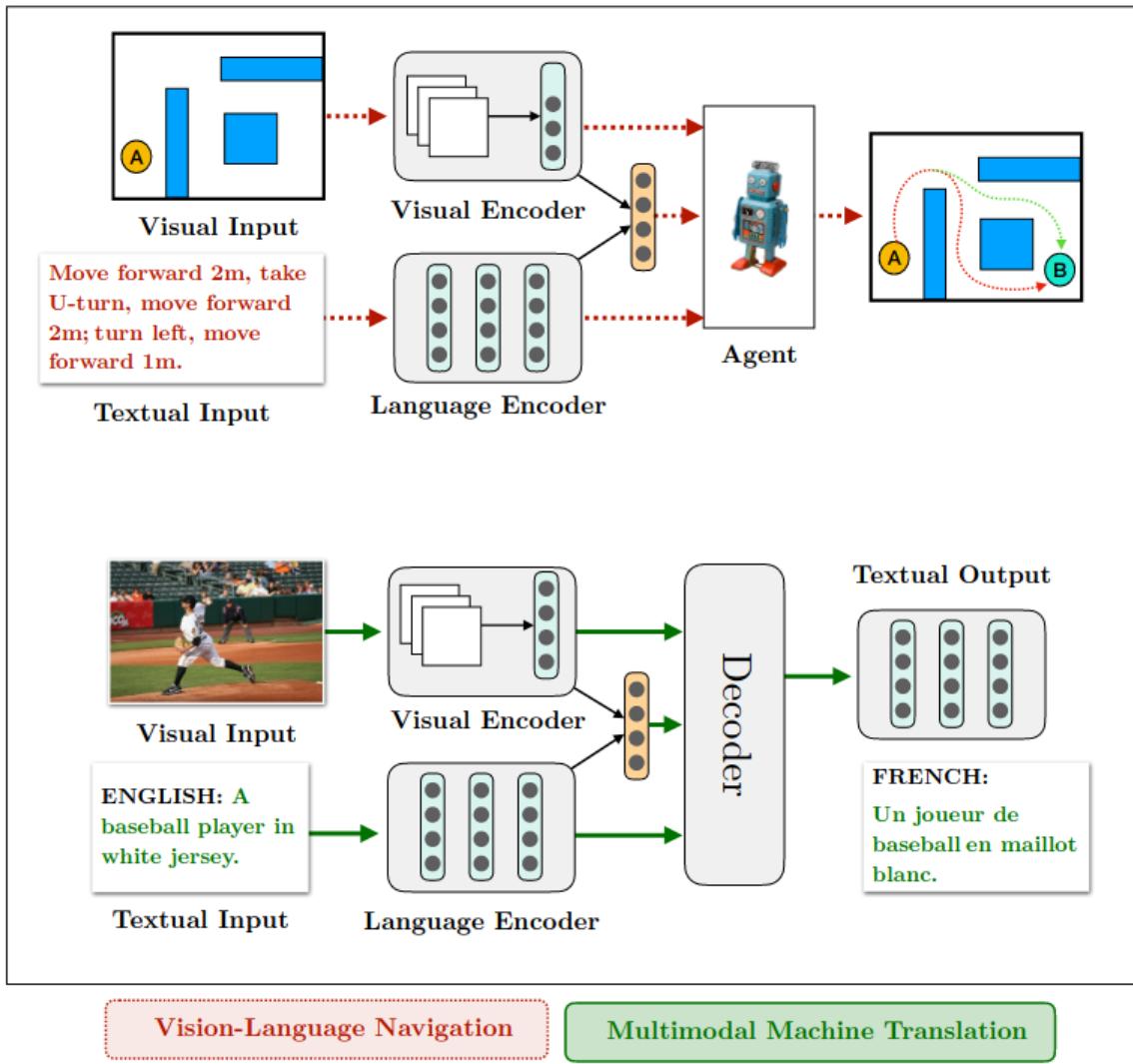
Multimodal Affective Computing      Natural Language for Visual Reasoning

**Figure 4: Overview of Retrieval Tasks.**



Visual Retrieval

**Figure 5: Overview of Other Tasks.**



## 51 Answering Questions about Data Visualizations using Efficient Bimodal Fusion

PReFIL(Parallel Recurrent Fusion of Image and Language) 算法. CQA 是一个 VQA, 涉及回答关于图表的问题. Chart 中很小的变更会导致回答的巨大改变, 所以适合作为研究推理机制的平台. CQA 常常需要 OCR 和关于给定视觉对象独特地分析词语.

Contributions

1. DVQA 的标注.
2. PReFIL 达到了超越人的性能. 开源.
3. 使用迭代 QA 来从图重建表.

Datasets

1. DVQA, 仅有条状图. Test-Familiar/Test-Novel: 后者是 training 中没有见过的例子. Open-ended. 需要考虑 OOV 问题, FigureQA 不需要,
2. FigureQA, 条状图 (横竖), 饼图, 点线图, 线图. 提供了 Val1/Test1 和 Val2/Test2, 后者颜色不存在于训练集. 所有问题都是二元 yes/no.

它们都是合成数据集, 所以缺少真实世界的数据演示.

已有方法

1. SANDY(SAN(stacked attention network) with DYnamic encoding). SAN 是 VQA 中广泛使用的框架. SAN 使用问题 (文本/表示) 来在卷积特征上 cast attention. 无法处理 OOV 问题. 为此, SANDY 使用了现有的 OCR 模块来识别任何 OOV 单词, 并且使用 dyn. encoding 来代表 OOV/chart-specific 单词. 这个 dyn. enc. 模块可以用到所有基于分类的 VQA 方法.
2. RN(relation network) 是 FigureQA 的创建者使用的. 编码任何两个图像中的 object 的二元作用. 每个 object 是卷积特征的一个 cell! RN 在 compositional reasoing (CLEVR) 上很有效.
3. FigureNet 是一个多步 FigureQA 算法. Spectral Segreagator 识别 chart 中的元素和颜色. Extract. Mod. 量化每个元素的值. 最后使用 NN 来预测答案. 它使用 FigureQA 里的具体标注来 pretrain 每个 module. 所以无法应用在饼图/线图.

