UTSA
The University of Texas at San Antonio™

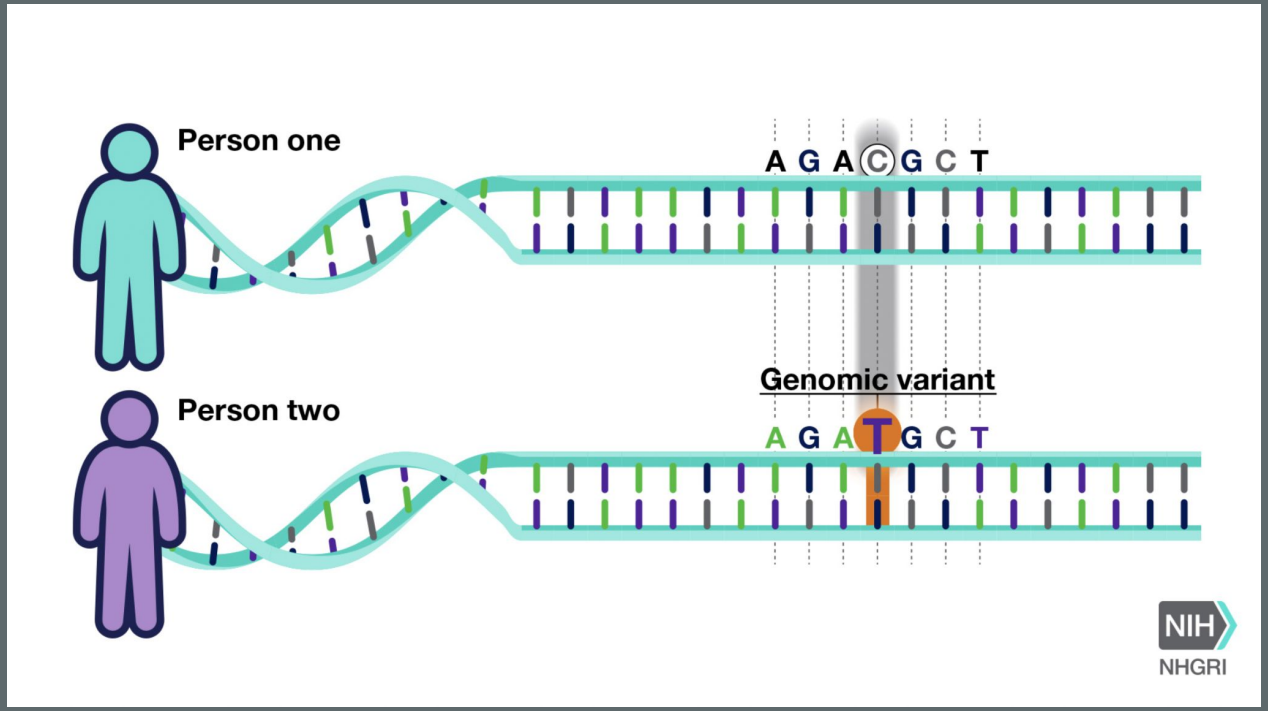
Protein Classification in Genomic Data with Deep Learning Network

Sthefanie Jofer Gomes Passo

Outline

- **Introduction**
  - Genome data and the challenge to manage it
- **Method**
  - Data and Preprocessing
- **Classification Model**
- **Experiments**
- **Conclusions and Future Research**

