CSCI 1850: Deep Learning in Genomics
Spring 2021

http://cs.brown.edu/courses/csci1850

April 08, 2021
Thursday

Instructor: Ritambhara Singh
Format: Online (Synchronous)
Time: TTh 10:30-11:50 AM
Section V: Other interesting applications
MAGAN: Aligning Biological Manifolds

• Motivation

• MAGAN Architecture and loss functions

• MNIST + Simulation results

• Class activity

• Biological Results

• Muddy points
Study of single-cells helps understand diseases

Measurements

- Gene Expression
- Open regions
- 3D Structure
- Microscopy Images

DNA

Blood stem cell
Myeloid stem cell
Lymphoid stem cell

Myeloblast
Lymphoblast

Red blood cells
Platelets
White blood cells
Data integration of single-cells is important.
Integration of single data is challenging

- **sci-RNA-seq**
  - Gene Expression
  - Usually No 1-1 correspondence between cells

- **sci-ATAC-seq**
  - Open regions
  - Usually No 1-1 correspondence between features
Think of it as comparing documents in different languages.

Usually No 1-1 correspondence between documents.

Usually No 1-1 correspondence between words.

Note: Completely made-up example.
Integrate single-cell data by projecting to a shared manifold

Where have we heard the term “manifold” before?

How can we use this alignment for downstream analysis?

Assumption: Data shares a common manifold structure
Questions?
MAGAN: Aligning Biological Manifolds

Matthew Amodio
matthew.amodio@yale.edu Smita Krishnaswamy
smita.krishnaswamy@yale.edu
GAN (Generative Adversarial Network): Recap

CycleGAN

apple → orange

orange → apple

Image courtesy: https://hardikbansal.github.io/CycleGANBlog/
CycleGAN

apple → orange

orange → apple

Image courtesy: https://hardikbansal.github.io/CycleGANBlog/
CycleGAN

Why include the autoencoder part (cycle)?
Superimposing Manifolds

Manifold A
- Helper T cells
- Killer T cells
- Regulatory T cells

Manifold B
- Helper T cells
- Killer T cells
- Regulatory T cells

Aligning Manifolds

Manifold A
- Helper T cells
- Killer T cells
- Regulatory T cells

Manifold B
- Helper T cells
- Killer T cells
- Regulatory T cells
MAGAN: Loss functions

\[ x_{12} = G_{12}(x_1) \]
\[ x_{121} = G_{21}(x_{12}) \]
\[ L_r = L_{\text{reconstruction}} = L(x_1, x_{121}) \]
\[ L_d = L_{\text{discriminator}} = -\mathbb{E}_{x_1 \sim P_X} \left[ \log D_2(x_{12}) \right] \]
\[ L_c = L_{\text{correspondence}} = L(x_1, x_{12}) \]
\[ L_{G_1} = L_r + L_d + L_c \]
MAGAN: Loss functions

\[ x_{21} = G_{21}(x_2) \]
\[ x_{212} = G_{12}(x_{21}) \]
\[ L_r = L(x_2, x_{212}) \]
\[ L_d = -\mathbb{E}_{x_2 \sim p_{x_2}} \left[ \log D_1(x_{21}) \right] \]
\[ L_c = L(x_2, x_{21}) \]
\[ L_{G_2} = L_r + L_d + L_c \]
MAGAN: Loss functions

\[ L_{D_1} = - \mathbb{E}_{x_1 \sim P_{X_1}} \left[ \log D_1(x_1) - \log D_1(x_{121}) \right] - \mathbb{E}_{x_2 \sim P_{X_2}} \left[ \log (1 - D_1(x_{21})) \right] \]
MAGAN: Loss functions

\[ L_{D_1} = -\mathbb{E}_{x_1 \sim P_{X_1}} [\log D_1(x_1) - \log D_1(x_{121})] \]
\[ - \mathbb{E}_{x_2 \sim P_{X_2}} [\log(1 - D_1(x_{21}))] \]

\[ L_{D_2} = -\mathbb{E}_{x_2 \sim P_{X_2}} [\log D_2(x_2) - \log D_2(x_{212})] \]
\[ - \mathbb{E}_{x_1 \sim P_{X_1}} [\log(1 - D_2(x_{12}))] \]
MAGAN: Correspondence Loss

Unsupervised Correspondence

e.g. a subset of shared features measured in both experiments
MAGAN: Correspondence Loss

Semi-supervised Correspondence

e.g. easy to acquire a very small number of labeled pairs

\[ L_c = MSE(G_{12}(x_{1i}), x_{2j}) + MSE(G_{21}(x_{2j}), x_{1i}) \]
MAGAN: Architecture

Generator: 3 layers

Discriminator: 5 layers

w/ Leaky ReLU
Questions?
Results: Simulated Data

**GAN w/o Correspondence**

\[X_1\]  \[\mathcal{G}_{12}(X_1)\]  \[X_2\]

**MAGAN**

\[X_1\]  \[\mathcal{G}_{12}(X_1)\]  \[X_2\]

Super imposing

Aligning
The semi-supervised correspondence loss with just a single labeled pair.
Questions?
Class activity [10 mins]

Single cell image data helps us capturing cell-cycle stages

Think:

• What could be the application of aligning single-cell image data with single-cell gene expression data?