### 51.1 Architecture

PReFIL 具有两个平行的 Q+I 融合分支. 每个分支从 LSTM 和 DenseNet 的不同层级拿特征. 每个融合分支 concat 问题特征到每个卷积 feature map 的每个元素上. 然后通过  $1 \times 1$  卷积 (i.e., MLPs) 来得到 question-specific 的双模态嵌入. recurrently aggregated(BiGRU), 送进分类器.

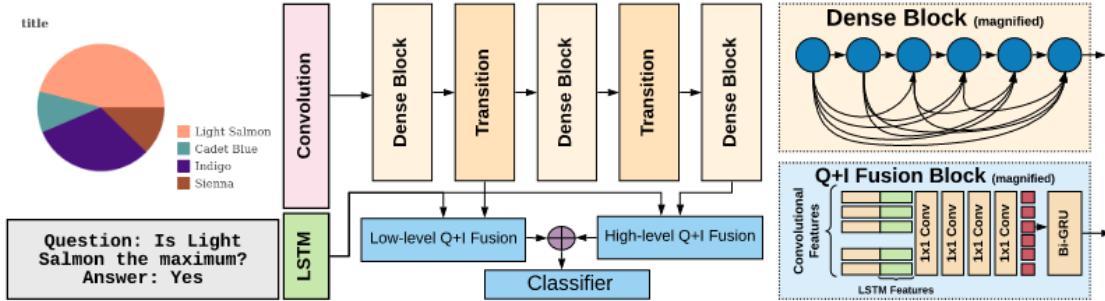


Figure 3. Components of our PReFiL model. Magnified views show the details of each dense block and Q+I fusion block.

### 51.1.1 Image Encoder

使用 DenseNet. Dense-block,..., Transition-block 更改特征尺寸 (conv/pooling). 同时使用 high-level/low-level 表示.

### 51.1.2 Parallel Fusion of I+L

首先 concat 语言 feature 到卷积 feature 的每个元素. 然后  $1 \times 1$  卷积. 这个模型很像早期 VQA 模型中 concat CNN 表示和 question 表示.

### 51.1.3 Recurrent Aggr. of bi-modal feat.

传统 CNN 的方法包括 max/mean-pooling, 另一个方法时直接 flatten(i.e., concat 每一维). 新的方法包括注意力加权和. 这些方法可能难以捕捉 features 之间的互动, 特别是对于 QA 这样的高阶任务. 我们使用 BiGRU 来集成, 具体上讲,  $F \in \mathbb{R}^{M \times N \times D}$ , 使用  $MN$  个  $D$  维向量作为输入. 同样也测试了 sum-pooling.

### 51.1.4 OCR Integration for DVQA

DVQA 需要 OCR, 有很多哦 OOV 词汇. 使用 SANDY 中同样的动态编码 scheme. 具体上, dyn. enc. 创建了一个 image-specific 的词典, 保存了 elements 的 spatial pos.. 在进行训练前, 预先进行 OCR 来得到这个词典. 接着如果问题中一个词在词典中出现, 则相应元素设为 1. 对于答案, 一部分分类层被用语 dyn. enc. 输出.

## 52 Seminar on CVPR2021

Deep Animation Interpolation: Segmentation-guided 光流 matching + RNN refinement + Synthesis

StableNet: 解决伪关系 ( $T \rightarrow X$ ,  $T \rightarrow Y$ , 导致看似  $X \rightarrow Y$  | 或者由于数据 bias). 试图给 sample 加上重加权, Correlation  $\Rightarrow$  Casuality. 在 DNN 中, 使用 RFF(随机傅立叶特征). 特征域到 RFF 域时

候的好性质：在后者线性独立，则在前者非线性独立。

AdCo：利用对抗样本作为负样本。

## 52.1 Subspace learning, Self-Expressive Model

设计两步：使用 Self-Expressive Model 得到亲和矩阵，然后使用谱聚类。

自表达模型：把数据点表示为其他数据的线性组合。若只有和本数据点相同子空间的对应系数为非 0，则说这是 subspace preserving 的。试图找到某种好的 reg. 来获得这种表示系数。 $l_1$  范数 SSC.

## 52.2 Kaleido-BERT, Fashion Domain VisLang Pre-training.

经典方法：单流，VL-BERT/Fashion-BERT，图像 token/patch 特征送到 BERT. 双流 BERT/Transformer：ViLBERT. 三流 V,L,V+L. MMFT-BERT.

通用 RoI 检测器不太使用与 Fashion 领域。

传统 Masking-预测的自监督（填空）。

图像侧的信息利用不充分（自监督中）。

Kaleido Patch Generation/KPG. saliency Net 提取前后景。多尺度 patch 送到 ResNet. Ablation Study 显示，这是最有效的提升。

Attention-based Align. Generation/AAG. 预训练图像 patch-词语 token 的对应。根据这个对应来 mask(Alignment Guided Matching/AGM)

Aligned Kaleido Path Modeling：对 patch 重拍/隐藏 etc. 来自监督

## 53 OSCAR: Object-Semantics Aligned Pre-training for Vision-Language Tasks

研究指出，VLP(Vision Language Pretraining) 可以给出 SoTA. 之前的 VLP 基于多层 Transformer, concat image region feature 到文本特征，使用注意力来学习文本到图片区域的对齐。缺少显式监督的随其信息导致这成为了一个弱监督学习任务。

启发于，现代物体检测框架可以非常准确，并且这些对象很有可能在文本中出现。在 6.5M 对 V+L 上 pretrain. 本工作是第一个使用 anchor point 在 VLP 的。

典型地，VLP 使用多层自注意力 Transformer 来学习跨模态表示。依赖于自注意力机制来学习图片-文本对应。缺点：

- Ambiguity. 常常使用 Faster RCNN 提取出的特征。overlapping.
- 缺少因果/信息。

### 53.1 OSCAR

输入为词-标签-图像 triplet( $w, q, v$ )。使用 pretrained-BERT 得到  $q, w$  之间的对齐，作为 OSCAR 的初始化；被检测出 tag 的图像区域也有高的初始注意力权重。

