Towards generalizable image analysis for systematic cell biology

Alan Moses
Outline

• Introduction:
  – Subcellular localization of proteins
  – Features and classifiers

• Self-supervised feature learning
  – Paired-cell inpainting
    Lu et al. *PLoS Comp Bio.* 2019

• Change detection
  – Classification “out-of-sample”
    Lu et al. *NeurIPS* 2019
  – Unsupervised approaches
    Lu et al. *eLife* 2018

Alex Lu is an expert on recent ML research in this area
Proteins go all over the cell

The cell appears red

one protein appears green

Yeast cells (Chong et al. Cell 2015)

The cell appears red and blue

Human cells (Thul et al. Science 2017)
Big data for cell biology

- GFP-tagging of ~4000 yeast proteins in 2003
- Automated microscopy and genetics
- Protein sub-cellular localization tells us a lot about protein function

Can we define a “localization pattern” for every protein? (in every cell type and condition?)
Proteome-scale microscopy data

Per experiment:
- Raw: $10^5$ images, 50GB
- Processed: $10^6$ to $10^7$ data points

$10^2$ image features

$10^1$ to $10^2$ cells (i.i.d.)

$10^3$ to $10^4$ proteins

“multiple instances”

Can we define a “localization pattern” for every protein? (in every cell type and condition?)

Human Protein Atlas
- 12,068 Proteins
- 81,312 Images
- 638,640 Single Cells

Thul et al. Science 2017

https://www.proteinatlas.org/
Proteome-scale microscopy data

Per experiment:
- Raw: $10^5$ images, 50GB
- Processed: $10^6$ to $10^7$ data points

$10^2$ image features

$10^1$ to $10^2$ cells (i.i.d.)

“multiple instances”

$10^3$ to $10^4$ proteins

- The “classification paradigm”:
  - Given a “training set” of proteins with “known” localization classes
  - Select features and train a classifier to assign a “localization” to all unknown proteins

- Need a training set
  - “experts” will “annotate” images

- Need features
  - “experts” will design these ahead of time

Boland & Murphy *Bioinformatics* 2001
Chen & Murphy *Bioinformatics* 2007
Reviewed in Newberg & Murphy 2009
one protein appears green
The cell appears red and blue

• Images were classified to one or more of 29 categories by “experts”
  Thul et al. Science 2017

• Images were reanalyzed by ~60,000 gamers in EVE online
  Sullivan et al. Nat. Biotech. 2018

• Subject of a kaggle competition, no method was comparable to experts
  Ouyang et al. Nat. Methods 2019

Proteins go where ever they want: many mixed, quantitative patterns
Proteome-scale microscopy data

Per experiment:
  Raw: $10^5$ images, 50GB
  Processed: $10^6$ to $10^7$ data points

$10^2$ image features

$10^1$ to $10^2$ cells (i.i.d)
“multiple instances”

$10^3$ to $10^4$ proteins

- The “classification paradigm”:
  - Given a “training set” of proteins with “known” localization classes
  - Select features and train a classifier to assign a “localization” to all unknown proteins

- Need a training set
  - “experts” will “annotate” images

- Need features
  - “experts” will design these ahead of time

Turns out to be a lot of work for “experts” and needs to be **redone** for each dataset/experiment

What about “discovery” of unseen patterns?
Proteome-scale microscopy data

Per experiment:
- Raw: $10^5$ images, 50GB
- Processed: $10^6$ to $10^7$ data points

- $10^2$ image features
- $10^1$ to $10^2$ cells (i.i.d.)
- "multiple instances"

- $10^3$ to $10^4$ proteins

- The “classification paradigm”:
  - Given a “training set” of proteins with “known” localization classes
  - Select features and train a classifier to assign a “localization” to all unknown proteins

- Need training:
  - "experts" will do this

- Need features:
  - "experts" will design these ahead of time

Unsupervised learning/clustering in the feature space

Self-supervised CNNs learn features for each dataset
Context Encoders: Feature Learning by Inpainting

