Deep Model Generalization for Medical Image Computing at Scale

DOU Qi

Department of Computer Science and Engineering
co-affiliated with T Stone Robotics Institute
The Chinese University of Hong Kong
Model Generalization in Real World Conditions

Data-driven method is sensitive to data mismatch

- Large-scale data always encounter data heterogeneity
- Medical imaging: different vendors, imaging protocols, patient population, etc.

An empirical study on the impact of scanner effects with brain imaging

Construct an age- and sex-matched dataset with T1-weighted brain MRI from n = 592 individuals, where 296 subjects (146 F) are taken each from the Cam-CAN and UKBB, to simulate a somewhat ‘best case scenario’ to remove population bias.

Very careful pre-processing is conducted, including: 1) reorientation, 2) skull stripping, 3) bias field correction, 4) intensity-based linear registration (rigid and affine) to MNI space, 5) whitening for intensity normalization

Site classification with random forest binary classifier

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>rigid</td>
<td>whitening</td>
<td>96.96%</td>
<td>0.4039</td>
<td>0.8296</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>affine</td>
<td>whitening</td>
<td>98.82%</td>
<td>0.3876</td>
<td>0.8397</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>rigid</td>
<td>graymatter</td>
<td>80.24%</td>
<td>0.6363</td>
</tr>
<tr>
<td>✓</td>
<td>non-linear</td>
<td>graymatter</td>
<td>96.62%</td>
<td>0.5675</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>rigid</td>
<td>graymatter</td>
<td>93.24%</td>
<td>0.4542</td>
</tr>
</tbody>
</table>


Related work: [Shafto et al., 2014; Taylor et al., 2017; Sudlow et al., 2015; Miller et al., 2016; Alfaro-Almagro et al., 2018]
Tackling Data Heterogeneity with Supervised Learning

A case study with prostate T2-weighted MRI image segmentation

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Case num</th>
<th>Field strength (T)</th>
<th>Resolution (in-plane/through-plane) (mm)</th>
<th>Coil</th>
<th>Manufacturer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site A</td>
<td>30</td>
<td>3</td>
<td>0.6-0.625/3.6-4</td>
<td>Surface</td>
<td>Siemens</td>
</tr>
<tr>
<td>Site B</td>
<td>30</td>
<td>1.5</td>
<td>0.4/3</td>
<td>Endorectal</td>
<td>Philips</td>
</tr>
<tr>
<td>Site C</td>
<td>19</td>
<td>3</td>
<td>0.67-0.79/1.25</td>
<td>No</td>
<td>Siemens</td>
</tr>
</tbody>
</table>

Methods | BFC | NF | Intensities | Site A | Site B | Site C | Overall |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Separate (A)</td>
<td>X</td>
<td>X</td>
<td>whitening</td>
<td>90.47</td>
<td>76.44</td>
<td>56.81</td>
<td>90.56</td>
</tr>
<tr>
<td>Separate (B)</td>
<td>X</td>
<td>X</td>
<td>whitening</td>
<td>70.11</td>
<td>90.52</td>
<td>50.25</td>
<td></td>
</tr>
<tr>
<td>Separate (C)</td>
<td>X</td>
<td>X</td>
<td>whitening</td>
<td>57.93</td>
<td>55.25</td>
<td>90.70</td>
<td></td>
</tr>
<tr>
<td>Joint</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>86.51</td>
<td>88.00</td>
<td>86.78</td>
<td>87.10</td>
</tr>
<tr>
<td>Joint</td>
<td>X</td>
<td>X</td>
<td>histogram</td>
<td>87.68</td>
<td>88.02</td>
<td>89.46</td>
<td>88.39</td>
</tr>
<tr>
<td>Joint</td>
<td>X</td>
<td>X</td>
<td>scaled</td>
<td>90.43</td>
<td>88.06</td>
<td>88.26</td>
<td>88.92</td>
</tr>
<tr>
<td>Joint</td>
<td>X</td>
<td>X</td>
<td>whitening</td>
<td>90.69</td>
<td>89.53</td>
<td>90.55</td>
<td>90.25</td>
</tr>
<tr>
<td>Joint</td>
<td>X</td>
<td>✓</td>
<td>whitening</td>
<td>90.76</td>
<td>89.46</td>
<td>90.91</td>
<td>90.37</td>
</tr>
<tr>
<td>Joint</td>
<td>✓</td>
<td>X</td>
<td>whitening</td>
<td>90.84</td>
<td>89.81</td>
<td>90.81</td>
<td>90.49</td>
</tr>
<tr>
<td>Joint</td>
<td>✓</td>
<td>✓</td>
<td>whitening</td>
<td>91.14</td>
<td>89.75</td>
<td>90.83</td>
<td>90.58</td>
</tr>
</tbody>
</table>