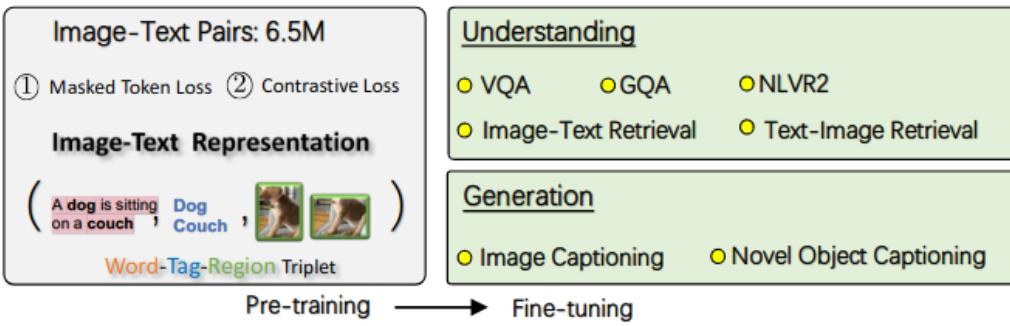


Fig. 1: OSCAR pipeline. The model takes a triple as input, is pre-trained with two losses (a masked token loss over words & tags, and a contrastive loss between tags and others), and fine-tuned for 5 understanding and 2 generation tasks (detailed in Sec. 4).

### OSCAR: Object-Semantics Aligned Pre-training for Vision-Language Tasks 3

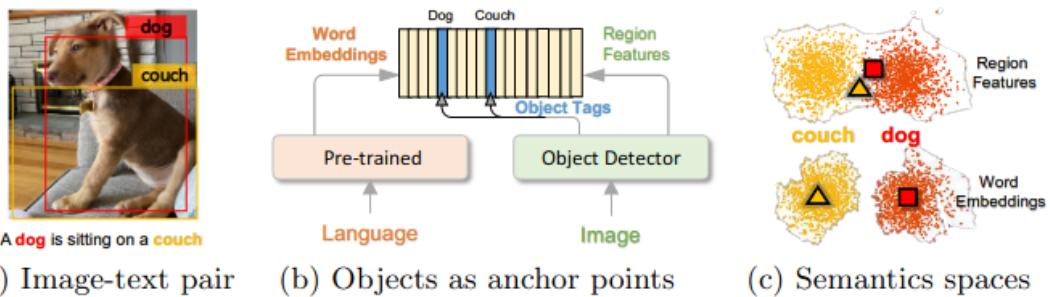


Fig. 2: Illustration on the process that OSCAR represents an image-text pair into semantic space via dictionary look up. (a) An example of input image-text pair (b) The object tags are used as anchor points to align image regions with word embeddings of pre-trained language models. (c) The word semantic space is more representative than image feature space. In this example, dog and couch are similar in the visual feature space due to the overlap regions, but distinctive in the word embedding space.

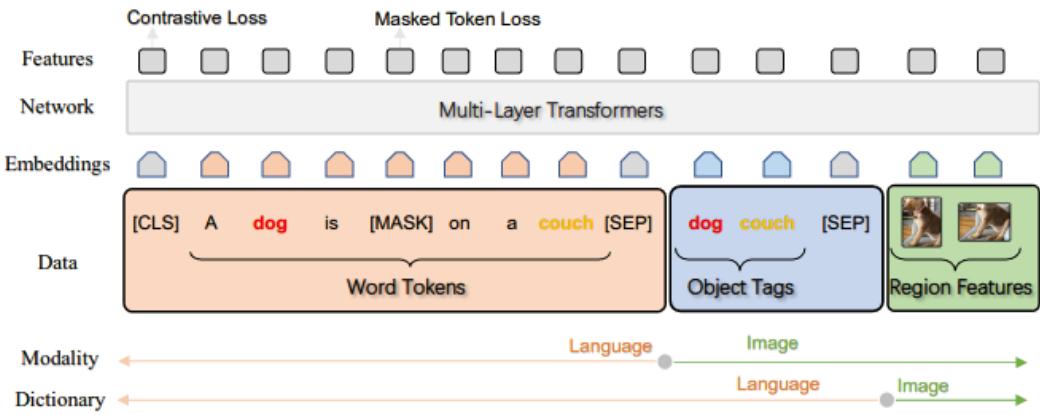


Fig. 3: Illustration of OSCAR. We represent the image-text pair as a triple [ word tokens , object tags , region features ], where the object tags (e.g., “dog” or “couch”) are proposed to align the cross-domain semantics; when removed, OSCAR reduces to previous VLP methods. The input triple can be understood from two perspectives: a *modality* view and a *dictionary* view.

具体上,  $\mathbf{v}, \mathbf{q}$  如下生成, 给定有  $K$  个对象区域的图片, Faster RCNN 用于提取图像特征  $(\mathbf{v}', \mathbf{z})$ ,  $\mathbf{v} \in \mathbb{R}^P$ , 而  $\mathbf{z}$  是一个 4/6 维向量 (左上/右下坐标 and/or 长宽), concat 二者. 使用线性映射来保证和词嵌入有相同的维度. 同时 Faster RCNN 用于检测一组高准确度的对象标签,  $\mathbf{q}$  是这些标签的词嵌入.

$$\mathbf{x} \triangleq [\underbrace{\mathbf{w}}_{\text{language}}, \underbrace{\mathbf{q}, \mathbf{v}}_{\text{image}}] = [\underbrace{\mathbf{w}, \mathbf{q}}_{\text{language}}, \underbrace{\mathbf{v}}_{\text{image}}] \triangleq \mathbf{x}' \quad (463)$$

用两种方式 interpret triplet

从字典角度看  $\mapsto$  Masked Token Loss(MTL). 定义  $\mathbf{h} \triangleq [\mathbf{w}, \mathbf{q}]$ . 每次迭代, mask 掉 15% 的 token, 用特殊的 MASK token 替换. MTL 的目标时预测 masked token. 于是 MTL 为

$$\mathcal{L}_{\text{MTL}} = -\mathbb{E}_{(\mathbf{v}, \mathbf{h}) \sim \mathcal{D}} \log p(h_i | \mathbf{h}_{\setminus i}, \mathbf{v}) \quad (464)$$

这和 BERT 类似.

