Deep Learning of the tissue-regulated splicing code

Leung, Xiong, Lee, Frey
University of Toronto

Deep learning

Black magic behind ⬇

- Object classification in photos
- Natural language processing
- Automatic game playing
- Academic research: biology, finance, health, sports, ….

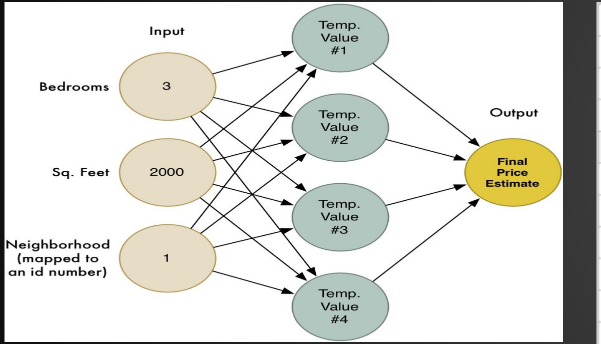
Big players ⬇

- **Google**, Microsoft, Facebook, tons of start-ups
- Yann LeCun, Geoffrey Hinton, Yoshua Bengio, GOD
- OpenAI and the fear of DL

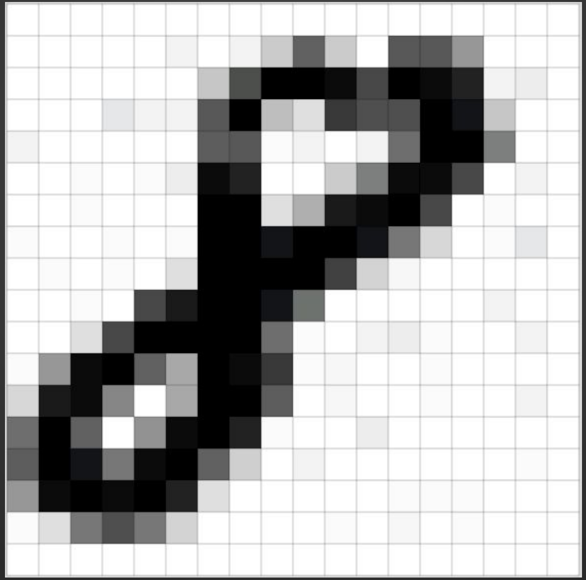
Deep learning Examples

DL Architectures: RNN, LSTM, FFNN, CNN, RBM and many variations

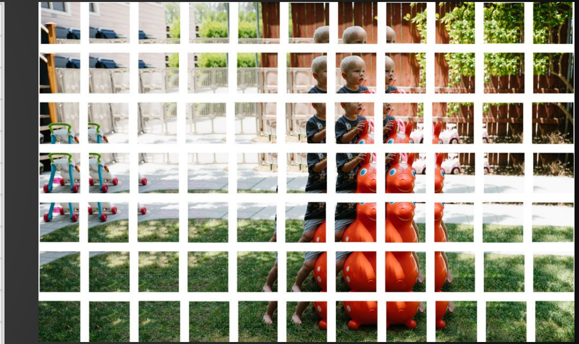
predicting house prices

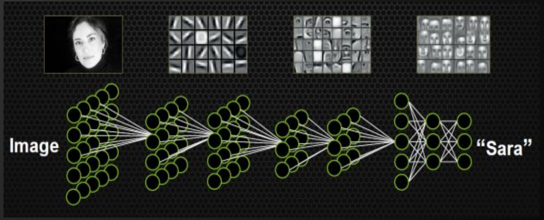


recognizing digits



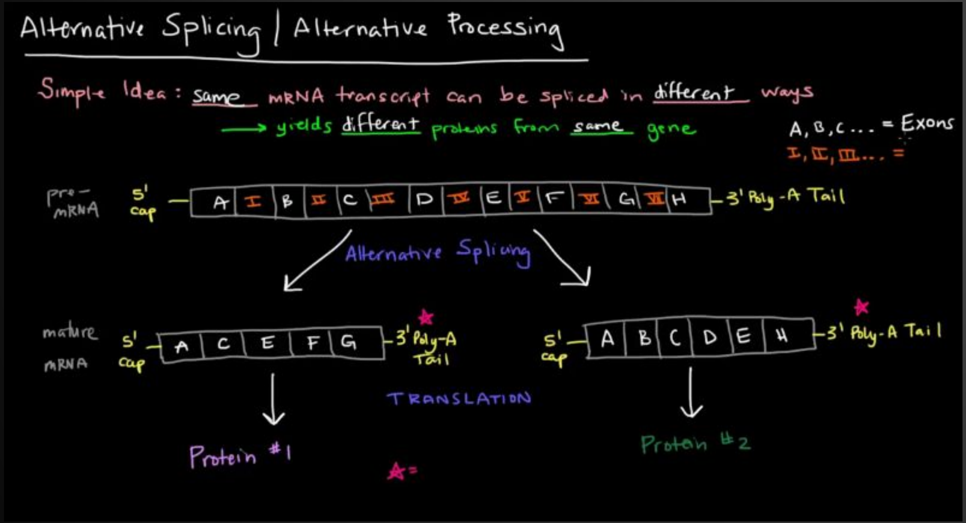
recognizing objects





Alternative splicing (AS)

Same mRNA transcript can be spliced in different ways



Alternative Splicing | Alternative Processing

Simple Idea: same mRNA transcript can be spliced in different ways

→ yields different proteins from same gene

A, B, C ... = Exons

I, II, III ... =

- 1 gene, 1 mRNA transcript
- 1 mRNA transcript, n proteins,