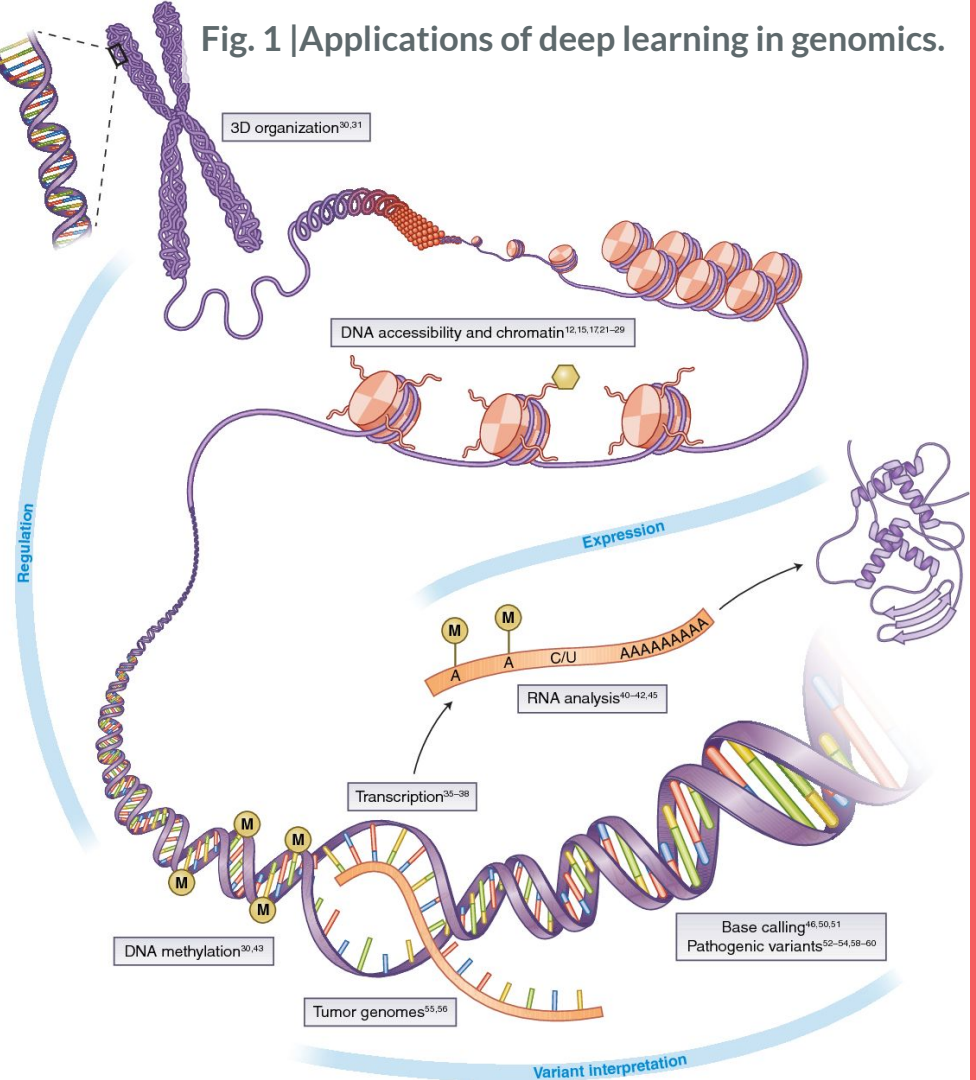
Introduction



Predicting phenotypes from genetic data is also a major area of interest in deep learning.

A first step in performing these types of predictions is to specify what genetic variants are present in an individual genome.

This **problem** has been addressed by DeepVariant, which applies a CNN to make variant calls from short read sequencing.



Can we apply genomic data in Machine Learning?

Deep learning methods are a class of machine learning techniques **capable of identifying highly complex patterns in large data-sets to genomic data.**

Similarly, the identification of transcription can be executed with the addition of features from the Encyclopedia of **DNA Elements (ENCODE) project**, as well as transcription-start-site sequencing and RNA-seq signals.

The methylation state of DNA, which also influences the **expression of genes**, has been inferred from three-dimensional genome topology (on the basis of Hi-C) and DNA sequence patterns.