从模态角度看: Contrastive Loss. 考虑图像模态  $\mathbf{h}' \triangleq [\mathbf{q}, \mathbf{v}]$  和语言模态  $\mathbf{w}$ . 在  $\mathbf{q}$  中 50% 堆积替换 tag. 在 encoder 输出后面加上一个 FC 层来预测 tag 是否包含了任何被替换的 tag ( $y=0$ ), 或者没有变化 ( $y=0$ ).

在跨模态预训练中, 使用对象标签来调整 BERT 的词嵌入. 总体 objective 是

$$\mathcal{L}_{\text{Pre-training}} = \mathcal{L}_{\text{MTL}} + \mathcal{L}_{\text{C}} \quad (465)$$

Reason: 虽然简单, 但是性能好. 简单地表现了两种 perspective.

Pretraining dataset: COCO, Conceptual Captions, SBU captions, flicker30k, GQA etc. 4.1M images, 6.5M triplets. OSCAR<sub>B</sub> 使用 BERT base, AdamW, 1.0M steps, 5e-5 lr, BS 768.

## 53.2 Adapting to V+L Tasks

**Image-Text Retrieval** 看作二分类. 对于任何 aligned text-image pair, 使用一个不同的 image/caption 来形成 unaligned pair. 没有使用 ranking loss, 二分类 loss 效果更好. 测试阶段, 输出的 prob. score 用来得出 ranking.

**Image Captioning** 使用 seq2seq objective. 随机 mask 15% 的 caption tokens 然后预测对应 token id. 类似 VLP, 使用了仅仅使用之前的 token feature 的自注意力来模拟单方向生成. 推理阶段, 把 image regions, obj. tags 和特殊 token[CLS] 作为输入, 并且模型的生成通过追加 [MASK] token, 然后输出的 token 替代 [MASK]token, 再追加 etc.. 模型输出 [STOP] 时则停止. 使用 beam search.

**Novel Obj. Captioning(NoCaps)** 从 Open Images dataset 中找到图像, 测试模型的 novel objects 的描述能力. 使用预测的 VG/Open Images 标签来生成 tag seq., 并且在 COCO 上训练 OSCAR.

**VQA** VQA 是多选回答. 使用 VQA 2.0, 给予 MSCOCO. 对于每个问题, 模型挑选 3129 个回答中的一个. 使用 [CLS] 的输出来作为特征, 并且送到线性分类器里. 视为 multi-label 分类问题, 给每个输出一个 soft target score, 取决于和 human answer 的相近度. 最后使用 CE, 使用预测分数和 soft-target 分数.

**GQA** 需要提供额外理由. 对于每个问题, 模型从 1852 个 candidate 里选择.

**Natural Language Visual Reasoning for Real, NLVR2** 选择一对图片/自然语言. 目标是决定这个自然语言语句是否为真.

## 54 In Defense of Grid Features for Visual Question Answering

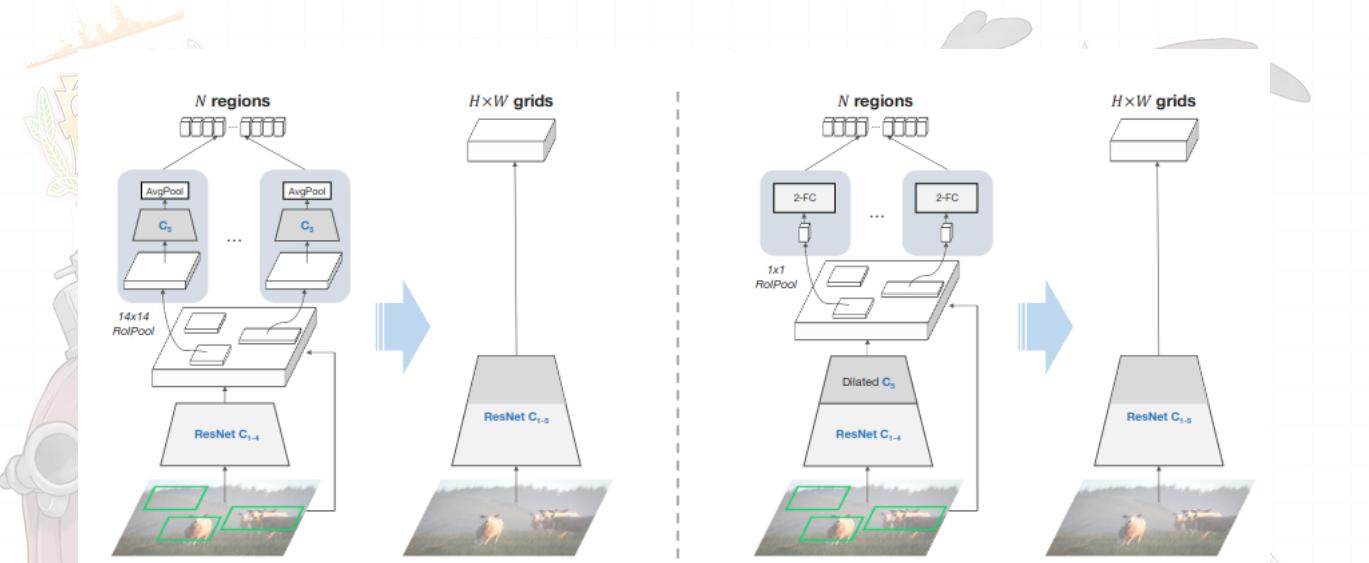
最近基于 region 的方法逐渐流行并超过了基于 grid 的方法. 为啥 region 方法好? 更好的对象定域性. 容易同时得到 high+low level 的信息?. 我们发现使用 obj. detectors 的 grid feature 而不是 region feature 性能不变, 小修改还能提高性能. 相容实验发现主要影响性能的是 pre-training 的数据集质量 + 输入图像的高分辨率, grid/region 只是小问题.