Related work: [Karani et al. MICCAI 2018; Gibson et al. MICCAI 2018; John et al. ISBI 2019]
Tackling Data Heterogeneity with Supervised Learning


Dice Coefficient (mean±std, %)

<table>
<thead>
<tr>
<th>Methods</th>
<th>Site A</th>
<th>Site B</th>
<th>Site C</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tian et al. [26]</td>
<td>88.23</td>
<td>88.23</td>
<td>---</td>
<td>90.56±2.88</td>
</tr>
<tr>
<td>Rundo et al. [9]</td>
<td>---</td>
<td>---</td>
<td>88.66</td>
<td>90.25±3.08</td>
</tr>
<tr>
<td>Separate</td>
<td>90.47±3.00</td>
<td>90.52±2.45</td>
<td>90.70±3.34</td>
<td>90.56±2.88</td>
</tr>
<tr>
<td>Joint</td>
<td>90.69±3.05</td>
<td>89.53±2.97</td>
<td>90.55±3.18</td>
<td>90.25±3.08</td>
</tr>
<tr>
<td>USE-Net [19]</td>
<td>90.90±2.41</td>
<td>90.17±2.61</td>
<td>90.73±2.36</td>
<td>90.60±2.50</td>
</tr>
<tr>
<td>Dual-Stream [47]</td>
<td>90.87±2.85</td>
<td>90.57±2.12</td>
<td>90.10±3.28</td>
<td>90.51±2.72</td>
</tr>
<tr>
<td>Series-Adapter [48]</td>
<td>90.80±2.72</td>
<td>89.92±2.80</td>
<td>91.24±2.21</td>
<td>90.65±2.71</td>
</tr>
<tr>
<td>Parallel-Adapter [23]</td>
<td>90.61±3.54</td>
<td>90.71±2.17</td>
<td>91.30±2.06</td>
<td>90.88±2.79</td>
</tr>
<tr>
<td>DSBN (ours)</td>
<td>90.98±2.69</td>
<td>90.67±2.22</td>
<td>91.07±1.86</td>
<td>90.91±2.36</td>
</tr>
<tr>
<td>MS-Net (ours)</td>
<td>91.54±2.01</td>
<td>91.24±1.97</td>
<td>92.18±1.62</td>
<td>91.66±1.95</td>
</tr>
</tbody>
</table>
Unpaired Multi-modal Learning with Knowledge Distillation

Distill activations per-class:
\[ z^i_c = \frac{1}{|S^i_c|} \sum_n \sum_{(w,h) \in S^i_c} z^i_{nwhi}, \]
\[ p^i_c = \frac{\exp(z^i_c/T)}{\sum_j \exp(z^j_c/T)}, \]

Minimize probability divergence:
\[ \mathcal{L}_{kd} = \frac{1}{C} \sum_c \left( \mathcal{D}_{KL}(q^a_c || q^b_c) + \mathcal{D}_{KL}(q^b_c || q^a_c) \right), \]
where \( \mathcal{D}_{KL}(q^a_c || q^b_c) = \sum q^a_c \log \frac{q^a_c}{q^b_c}. \)

Tackling Data Heterogeneity with Supervised Learning

Tackling Data Heterogeneity with UDA

Unsupervised domain adaptation through pixel-level alignment

Segmenter

Image Transformation Network

Inference


Related work: [Y. Huo et al., ISBI 2018; Z. Zhang et al. CVPR 2018; Y. Zhang et al. MICCAI 2018]
Tackling Data Heterogeneity with UDA

Image-to-image transformation with generative adversarial nets


Related work: [Y. Huo et al., ISBI 2018; Z. Zhang et al. CVPR 2018; Y. Zhang et al. MICCAI 2018]
Unsupervised Domain Adaptation: Feature-level Alignment

Train a source domain segmentation model

- joint cross-entropy loss and dice loss

$$L_{seg} = - \sum_{i=1}^{N^s} \sum_{c \in C} w_c \cdot y_{i,c}^s \log(\hat{p}_{i,c}^s) - \lambda \sum_{c \in C} \frac{\sum_{i=1}^{N^s} 2y_{i,c}^s \hat{y}_{i,c}^s}{\sum_{i=1}^{N^s} y_{i,c}^s \hat{y}_{i,c}^s + \sum_{i=1}^{N^s} \hat{y}_{i,c}^s \hat{y}_{i,c}^s}$$
Unsupervised learning with adversarial loss