Method

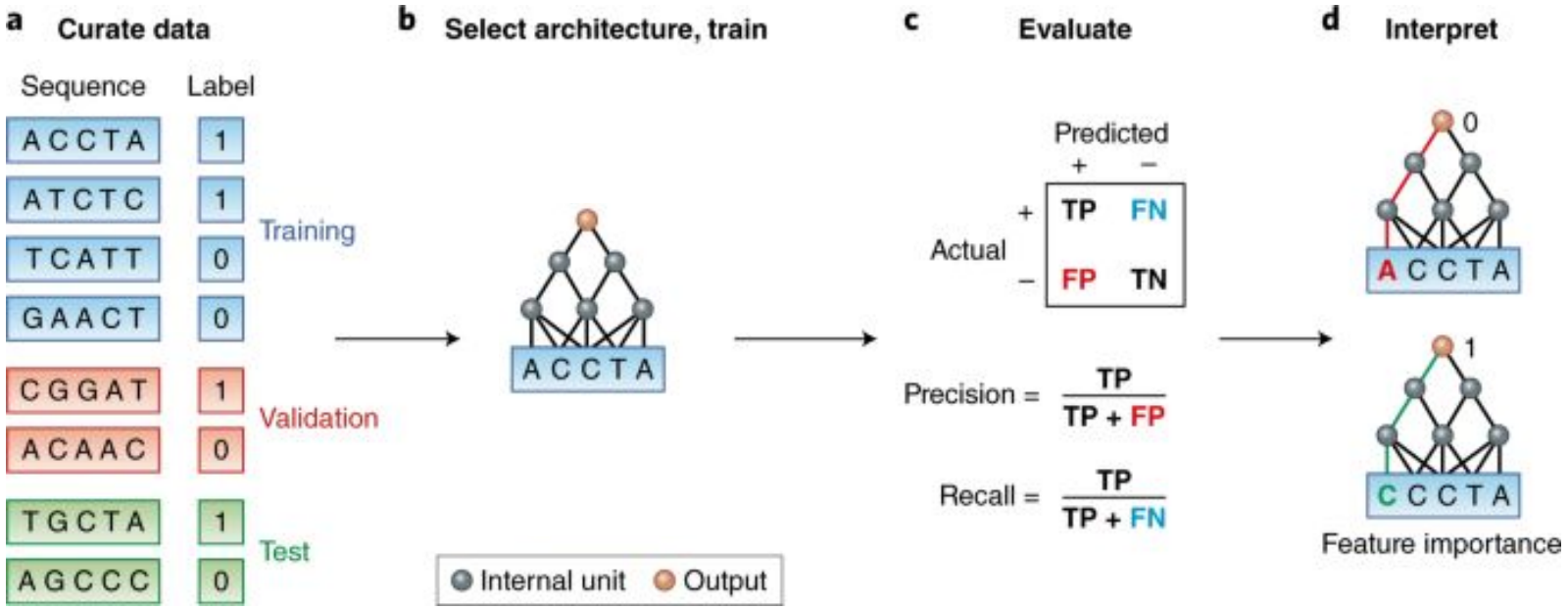


Fig. 2 |Deep learning workflow in genomics.

Method

DNA Sequence							
C	C	G	...	A	G	T	A

One hot encoding of Sequence							
0	0	0	...	1	0	0	1
1	1	0	...	0	0	0	0
0	0	1	...	0	1	0	0
0	0	0	...	0	0	1	0

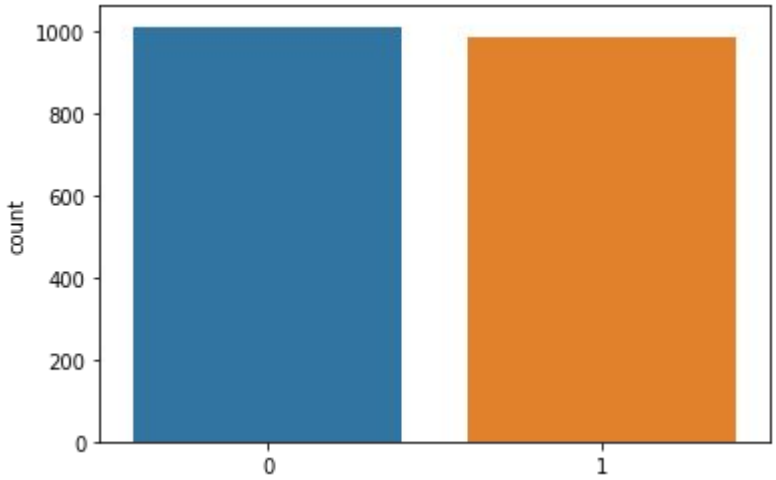
Preprocessing:

- **Encyclopedia of DNA Elements (ENCODE) project**
- We had 2000 samples
- 50 feature of the bases A, C, G and T from the genoma
- After we apply the one hot encoding the data that was (2000, 50, 1) became (2000, 50, 4)
- 25% of the features are separated to test the network (ZOU, 2019)

Method

Data Balance:

Figure 3: Genomas that have certain protein or not (*CGACCGAACTCC*)



- How long it takes to exactly search in a combinatory (brute force) way or high level way (machine learning)?
  - Combinatory Analysis:

$$\frac{50!}{11!(50-11)!} = \frac{5.81506\ldots E19}{479001600} = 121399651099.99998 = 12 \times 10^{10}$$

  - Machine learning:

$$CostFunction = Sum[(Actual - Predicted)^2] \times (\frac{1}{NumberOfObservations})$$

$$CostFunction = Sum[(500-400)^2] \times (\frac{1}{50}) = 10000 \times \frac{1}{50} = 200$$

- Human don't do exhausted researches

Classification Model

Figura 7: State of Art with Modified Dimensionalities.

ZOU (2019)	+ Average Pool	+ Global Max	+ Global Average
Convolution 1D (39 x 32)	Convolution 1D (39 x 32)	Convolution 1D (39 x 32)	Convolution 1D (39 x 32)
Max Pooling 1D (9 x 32)	Average Pooling 1D (9 x 32)	Global Max Pooling 1D (9 x 32)	Global Average Pooling 1D (9 x 32)
Flatten (288)	Flatten (288)	Flatten (288)	Flatten (288)
Dense (16)	Dense (16)	Dense (16)	Dense (16)
Dense (2)	Dense (2)	Dense (2)	Dense (2)