### 54.1 Related Work

现代 VQA 框架的性能依赖于视觉特征的提取, VGG $\leftrightarrow$  ResNet  $\mapsto$  bottom-up attention(Transformer, BERT?).

很多 VQA 方法都包含了 pre-training, 视觉模型在 ImageNet+VG, 语言模型在词嵌入. 一些方法提出了共同的训练 V+L, 特别是利用 BERT(Linear Transformer variants etc.) 来把 regions/words 看作 token, 用于 masked prediction.

R-CNN(基于 region refinement) vs. 一阶段对象检测器



**Figure 2: From regions to grids.** **Left:** We convert the original region feature extractor used by bottom-up attention [2] back to the ResNet [15] grid feature extractor for the *same* layer (see Sec. 3.2, weights in blue are transferred), and find it works surprisingly well for VQA [11]. **Right:** We build a detector based on  $1 \times 1$  RoIPool while keeping the output architecture *fixed* for grid features (see Sec. 3.3), and the resulting grid features consistently perform at-par with region features.

## 54.2 From Regions to Grids

### 54.2.1 Bottom-Up Attention w/ Regions

一般使用 Faster R-CNN, 在 cleaned ver. VG 上训练. 对于这些方法, 要获得自底向上注意力特征, 进行如下两步

1. Region Selection. 通过一个 Region Proposal Net., 提出候选 region(Regions of Interest, RoIs), 接着通过一个 score comp., 选择 top-N 的区域, 并且两步都使用 NMS.
2. 给出了上述步骤的 regions, 使用 RoIPool 来得到 region-feature.

由于 VG 数据集的复杂性和 Faster R-CNN, 这两步计算上都很昂贵.

### 54.2.2 Grid Features from the Same Layer

Faster R-CNN 是 c4 模型 w/ 属性分类分支的变种. 首先使用 ResNet 的  $C_4$  blocks 来得到 feature map, 接着 per-region feature 先  $14 \times 14$  RoIPool, 再应用  $C_5$ , 最后 avg-pool 来得到每个 region 的 F. 我们直接使用  $C_5$  在 grids 来得到特征.

### 54.2.3 $1 \times 1$ RoIPool for Improved Grid F.

这意味着用一个一维向量来表示一个 region, 而不是 Faster RCNN 里的 HWC 三维. 使用  $1 \times 1$  RoIPool 会降低物体检测的性能, 对于 VQA, 这要求这个特征尽可能单独地编码信息. 由于预训练的  $C_5$  输入不适用, 使用最近的直接使用整个  $C_5$  的 ResNet 的工作<sup>27</sup>.

<sup>27</sup>Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. De-formable convnets v2: More deformable, better results. InCVPR, 2019.

### 54.3 Comparison: Region v. Grids

具体的 VQA 模型使用的是 co-attention model<sup>28</sup>  
推理速度远快于基于 regions 的方法.

### 54.4 Why Work?

发现使用更大的图像, 更高的精度.

不同的预训练任务上, detection w/ attr. > detection w/o attr. > classification w/ tag > cls. w/ label.

### 54.5 Generalization

使用 Res-NeXt 改进了性能. 在别的 framework 上使用 R/G 特征也提高了性能.

## 55 Beyond Bilinear: Generalized Multimodal Factorized High-order Pooling for Visual Question Answering

要进行 VQA, 需要解决三个问题: 提取两边特征, 结合特征, 用来预测答案.

直觉上, 使用注意力机制是很自然的. 所以他们使用 co-attention 模块来同时学习两边的 attention. 很多模型直接使用线性特征融合 (concat/add), 显然不好. 最近 bilinear pooling 似乎用来聚合不同的 CNN 特征, 但是对与 V+L 表示可能输出维度太高. 于是有 Multi-modal Compact Bilinear(MCB) pooling, Multi-modal Low-rank Bilinear(MLB) pooling. 然而 MCB 需要非常高的特征维度来保证性能, 同时 MLB 需要非常多的迭代数来收敛. 他们提出 MFB(Multi-modal Factorized Bilinear), 以及使用高阶设置来得到 Multi-modal Factorized High-order pooling (MFH).

### 55.1 Related Work: Multi-modal Bilinear Models for VQA

(Fukui et al., 2016) 提出了 MCB, 使用特征向量的外积来得到非常高维的二次融合特征. 为了减少计算代价, 使用了一个基于采样的估计, 基于事实: 向量的可以表示为卷积. (Kim et al., 2017) 使用了 MLB, 基于向量的 Hadamard 积, 在一个共同的 low-rank 空间里:

$$z = \text{MLB}(x, y) = (U^T x) \circ (V^T y) \quad (466)$$

为了提供非线性, z 后面接上 activation(like tanh). 实验说明 MLB 可能有非常低的收敛速度 (250k its/140 epochs, 200 bs)

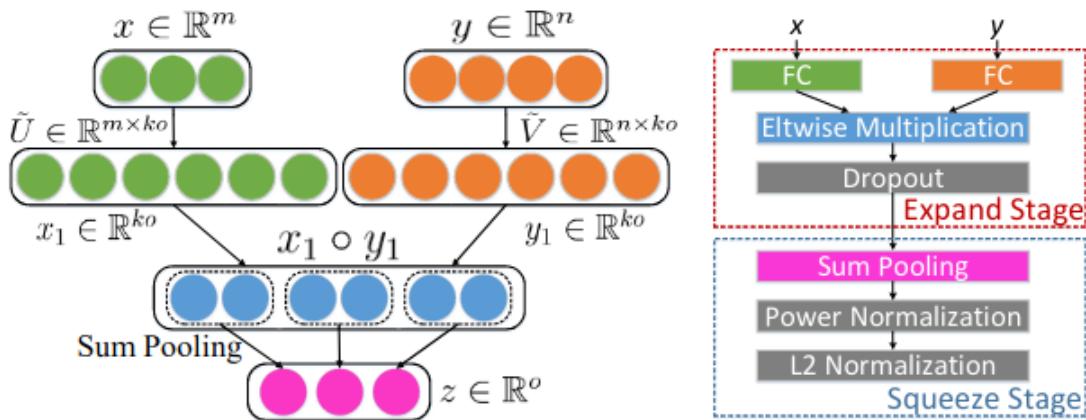
<sup>28</sup>Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. TNNLS, 2018.