Deepak Pathak  Philipp Krähenbühl  Jeff Donahue  Trevor Darrell  Alexei A. Efros
University of California, Berkeley
{pathak,philkr,jdonahue,trevor,efros}@cs.berkeley.edu

Abstract

We present an unsupervised visual feature learning algorithm driven by context-based pixel prediction. By analogy with auto-encoders, we propose Context Encoders – a convolutional neural network trained to generate the contents of an arbitrary image region conditioned on its surroundings. In order to succeed at this task, context encoders need to both understand the content of the entire image, as well as produce a plausible hypothesis for the missing part(s). When training context encoders, we have experimented with both a standard pixel-wise reconstruction loss, as well as a reconstruction plus an adversarial loss. The latter produces much sharper results because it can better handle multiple modes in the output. We found that a context encoder learns a representation that captures not just appearance but also the semantics of visual structures. We quantitatively demonstrate the effectiveness of our learned features for CNN pre-training on classification, detection, and segmentation tasks. Furthermore, context encoders can be used for semantic inpainting tasks, either stand-alone or as initialization for non-parametric methods.
Paired-cell inpainting

Cells in the same image usually have similar patterns

Lu et al. PLoS Comp Bio. 2019
Paired-cell inpainting

Lu et al. *PLoS Comp Bio.* 2019
• The CNN can actually learn to do this for Many subcellular localization patterns
• Robust to morphological differences
• Not just memorizing: these are all examples where the pair of cells was never seen during training, and actually the localization pattern is totally different
Paired-cell inpainting

Lu et al. *PLoS Comp Bio.* 2019
Outperforms other feature sets

Single yeast cell classification benchmark

Overall accuracy approaches supervised CNN

Lu et al. *PLoS Comp Bio.* 2019
Evidence that self-supervised features are better for discovery

Labels from the single cell benchmark

UMAP of supervised CNN features
Evidence that self-supervised features are better for discovery

Labels from the single cell benchmark

UMAP of supervised CNN features

UMAP of paired-cell inpainting features

Alex Lu unpublished
Evidence that self-supervised features are better for discovery

There are “classes” that the supervised features have learned to *ignore*

Alex Lu *unpublished*
Discovery of rare & classification of difficult patterns in the human proteome atlas

Discovery of rare & classification of difficult patterns in the human proteome atlas


“Nucleolar rim”: very rare pattern with no training data

~12,000 Human Proteins

Averaged Features

Splicing “Speckles”

SYF2

NCL

EN1

ETV4

PARK2

Other “Speckles”
Proteome-scale microscopy data

Per experiment:
- Raw: $10^5$ images, 50GB
- Processed: $10^6$ to $10^7$ data points

10^2 image features

10^1 to 10^2 cells (i.i.d.)
“multiple instances”

10^3 to 10^4 proteins

• The “classification paradigm”:
  - Given a “training set” of proteins with “known” localization classes
  - Select features and train a classifier to assign a “localization” to all unknown proteins

• Need a training set
  - “experts” will “annotate” images

• Need features
  - “experts” will design these ahead of time

Unsupervised learning/clustering in the feature space

Self-supervised CNNs learn features for each dataset

Reviewed in Newberg & Murphy 2009
Paired-cell inpainting features can be learned from any dataset with segmentable cells and two channels.

- All three experiments measured the “same thing”
- Similar class-relationships are learned from visually different microscopy images

Paired-cell inpainting features can be learned from any dataset with segmentable cells and two channels.
Outline

• Introduction:
  – Subcellular localization of proteins
  – Features and classifiers

• Self-supervised feature learning
  – Paired-cell inpainting

• Change detection
  – Classification “out-of-sample”
  – Unsupervised approaches

Alex is an expert on recent ML research in this area
Proteome-scale microscopy data

Per experiment:
Raw: $10^5$ images, 50GB
Processed: $10^6$ to $10^7$ data points

$10^2$ image features

$10^3$ to $10^4$ proteins

$10^1$ to $10^2$ cells (i.i.d)

“multiple instances”

“wt”

“standard conditions”

“control”

Repeat the experiment!

