Explaining gene expression variation to neural network derived image features

Anonymous Author(s)
Affiliation
Address
email

Abstract

How does gene expression influence visual characteristic of tissues? Can the variation in expression be explaining using data extracting from these images alone? To what extent is this possible? In this work, I am to answer these questions, by employing neural networks on a dataset of biomedical images annotated with genotype and gene expression data.

1 Introduction

Biomedical images are routinely used by doctors in pathology to diagnose diseases. For example, lung biopsies are taken from patients when grading different levels of lung cancers. Specifically for lung cancer, a great deal of work has focussed on the identification of specific tissue characteristics which are indicative of metastasis. Indeed, this has been the topic of many supervised machine learning competition and approaches. We can now accurately identify metastatic regions from high resolution histopathology lung images, with the state of the art being achieve by Convolutional Neural Networks (CNNs).

What other types of information can be estimated from these high resolution images? We explore this direction using a datasets of.....

2 Results

We find that across 10 different tissue types

Plan

Q1: How much variability is shared between image features and expression?

1. ~~Calculate PCA of image features and expression~~
2. ~~Calculate cross correlations between the PCs until 99.9% variance is explained.~~
3. ~~Use the cross correlation measurements to derive total fraction of variance explained of images by expression: calculate $\sum_i \lambda_i \sum_j r_{ij}^2$ where λ_i is the ith eigenvalue of the image feature PCA, r_{ij}^2 is the correlation of image feature PC i with expression feature PC j.~~
4. ~~Make all of the above one-line-to-run for any combination of patch size, feature layer, tissue~~
5. ~~Report the value for each patch size, feature layer, tissue;~~ interpret findings

Submitted to 31st Conference on Neural Information Processing Systems (NIPS 2017). Do not distribute.

Q2: How much of the shared variability is due to technical confounders?

1. ~~Regress out known confounders C from the image feature F and expression data E. I.e., fit models $F \sim C + \epsilon$ and $E \sim C + \epsilon$. Use log scale for Ischemic time, monitor all scatter plots for confounders that are included in the model that the linear model assumptions hold. Decide whether to retain full model of all confounders x all genes/features, retain only significant associations, or do forward feature selection.~~
2. ~~Repeat steps Q1:1-5 above using the residuals of this model.~~
3. ~~Create a scatter plot of total shared variance vs. technical shared variance [again, all tissues etc.]~~
4. Interpret findings in context of many tissues, scales, feature layers

Q3: What genes share expression variability to visual features, and at which scales?

1. ~~Using residuals of technical confounders, fit linear model of expression vs. image feature - calculate effect sizes and p-values~~
2. ~~Apply multiple testing correction to derive a list of image feature associated genes.~~
3. ~~Sort genes based on variance explained by each image feature, perform GSEA or gProfiler ranked gene list analysis to test whether particular aspects of biology are enriched.~~ Interpret
4. Assess number of associations [total and per-feature] for different tissues, feature scales. Assess overlap of findings from gene list analyses. Interpret.

Q3A: What genes are influenced by technical confounders?

1. For each confounder, calculate the fraction of variance it explains for each gene expression level.
2. Sort genes based on variance explained by confounder, perform GSEA or gProfiler ranked gene list analysis to test whether particular aspects of biology are enriched.]

Once these done

- Q4: What do the image features represent?
- Q5: Is there a genetic basis to the image features?
- Q6: Are there gene expression levels that cause image feature changes, or vice versa?
- Q7: Are there image features that are associated with clinical annotations? Are any causal?

Methods

Assessing shared variability between image features and expression

We calculate principle components that capture 99.9% of variation in the image features and expression features respectively. We repeat this after fitting a linear model containing known technical confounders to both the image features and expression, and regressing out these effects.

We choose the top 2000 transcripts with mean expression greater than one, and look for associations with the image features across tissue samples. For image features generated from Lung tissue, we plot a QQ-plot with

Results

Variability in expression explains a substantial amount of variability in the image features (Figure ???). Furthermore, almost 50% of the variability in expression, and almost 40% of the variability in the image features, is explained by 51 known technical factors (Figure ???)