domain adaptation module (generator):
$$\min_{\mathcal{M}} \mathcal{L}_\mathcal{M}(X^t, D) = - \mathbb{E}_{(M_A(x^t), F_H(x^t)) \sim P_g} [\mathcal{D}(M_A(x^t), F_H(x^t))]$$

domain critic module (discriminator):
$$\min_D \mathcal{L}_\mathcal{D}(X^s, X^t, \mathcal{M}) = \mathbb{E}_{(M_A(x^t), F_H(x^t)) \sim P_g} [\mathcal{D}(M_A(x^t), F_H(x^t))] - \mathbb{E}_{(M_A(x^s), F_H(x^s)) \sim P_g} [\mathcal{D}(M_A(x^s), F_H(x^s))], \text{s.t. } \|D\|_L \leq K$$


Related work: [K Kamnitsas et al. IPMI 2017]
Unsupervised Domain Adaptation: Feature-level Alignment

Unsupervised learning with adversarial loss

domain adaptation module (generator): \[ \min_{M} \mathcal{L}_{M}(X^t, D) = -\mathbb{E}_{(M_A(x^t), F_H(x^t)) \sim P_g} [D(M_A(x^t), F_H(x^t))] \]

domain critic module (discriminator): \[ \min_{D} \mathcal{L}_{D}(X^s, X^t, M) = \mathbb{E}_{(M_A(x^t), F_H(x^t)) \sim P_g} [D(M_A(x^t), F_H(x^t))] - \mathbb{E}_{(M_A(x^s), F_H(x^s)) \sim P_g} [D(M_A(x^s), F_H(x^s))], \text{s.t. } \|D\|_{L \leq K} \]


Related work: [K Kamnitsas et al. IPMI 2017]
Unsupervised Domain Adaptation: Feature-level Alignment

Related work: [DANN, Ganin et al. JMLR 2016; ADDA, Tzeng et al. CVPR 2017; CycleGAN, Zhu et al. ICCV 2017]
Unsupervised Domain Adaptation: Synergistic Alignment


Related work: [DANN, Ganin et al. JMLR 2016; ADDA, Tzeng et al. CVPR 2017; CycleGAN, Zhu et al. ICCV 2017]
### Cardiac MRI → Cardiac CT

<table>
<thead>
<tr>
<th>Method</th>
<th>Dice</th>
<th>ASD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised training</td>
<td></td>
<td></td>
</tr>
<tr>
<td>W/o adaptation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PaS-AdaNet [44]</td>
<td>28.4</td>
<td>27.7</td>
</tr>
<tr>
<td>SysSeg-Net [37]</td>
<td>21.6</td>
<td>20.7</td>
</tr>
<tr>
<td>CyCADA [18]</td>
<td>21.6</td>
<td>20.7</td>
</tr>
<tr>
<td>SIFA (Ours)</td>
<td>21.6</td>
<td>20.7</td>
</tr>
</tbody>
</table>

### Abdominal MRI → Abdominal CT

<table>
<thead>
<tr>
<th>Method</th>
<th>Dice</th>
<th>ASD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised training</td>
<td></td>
<td></td>
</tr>
<tr>
<td>W/o adaptation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SysSeg-Net [37]</td>
<td>21.6</td>
<td>20.7</td>
</tr>
<tr>
<td>CyCADA [18]</td>
<td>21.6</td>
<td>20.7</td>
</tr>
<tr>
<td>SIFA (Ours)</td>
<td>21.6</td>
<td>20.7</td>
</tr>
</tbody>
</table>
Harmonizing Transferability and Discriminability for Adapting Object Detectors

HTCN: Hierarchical Transferability Calibration Network

(a) Source Only  
(b) SWDA (CVPR 2019)  
(c) HTCN (Ours)
Domain Generalization

**Problem setting:** train on multiple source domains and directly generalize to unseen domains

Regularization for generic semantic features

- adversarial feature alignment for domain invariance [Li et al. ECCV 2018]
- decompose networks parameters to domain-specific/invariant [Khosla ECCV 2012]
- data augmentation based methods [Shankar et al. ICLR 2018; Volpi et al. NeurIPS 2018]
- multi-task or self-supervised signals [Ghifary et al. ICCV 2015; Carlucci et al. CVPR 2019]
Tackling Data Heterogeneity for Domain Generalization

Domain Generalization with Gradient-based Meta-learning

Model-agnostic learning: MAML (model-agnostic meta-learning) [Finn et al. ICML 2017]

Applying to domain generalization:

- MLDG: directly applying episodic training paradigm [Li et al. AAAI 2018]
- Episodic training with alternative model updates [Li et al. ICCV 2019]