## 55.2 Generalized Multi-modal Factorized High-order Pooling

对于视觉特征  $x \in \mathbb{R}^m$  和语言特征  $y \in \mathbb{R}^n$  最基本的 bilinear fusion

$$z_i = x^T W_i y \quad (467)$$

同时带来了巨大数量的参数.



(a) Multi-modal Factorized Bilinear Pooling

(b) MFB module

Fig. 2. The flowchart of Multi-modal Factorized Bilinear Pooling and complete MFB module.

启发于单模态问题中的矩阵分解技巧, 使用 low-rank approx.

$$\begin{aligned} z_i &= x^T U_i V_i^T y = \sum_{d=1}^k x^T u_d v_d^T y \\ &= 1^T (\tilde{U}_i^T x \circ \tilde{V}_i^T y) \end{aligned} \quad (468)$$

其中

$$U_i = [u_1, \dots, u_k] \in \mathbb{R}^{m \times k} \quad (469)$$

并且令  $V = [V_1, \dots, V_o] \in \mathbb{R}^{n \times k \times o}$  编程上, 写成

$$z = \text{SumPool}(\tilde{U}^T x \circ \tilde{V}^T y, k) \quad (470)$$

之后接上 dropout, power norm.  $z \leftarrow \text{sign}(z)|z|^{0.5}$  和 l2 norm.  $(z \leftarrow z/\|z\|)$ .

接着 MFH. 设 MFB 里的中间表示为

$$z_{\text{exp}} = \text{MFB}_{\text{exp}}(x, y) = \text{Dropout}(\tilde{U}^T x \circ \tilde{V}^T y) \in \mathbb{R}^{ko} \quad (471)$$

以及

$$z = \text{MFB}_{\text{sqz}}(z_{\text{exp}}) = \text{Norm}(\text{SumPool}(z_{\text{exp}})) \in \mathbb{R}^o \quad (472)$$

使用多层 cascade 结构

$$z_{exp}^i = \text{MFB}_{exp}^i(x, y) = z_{exp}^{i-1} \circ \left( \text{Dropout} \left( \tilde{U}^{i^T} x \circ \tilde{V}^{i^T} y \right) \right) \quad (473)$$

其中第 0 层  $z_{exp}^0 \in \mathbb{R}^{k^{ho}}$  concat 所有层的特征

$$z = \text{MFH}^p = [z^1, z^2, \dots, z^p] \in \mathbb{R}^{op} \quad (474)$$

### 55.3 Net Architecture

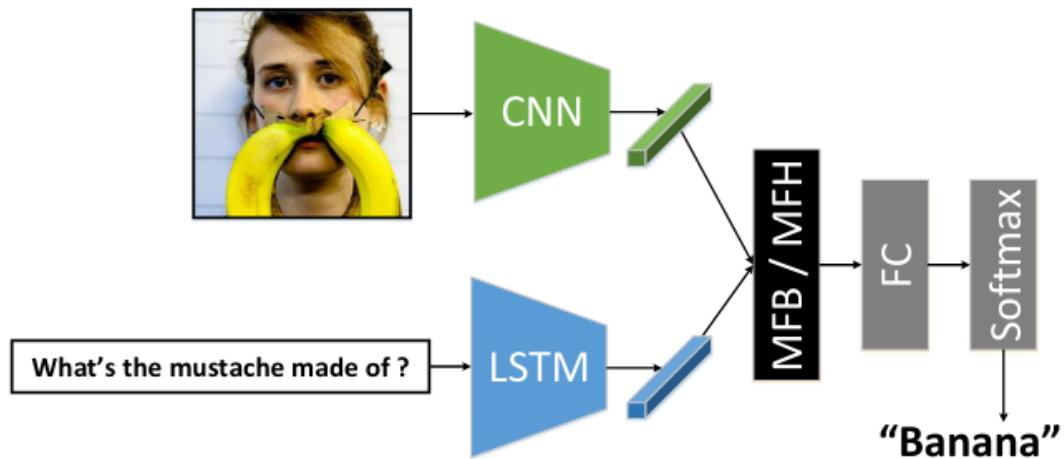


Fig. 4. The baseline network architecture with MFB or MFH and without the attention mechanism for VQA.

对于 baseline 模型, CNN 骨架是 ImageNet 上预训练的 ResNet-152,  $448 \times 448$  输入, 2048d  $C_5$  特征。词语 tokenize 后吧 one-hot embedding 送进 embedding 层后, 送入 1024d LSTM。对于分类器, 直接使用 top-N 最常出现的类别, 因为类别服从长尾分布。

MFH/MFB+CoAtt 模型! LSTM 生成特征, 使用自注意力得到 Question Att. Feat., 用于和 Image Feat. 进行 MFB/MFH, 再得到 Image Att. Coeff., 在得到带注意力的 Image Feat., Image Att. Feat. 和 Question Att. Feat. MFB/MFH 融合送到 FC 来进行分类。

具体上, 取到  $N = 3000$ (对于 VQAv1/v2). 相容实验中的发现:

1. 缺少  $L_2$  norm 降低了性能 (-3%),  $L_p$  norm 影响并不大。
2.  $L_2$  norm 明显让 neuron value 变得稳定, 让模型更稳定。