• Suppose you use some alignment algorithm to align these datasets, how would you validate your alignment results? (e.g. what information would you need?)

Pair

Share: 
https://docs.google.com/document/d/1bKOgeGGwq9rlzMZ1kVEpadkTysvUXNtjG2zCrde9po/edit?usp=sharing

Questions?
Results: Biological data

Measurements of protein abundance

- 2 batches
- 75,000 cells
- 34 proteins

- Correspondence loss calculated using features

Which correspondence loss setting is this (according to definitions in the paper? (a) Unsupervised (b) Semi-supervised
Results: Biological data (Setting 1)
Results: Biological data (Setting 1)
Results: Biological data (Setting 2)

Measurements of protein abundance

- 2 experiments
- Exp1: 35 proteins
- Exp2: 31 proteins
- 16 common protein measurements

- Correspondence loss calculated using 15 common protein measurements

- Left 1 common protein out for validation

Which correspondence loss setting is this (according to definitions in the paper? (a) Unsupervised (b) Semi-supervised
Results: Biological data (Setting 2)

Measurements of protein abundance

- 2 experiments
- Exp1: 35 proteins
- Exp2: 31 proteins
- 16 common protein measurements
- Correspondence loss calculated using 15 common protein measurements
- Left 1 common protein out for validation
Results: Biological data (Setting 3)

Measurements of protein abundance + gene expression
- 1 experiment with 2830 measurements
- 12 measurements: Protein abundance
- 12496 measurements: Gene expression

Correspondence loss calculated using 10 cells

Which correspondence loss setting is this (according to definitions in the paper)?
(a) Unsupervised
(b) Semi-supervised
Results: Biological data (Setting 3)

Measurements of protein abundance + gene expression

- 1 experiment with 2830 measurements
- 12 measurements: Protein abundance
- 12496 measurements: Gene expression

Correspondence loss calculated using 10 cells

<table>
<thead>
<tr>
<th>Paired CyTOF &amp; scRNA-seq</th>
<th>Without Correspondence Loss</th>
<th>With Correspondence Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE($x_1, G_{21}(x_2)$)</td>
<td>99.3</td>
<td>22.0</td>
</tr>
<tr>
<td>MSE($x_2, G_{12}(x_1)$)</td>
<td>33.7</td>
<td>7.1</td>
</tr>
</tbody>
</table>
Questions?
Integrated gradients (Muddy points)

\[
\phi_i^{IG}(f, x, x') = \left( x_i - x'_i \right) \times \int_{\alpha=0}^{1} \frac{\delta f(x' + \alpha(x - x'))}{\delta x_i} d\alpha
\]

From baseline to input...

Uniformly scale from baseline to input image

Baseline (all zeros)

Input image

(\(\alpha = 0\) )

(\(\alpha = 0.3\) )

(\(\alpha = 1\) )

Muddy point: Were the proteins target proteins or proteins found in the drugs?
Decagon Decoder (Muddy points)

Tensor factorization

\[ g(v_i, r, v_j) = \begin{cases} z_i^T D_{r} z_j & \text{if } v_i \text{ and } v_j \text{ are drugs} \\ z_i^T M_r z_j & \text{if } v_i \text{ and } v_j \text{ are both proteins, or, } v_i \text{ and } v_j \text{ are a protein and a drug} \end{cases} \]
Questions?
Upcoming

Course website: [http://cs.brown.edu/courses/csci1850](http://cs.brown.edu/courses/csci1850)

• Final project presentations next week (April 13 and April 15) – sign up!
  • Additions over mid-term – Sequences and interpretation
  • 7-minute presentations (~7 slides) + Q/A
  • Practice the timing with your teammates
  • Everyone is highly encouraged to attend and ask questions!

• Final project reports and code due on **April 20 at 11:59 PM**
  *(hard deadline)*
What was the clearest point today?

What was the muddiest point today?

https://forms.gle/wWWx74yJWfbTLV2z6