MASF: Model-Agnostic Learning of Semantic Features

Episodic training paradigm

Available domains: \( D = \{D_1, D_2, ..., D_K\} \)
Neural network is composed of: \( F_\psi \circ T_\theta \)

Learning with explicit simulation of domain shift:

At each iteration, split into meta-train \( D_{tr} \) and meta-test \( D_{te} \)
Update the parameters one or more steps with gradient descent:

\[
(\psi', \theta') = (\psi, \theta) - \alpha \nabla_{\psi, \theta} L_{task}(D_{tr}; \psi, \theta)
\]

Then, apply meta-learning step, to enforce certain properties to be exhibited on held-out domain \( D_{te} \), to regularize semantic features

Global Class Alignment

Inter-class relationships concept is domain-invariant and transferable

- In each domain, compute class-specific mean feature vector:
  \[
  \bar{z}^{(k)}_c = \frac{1}{N^{(c)}_k} \sum_{n:y^{(k)}_n=c} F_{\psi'}(x^{(k)}_n) \approx \mathbb{E}_{D_k}[F_{\psi'}(x) \mid y = c]
  \]

- Compute soft label distribution: \( s_c^{(k)} = \text{softmax}(T_{\theta'}(\bar{z}^{(k)}_c)/\tau) \)

- With \((D_i, D_j) \in D_{tr} \times D_{te}\), regularize consistency of inter-class alignment:
  \[
  \ell_{\text{global}}(D_i, D_j; \psi', \theta') = \frac{1}{C} \sum_{c=1}^{C} \frac{1}{2} [D_{\text{KL}}(s^{(i)}_c \parallel s^{(j)}_c) + D_{\text{KL}}(s^{(j)}_c \parallel s^{(i)}_c)]
  \]

(Note: complexity of pairs is controllable via mini-batch sampling in large-scale scenarios.)
MAFS: Model-Agnostic Learning of Semantic Features

Local Sample Clustering

Feature clusters with domain-independent class-specific cohesion and separation

Use a *metric-learning* approach, with an embedding network and operates in semantic feature space:

- obtain a learnable distance function:
  \[ d_\phi(z_n, z_m) = \| e_n - e_m \|_2 = \| M_\phi(z_n) - M_\phi(z_m) \|_2 \]

- metric-learning can rely on contrastive loss [Hadsell et al. CVPR 2006]:
  \[ \ell_{\text{con}}^{n,m} = \begin{cases} d_\phi(z_n, z_m)^2, & \text{if } y_n = y_m \\ (\max\{0, \xi - d_\phi(z_n, z_m)\})^2, & \text{if } y_n \neq y_m \end{cases} \]

- or triplet loss [Schroff et al. CVPR 2015]:
  \[ \ell_{\text{tri}}^{a,p,n} = \max\{0, d_\phi(z_a, z_p)^2 - d_\phi(z_a, z_n)^2 + \xi\} \]
MASF: Model-Agnostic Learning of Semantic Features
MASF: Model-Agnostic Learning of Semantic Features

Medical application of brain tissue segmentation

- data acquisition differences in scanners, imaging protocols, and many other factors
- posing severe limitations for translating learning-based methods in clinical practice
- segmentation of 3 brain tissues: white matter, gray matter and cerebrospinal fluid
- 4 domains corresponding to 4 hospitals
Shape-awareness in MASF scheme for segmentation tasks

- Encourage complete segmentation shape at domain shift
- Learn domain-invariant contour-relevant and background-relevant embedding

Influence of training domain numbers on generalization

Datasets

1. Multi-Modality Whole Heart Segmentation (MMWHS) Challenge
   http://www.sdspeople.fudan.edu.cn/zhuangxiahai/0/mmwhs/
   https://github.com/carrenD/Medical-Cross-Modality-Domain-Adaptation

2. MICCAI 2019 MS-CMRSeg Multi-sequence Cardiac MR Segmentation Challenge
   https://zmiclab.github.io/mscmrseg19/

3. MICCAI iSeg 2019 Challenge 6-month Infant Brain MRI segmentation from Multiple Sites
   http://iseg2019.web.unc.edu

4. ISBI 2019 CHAOS Challenge CT-MRI Abdominal Multi-Organ Segmentation
   https://chaos.grand-challenge.org

5. Prostate Segmentation, with several public datasets,
   i.e., NCI-ISBI 2013 dataset, I2CVB dataset (include multiple sites), PROMISE12 dataset (include multiple sites)

6. Chest X-Ray, with several public datasets,
   i.e., ChestX-ray14 NIH, CheXpert, PadChest, Mimic-CXR
   MIDL 2019: https://openreview.net/forum?id=S1gvm2E-t4
Acknowledgement
Thanks for your attention!

Q & A