- difference in amino acid sequence, difference in biological functions
- AS and the human genome: synthesis of many more proteins than the 20,000 protein-coding genes

DL applied to AS

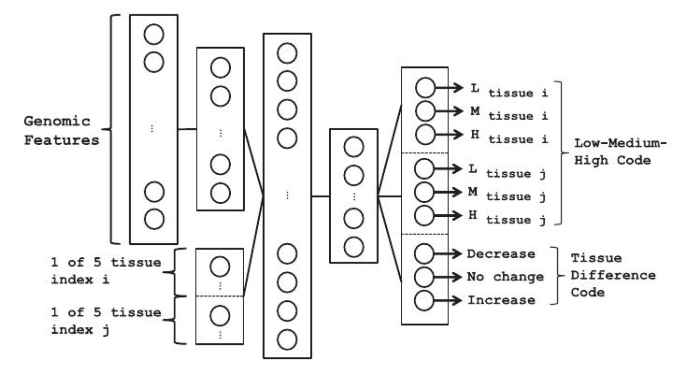
- DNN used to model RNA-Seq data from mouse
- Relatively cheaper computationally
- Works better with sparse data
- Large (many hidden variables) and deep (multiple hidden layers) neural nets can improve the predictive performances of splicing code compared to BNN or MCMC
- PSI (percent splicing index) prediction for each tissue
    - Ratio between exon inclusions and inclusion + exclusion
- Difference in PSI between tissue pairs

Methods & Model

- ➔ No data no love: 11019 mouse exons along with PSI values
- ➔ 5 tissue types: whole brain, heart, kidney, liver and testis
- ➔ Model output: activation of each hidden layer

$$a_v^l = f\left(\sum_m^{M^{l-1}} \theta_{v,m}^l a_m^{l-1}\right)$$

⬆ Output activation of each hidden unit v in layer l: sum of weighted outputs from previous layer using a nonlinear function f



⬅ Input layer: 1393 genomic features describes exon, neighboring introns and adjacent exons