$10^1$ to $10^2$ cells (i.i.d)

“multiple instances”

“drug X”

Changes in protein localization could tell us specific molecular pathways affected by the drug
Proteome-scale microscopy data

Per experiment:
- Raw: $10^5$ images, 50GB
- Processed: $10^6$ to $10^7$ data points

- $10^2$ image features
- $10^3$ to $10^4$ proteins

Repeat the experiment!

“control”

“drug X”

“drug Y”

10^1 to 10^2 cells (i.i.d)

“multiple instances”

Train classifiers and select features on the control then look for proteins/markers that change in each experiment.
In first studies, images were compared by eye.

A novel single-cell screening platform reveals proteome plasticity during yeast stress responses.

Dissecting DNA damage response pathways by analysing protein localization and abundance changes during DNA replication stress.


A novel single-cell screening platform reveals proteome plasticity during yeast stress responses

JCB 2013

1Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, Israel 76100
2Harvard-MIT Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, MA 02139
3Whitehead Institute for Biomedical Research, Cambridge, MA 02142
Why is automated change detection in microscopy images so hard?

Hormone (alpha factor) treatment time course

- Protein is usually in the cytoplasm.
- Occasionally goes to the nucleus.

“Incomplete penetrance” or cell-to-cell variability.
Single-cell classification approach

• We can deal with incomplete penetrance:
  – Train single-cell classifiers based on the control images
  – Look for changes in class frequencies between control and treatment
  – Subcellular localization changes are inferred when the differences exceed sampling variance of a multinomial

  e.g., Chong et al Cell 2015
Mouse cell-lines with one marker each

“control”
N=11,000 cells
(40,000 more were used to train the classifiers, not included here)

“drug X”
N=32,600 cells

“drug Y”
N=30,800 cells

Are these drugs affecting (different) secretory pathways?

With Wiebke Schormann and David Andrews
Classifiers are letting us down

• Actually no treatments were given to the cells
  – “drug X” was a different day
  – “drug Y” was a different microscope
• “co-variate shift” or “out-of-sample” effect leads to performance degradation
Classifiers are letting us down

• Actually no treatments were given to the cells
  – “drug X” was a different day
  – “drug Y” was a different microscope
• “co-variate shift” or “out-of-sample” effect leads to performance degradation

UMAP of paired-cell inpainting features (learned on the “control” data)

Classes are still well-separated in the feature space, but decision boundary was crossed

Alex Lu unpublished
Classifiers are unreliable on new data sets

- COOS-7 dataset designed to test generalization to new datasets  
  Lu et al. NeurIPS 2019
- All classifiers/features failed to generalize  
  Lu et al. NeurIPS 2019

- ML: New “robust” classifiers?

- Unsupervised change detection
Political Messages

Open access: “real scientists do it in public”
www.plos.org  www.biomedcentral.com

Keep peer-review sustainable: \( \frac{\text{papers reviewed}}{\text{papers submitted}} \geq 3 \)

#blessed:  

Yolanda Chong  
Brenda Andrews  
Helena Friesen  
Judice Koh  
Grant Brown

Wiebke Schormann  
David Andrews

Marzyeh Ghassemi

Alex Lu (Graduating this spring)
Amy Lu (now at insitro)
Louis-Francois Handfield (now at Sanger)
Covariate shifts, out-of-sample effects affect markers/patterns/cell lines differently

Our approach: local statistics

Handfield et al. *Bioinformatics* 2014
Lu et al. *eLife* 2018
Features seem highly robust to morphological variation
K nearest neighbours (or adaptive bandwidth kernel regression) can be used to compute local expectation. In practice, we compare the observations for protein p to 50 most similar proteins.
Why is automated change detection in microscopy images so hard?

Cancer drug (Ramapycin) treatment time course

Before

Cells now have multiple vacuoles, rarely seen before the drug

1 hour

Cells now mostly have multiple vacuoles, and some have giant vacuoles

5 hours

Dark areas are vacuoles

These changes affect all the cells, and quantitatively (and differently) affect all the protein localization patterns