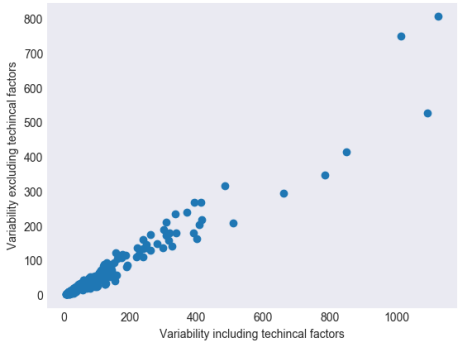

Figure 1: Sample figure caption.

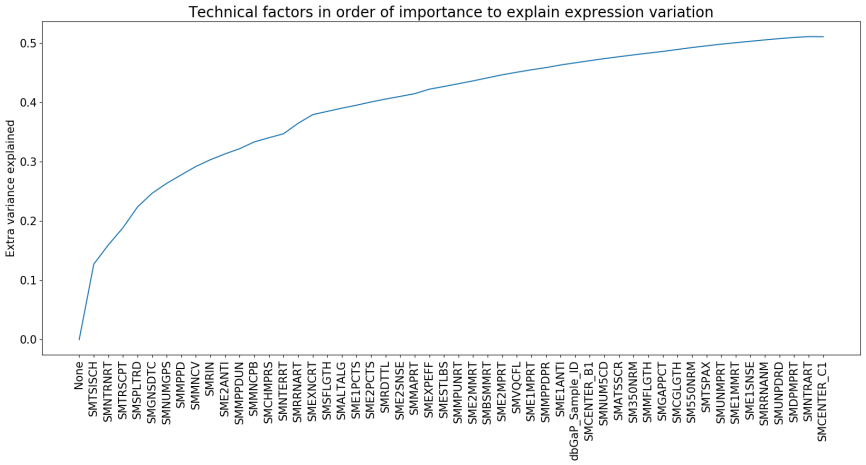

Figure 2: Sample figure caption.

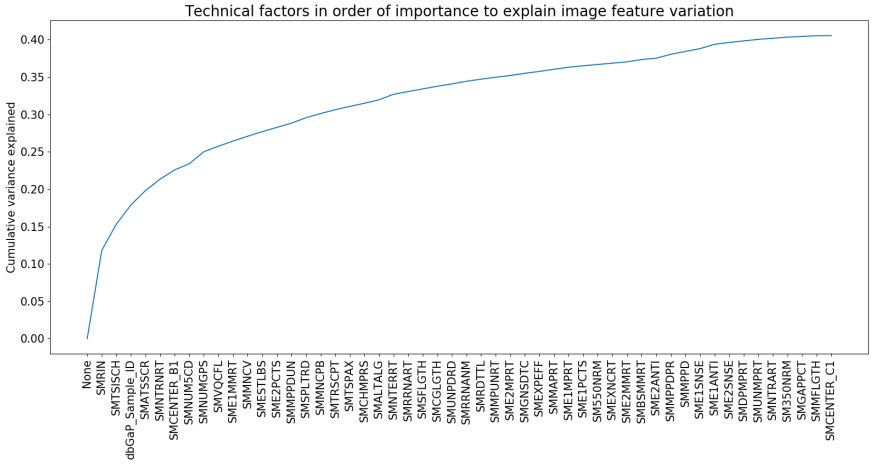

Figure 3: Sample figure caption.

Table 1: Enriched Ontology terms for image feature 501

p-value	Ontology	Genes
1.21e-15	lamellar body	LAMP3,CTSH,SFTPA1,NAPSA,SFTPD
6.87e-09	Defective ABCA3 causes pulmonary surfactant metabolism dysfunction type 3 (SMDP3)	SFTPA1,SFTPD,SFTPB,SFTPA2

Table 2: Enriched Ontology terms for image feature 3

p-value	Ontology	Genes
0.0498	BRONCHIECTASIS WITH OR WITHOUT ELEVATED SWEAT CHLORIDE 3; BESC3;;CYSTIC FIBROSIS-LIKE SYNDROME	SCNN1G