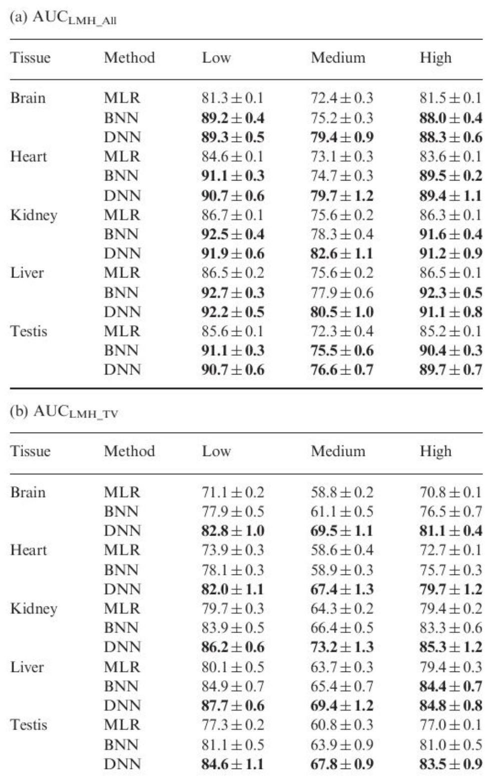
⬇ Outputs of last layer is used as input into a softmax function: represents the probability of each splicing pattern k

$$h_k = \frac{\exp\left(\sum_m \theta_{k,m}^{last} a_m^{last}\right)}{\sum_{k'} \exp\left(\sum_m \theta_{k',m}^{last} a_m^{last}\right)}$$

Training & Predictions

- First hidden layer trained as an autoencoder to reduce dimensionality of the feature space
- AE trained by supplying the input through a nonlinear layer
- Works well with DNNs since nonlinear techniques are more likely to discover better representation of the features
- Prediction: PSI value given a particular tissue type set of genomic features: low, medium, high (LMH). This represents the probability that a given exon and tissue type has a PSI value within these intervals
- Implemented in Python using Gnumpy (GPU library)

Results

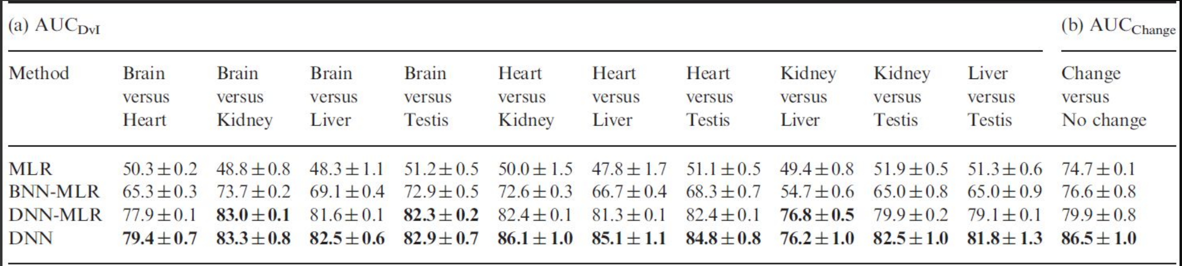


(a) AUC$_{\mathrm{LMH\_All}}$

Tissue	Method	Low	Medium	High
Brain	MLR	81.3 ± 0.1	72.4 ± 0.3	81.5 ± 0.1
	BNN	**89.2 ± 0.4**	75.2 ± 0.3	**88.0 ± 0.4**
	DNN	**89.3 ± 0.5**	**79.4 ± 0.9**	**88.3 ± 0.6**
Heart	MLR	84.6 ± 0.1	73.1 ± 0.3	83.6 ± 0.1
	BNN	**91.1 ± 0.3**	74.7 ± 0.3	**89.5 ± 0.2**
	DNN	**90.7 ± 0.6**	**79.7 ± 1.2**	**89.4 ± 1.1**
Kidney	MLR	86.7 ± 0.1	75.6 ± 0.2	86.3 ± 0.1
	BNN	**92.5 ± 0.4**	78.3 ± 0.4	**91.6 ± 0.4**
	DNN	**91.9 ± 0.6**	**82.6 ± 1.1**	**91.2 ± 0.9**
Liver	MLR	86.5 ± 0.2	75.6 ± 0.2	86.5 ± 0.1
	BNN	**92.7 ± 0.3**	77.9 ± 0.6	**92.3 ± 0.5**
	DNN	**92.2 ± 0.5**	**80.5 ± 1.0**	**91.1 ± 0.8**
Testis	MLR	85.6 ± 0.1	72.3 ± 0.4	85.2 ± 0.1
	BNN	**91.1 ± 0.3**	**75.5 ± 0.6**	**90.4 ± 0.3**
	DNN	**90.7 ± 0.6**	**76.6 ± 0.7**	**89.7 ± 0.7**