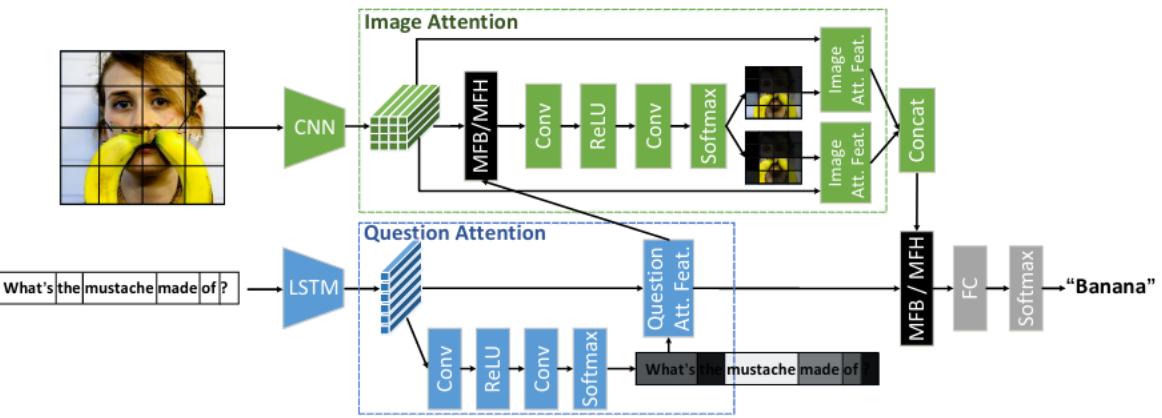


Fig. 5. The co-attention network architecture with MFB or MFH for VQA. Different from the network of MFB baseline, the images and questions are firstly represented as the fine-grained features respectively. Then, *Question Attention* and *Image Attention* modules are jointly modeled in the framework to provide more accurate answer predictions. For both the image and question attention modules, multiple attention maps (see the example in the image attention module) can be adapted to further improve the representation capacity of the fine-grained features.

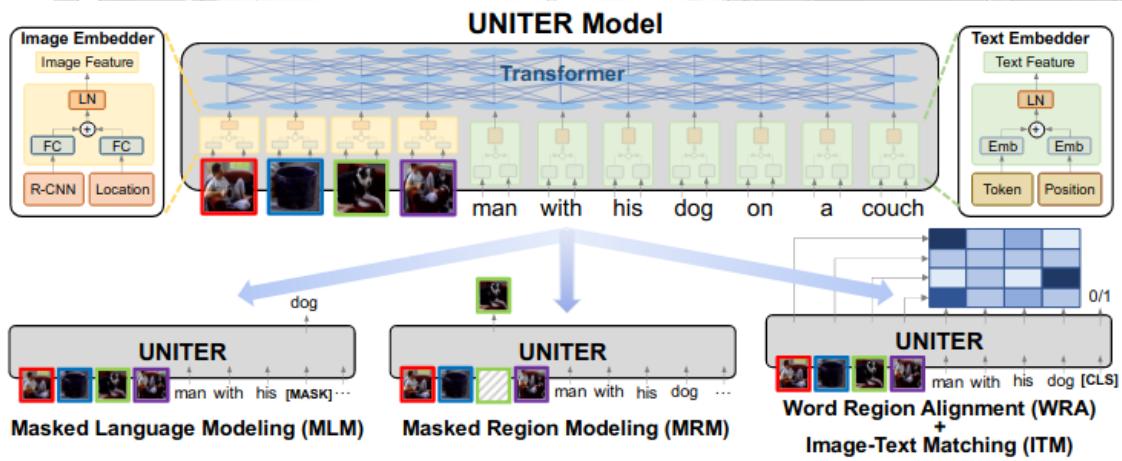


Fig. 1: Overview of the proposed UNITER model (best viewed in color), consisting of an Image Embedder, a Text Embedder and a multi-layer Transformer, learned through four pre-training tasks

## 56 UNITER UNiversal Image-Text Representation Learning

MCB, BAN, DFAF 提出了 VQA 的模态融合方法. SCAN, MAttNet 研究了图像和语言的潜在对齐 (在 Retrieval, Referring Exp. Comprehension). 他们提出一个通用 V+L 表示学习方法. 使用 Transformer, 启发于 BERT, 我们使用这些任务 pretrain UNITER:

1. Masked Language Modeling (MLM) conditioned on image
2. Masked Region Modeling (MRM) conditioned on Text
  - Masked Region Classification (MRC)
  - Masked Region Feature Regression (MRFR)
  - Masked Region Classification with KL-divergence (MRC-kl)
3. Image-Text Match-ing (ITM)
4. Word-Region Alignment (WRA).

并且使用 conditional masking 和基于最佳输送 (OT) 的 WRA 任务. 具体的, 我们尝试最小化从 image regions embed. 到 words embed. 的最佳输送.

Pre-trained 语言模型大大改进了 NLP 的能力 (ELMo, BERT, GPT-2, XLNet, RoBERTa, ALBERT), 特点是在大型数据集上进行预训练, 同时使用了 Transformer 架构. 同时多模态预训练也有很多工作, VideoBERT, CBT 使用 BERT 同时学习 video-text pairs. ViLBERT, LXMERT 则使用了双流结构, Transformers 分别用于图像和文字, 在使用第三个 Transformer 来融合. 另一边 B2T2, VisualBERT, Unicoder-VL, VL-BERT 是单流结构. (Gan et al., 2020) 提出了使用多任务和对抗训练来提高性能. VALUE 使用了 probe tasks 来理解预训练模型.

具体上, Image Embedder 使用 Faster RCNN 提取出的特征 (ROIPool 过), 以及区域位置特征送到两个 FC, 再相加, 再 LN. Text Embedder 类似 BERT, 将输入句子 token 成 WordPieces. 每个 subword 的 embedding 是词嵌入和位置嵌入加和, 再 LN.<sup>29</sup>

<sup>29</sup>LN, Layer Normalization

以 MLP 为例, feature 是  $a \in \mathbb{R}^{N \times H}$

$$\mu^l = \frac{1}{H} \sum_{i=1}^H a_i^l \quad \sigma^l = \sqrt{\frac{1}{H} \sum_{i=1}^H (a_i^l - \mu^l)^2} \quad (475)$$

归一化

$$\hat{a}^l = \frac{a^l - \mu^l}{\sqrt{(\sigma^l)^2 + \epsilon}} \quad (476)$$

增加 gain, bias

$$h^l = f(g^l \odot \hat{a}^l + b^l) \quad (477)$$

对于 NLP/GNN,  $f \in \mathbb{R}^{B \times T \times N} \mapsto \mathbb{R}^{B \times T \times 1}$ , 对于 MLP,  $f \in \mathbb{R}^{B \times C \times N} \mapsto \mathbb{R}^{B \times 1 \times 1}$ , 相较之下 BN  $f \in \mathbb{R}^{B \times C \times N} \mapsto \mathbb{R}^{1 \times C \times 1}$ .

## 56.1 Pre-training Tasks

**Masked Language Modeling (MLM)** 图片区域  $\{\mathbf{v}_1, \dots, \mathbf{v}_K\}$ , 词  $\mathbf{w} = \{\mathbf{w}_1, \dots, \mathbf{w}_T\}$ , mask 指标  $\mathbf{m} \in \mathbb{N}^M$ . MLM 中随机 mask 15% 的词, 使用 [MASK]<sup>30</sup>. 目标是用周围的词预测 masked words

$$\mathcal{L}_{\text{MLM}}(\theta) = -\mathbb{E}_{(\mathbf{w}, \mathbf{v}) \sim D} \log P_{\theta}(\mathbf{w}_m | \mathbf{w} \setminus \mathbf{m}, \mathbf{v}) \quad (478)$$

**Remark** As if OSCAR does.

**Image-Text Matching (ITM)** 加入了特殊 Token [CLS], 输入的是一些 regions 和词, 输出是否是匹配的文字/图像对. 具体的, 使用 [CLS] 对应的输出作为 joint repr., 送到 FC+sigmoid 里的道  $[0, 1]$  的分数. 表示为  $s_{\theta}(\mathbf{w}, \mathbf{v})$ . 使用二分类 CE

$$\mathcal{L}_{\text{ITM}}(\theta) = -\mathbb{E}_{(\mathbf{w}, \mathbf{v}) \sim D} [y \log s_{\theta}(\mathbf{w}, \mathbf{v}) + (1 - y) \log (1 - s_{\theta}(\mathbf{w}, \mathbf{v}))] \quad (479)$$

训练中总是 50% 几率送进 paired/unpaired 的对.

**Remark** As if OSCAR does.

**Word-Region Alignment (WRA)** 使用 transport plan  $\mathbf{T} \in \mathbb{R}^{T \times K}$  来表示  $\mathbf{w}/\mathbf{v}$  的对齐. 这是好的, 因为

- 自正则化, 元素和为 1
- 稀疏, 精确解必然只包含  $2r-1$  个非 0 元素
- 效率, 相比 linear solvers, 可以通过迭代法求解

具体的, 把  $\mathbf{w}, \mathbf{v}$  看作离散分布  $\boldsymbol{\mu}, \boldsymbol{\nu}$ , 有  $\boldsymbol{\mu} = \sum_{i=1}^T \mathbf{a}_i \delta_{\mathbf{w}_i}$  and  $\boldsymbol{\nu} = \sum_{j=1}^K \mathbf{b}_j \delta_{\mathbf{v}_j}$  并且权值为单形上的顶点  $\mathbf{a} = \{\mathbf{a}_i\}_{i=1}^T \in \Delta_T$  and  $\mathbf{b} = \{\mathbf{b}_j\}_{j=1}^K \in \Delta_K$ , 那么定义两个离散分布之间的 OT 距离

$$\mathcal{L}_{\text{WRA}}(\theta) = \mathcal{D}_{ot}(\boldsymbol{\mu}, \boldsymbol{\nu}) = \min_{\mathbf{T} \in \Pi(\mathbf{a}, \mathbf{b})} \sum_{i=1}^T \sum_{j=1}^K \mathbf{T}_{ij} \cdot c(\mathbf{w}_i, \mathbf{v}_j) \quad (480)$$

其中  $c$  是运输 cost, 实验中使用的是 cosine unsim.(cos. dist.)

$$c(\mathbf{w}_i, \mathbf{v}_j) = 1 - \frac{\mathbf{w}_i^\top \mathbf{v}_j}{\|\mathbf{w}_i\|_2 \|\mathbf{v}_j\|_2} \quad (481)$$

$\mathbf{T}$  的精确求解 intractable, 使用 IPOT 算法进行估计.

**Masked Region Modeling (MRM)** 类似 MLM, 15% 随机 drop feature. 训练来重建缺失的 visual feature, 但是由于这些特征是连续和高维的, 使用三个变种, 同一个 objective:

$$\mathcal{L}_{\text{MRM}}(\theta) = \mathbb{E}_{(\mathbf{w}, \mathbf{v}) \sim D} f_{\theta}(\mathbf{v}_m | \mathbf{v}_{\setminus m}, \mathbf{w}) \quad (482)$$

<sup>30</sup>和 BERT 一致, 10% 变成随机词, 10% 不变, 80% 变成 [MASK]



1. **Masked Region Feature Regression (MRFR)** 使用一个 FC 层把 Transformer 输出转换到输入同尺寸的空间上, 再使用 l2 回归

$$f_{\theta}(\mathbf{v}_m | \mathbf{v} \setminus m, \mathbf{w}) = \sum_{i=1}^M \| h_{\theta}(\mathbf{v}_m^{(i)}) - r(\mathbf{v}_m^{(i)}) \|_2^2 \quad (483)$$

2. **Masked Region Classification (MRC)** 预测每个 masked 对象的 object semantic class. 使用 CE

$$f_{\theta}(\mathbf{v}_m | \mathbf{v} \setminus m, \mathbf{w}) = \sum_{i=1}^M \text{CE}(c(\mathbf{v}_m^{(i)}), g_{\theta}(\mathbf{v}_m^{(i)})) \quad (484)$$

3. **Masked Region Classification with KL-Divergence (MRC-kl)** 使用 soft-label, 即 object detector 的原始输出来做 CE/KL

$$f_{\theta}(\mathbf{v}_m | \mathbf{v} \setminus m, \mathbf{w}) = \sum_{i=1}^M D_{KL}(\tilde{c}(\mathbf{v}_m^{(i)}) \| g_{\theta}(\mathbf{v}_m^{(i)})) . \quad (485)$$

## 57 Vusial Commonsense R-CNN