(b) AUC$_{\mathrm{LMH\_TV}}$

Tissue	Method	Low	Medium	High
Brain	MLR	71.1 ± 0.2	58.8 ± 0.2	70.8 ± 0.1
	BNN	77.9 ± 0.5	61.1 ± 0.5	76.5 ± 0.7
	DNN	**82.8 ± 1.0**	**69.5 ± 1.1**	**81.1 ± 0.4**
Heart	MLR	73.9 ± 0.3	58.6 ± 0.4	72.7 ± 0.1
	BNN	78.1 ± 0.3	58.9 ± 0.3	75.7 ± 0.3
	DNN	**82.0 ± 1.1**	**67.4 ± 1.3**	**79.7 ± 1.2**
Kidney	MLR	79.7 ± 0.3	64.3 ± 0.2	79.4 ± 0.2
	BNN	83.9 ± 0.5	66.4 ± 0.5	83.3 ± 0.6
	DNN	**86.2 ± 0.6**	**73.2 ± 1.3**	**85.3 ± 1.2**
Liver	MLR	80.1 ± 0.5	63.7 ± 0.3	79.4 ± 0.3
	BNN	84.9 ± 0.7	65.4 ± 0.7	**84.4 ± 0.7**
	DNN	**87.7 ± 0.6**	**69.4 ± 1.2**	**84.8 ± 0.8**
Testis	MLR	77.3 ± 0.2	60.8 ± 0.3	77.0 ± 0.1
	BNN	81.1 ± 0.5	63.9 ± 0.9	81.0 ± 0.5
	DNN	**84.6 ± 1.1**	**67.8 ± 0.9**	**83.5 ± 0.9**



(a) AUC$_{\mathrm{DvI}}$											(b) AUC$_{\mathrm{Change}}$
Method	Brain versus Heart	Brain versus Kidney	Brain versus Liver	Brain versus Testis	Heart versus Kidney	Heart versus Liver	Heart versus Testis	Kidney versus Liver	Kidney versus Testis	Liver versus Testis	Change versus No change
MLR	50.3 ± 0.2	48.8 ± 0.8	48.3 ± 1.1	51.2 ± 0.5	50.0 ± 1.5	47.8 ± 1.7	51.1 ± 0.5	49.4 ± 0.8	51.9 ± 0.5	51.3 ± 0.6	74.7 ± 0.1
BNN-MLR	65.3 ± 0.3	73.7 ± 0.2	69.1 ± 0.4	72.9 ± 0.5	72.6 ± 0.3	66.7 ± 0.4	68.3 ± 0.7	54.7 ± 0.6	65.0 ± 0.8	65.0 ± 0.9	76.6 ± 0.8
DNN-MLR	77.9 ± 0.1	**83.0 ± 0.1**	81.6 ± 0.1	**82.3 ± 0.2**	82.4 ± 0.1	81.3 ± 0.1	82.4 ± 0.1	**76.8 ± 0.5**	79.9 ± 0.2	79.1 ± 0.1	79.9 ± 0.8
DNN	**79.4 ± 0.7**	**83.3 ± 0.8**	**82.5 ± 0.6**	**82.9 ± 0.7**	**86.1 ± 1.0**	**85.1 ± 1.1**	**84.8 ± 0.8**	**76.2 ± 1.0**	**82.5 ± 1.0**	**81.8 ± 1.3**	**86.5 ± 1.0**

- ➔ AUC (area under the curve) comparison of multinomial linear regression, bayesian neural network, and deep neural network
- ➔ DNN predicts both LMH and DNI at the same time, but BNN can only predict LMH
- ➔ Methods were combined for fair comparison

Conclusion

- Introduced a computational model that extends the previous splicing models with new prediction targets and improved tissue-specificity
- Used a learning algorithm that scales well with high volume data and large number of hidden variables
- DNNs can be trained rapidly with the aid of GPUs
- Deep architectures can be beneficial when dealing with sparse biological datasets
- Input features can be analyzed in terms of the predictions of the model to gain insights into inferred tissue-regulated splicing code