Max Pooling			
2	8	3	6
8		**6**	

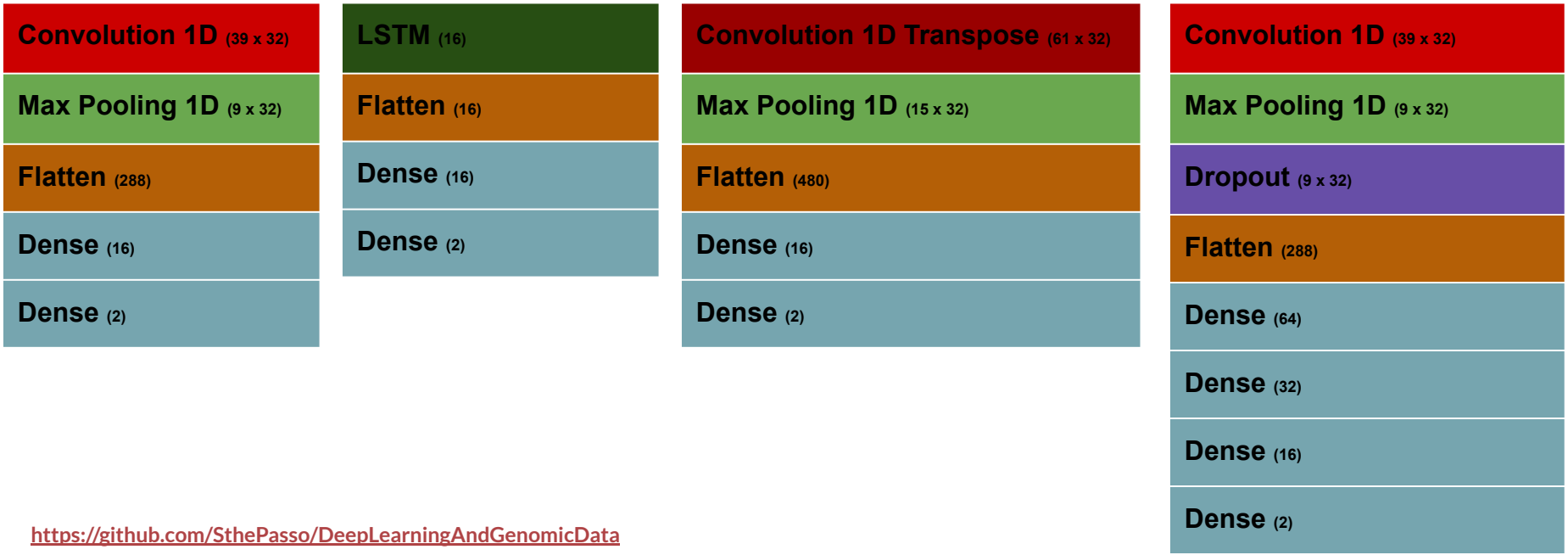
Average Pool			
2	8	3	6
5		**4.5**	

Global Max			
2	8	3	6
8			

Global Average			
2	8	3	6
4.75			

Classification Model

Figura 7: State of Art with Modified Techniques



https://github.com/SthePasso/DeepLearningAndGenomicData

Experiments

Comparing the genome classifications models with **modified dimensionalities**.

Method	Classe	TP	TN	FP	FN	Acc(%)	se(%)	sp(%)	+p(%)
ZOU [2019]	No	248	239	11	2	**98.20**	**99.21**	**97.15**	**97.30**
	Yes	239	248	2	11		**97.15**	**99.21**	**99.17**
Average Pooling	No	200	177	59	64	75.40	75.76	75.00	77.22
	Yes	177	200	64	59		75.00	75.76	73.44
Global Max Pooling	No	183	189	76	52	73.60	75.92	71.37	71.81
	Yes	189	183	52	76		71.37	75.92	75.52
Global Average Pooling	No	183	189	76	52	74.40	77.87	71.32	70.66
	Yes	189	183	52	76		71.32	77.87	78.42

Experiments

Comparing the genome classifications models with **modified techniques**.

Method	Classe	TP	TN	FP	FN	Acc(%)	se(%)	sp(%)	+p(%)
ZOU [2019]	No	248	239	11	2	98.20	99.21	97.15	97.30
	Yes	239	248	2	11		97.15	99.21	99.17
LSTM	No	255	240	4	1	99.00	99.61	98.36	98.46
	Yes	240	255	1	4		98.36	99.61	99.59
Convolution Transpose	No	258	240	1	1	99.60	99.61	99.59	99.61
	Yes	240	258	1	1		99.59	99.61	99.59
Dropout and Dense	No	259	241	0	0	**100.00**	**100**	**100**	**100**
	Yes	241	259	0	0		**100**	**100**	**100**

Conclusions and Future Research

- Different mathematics Techniques will be better for different kinds of datas
- **Application of ODE techniques to replace the Deep Learning network or part of it to my current PhD project**
- Find other genome baselines to similar applicabilities

Thanks for the attention

SOLI DEO GLORIA

Sthefanie.Passo@utsa.edu

https://github.com/SthePasso/DeepLearningAndGenomicData

Referências

AAMI. (1999). Association for the Advancement of Medical Instrumentation and American National Standards Institute, Testing and Reporting Performance Results of Cardiac Rhythm and ST-segment Measurement Algorithms, ANSI/AAMI, The Association, pp. 1–36 https://books.google.it/books?id=gzPdtgAACAAJ.

AHRQ. (2012). Weighted national estimete. HCUP National Inpatient Sample. Agency for Healthcare Research and Quality. [Online] https://www.ahrq.gov/. Acesso em: 25/04/2019.
Banerjee, R., et al. (2014) Photoecg: Photoplethysmographyto estimate ecg parameters. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pgs. 4404–4408.

CDC. CDC 24/7: Saving Lives, Protecting People. (2017). Centers for Disease Control and Prevention. [Online] https://www.cdc.gov/. Acesso em: 14/05/2020.

CDCP. About multiple cause of death 1999-2011. 2014. Centers for Disease Control and Prevention. [Online] https://www.cdc.gov/. Acesso em: 25/04/2019.

Fang, J. et al, (2019) Awareness of Heart Attack Symptoms and Response Among Adults — United States, 2008, 2014, and 2017. MMWR - Morbidity and Mortality Weekly Report. [Online] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6366680/.

Related Works

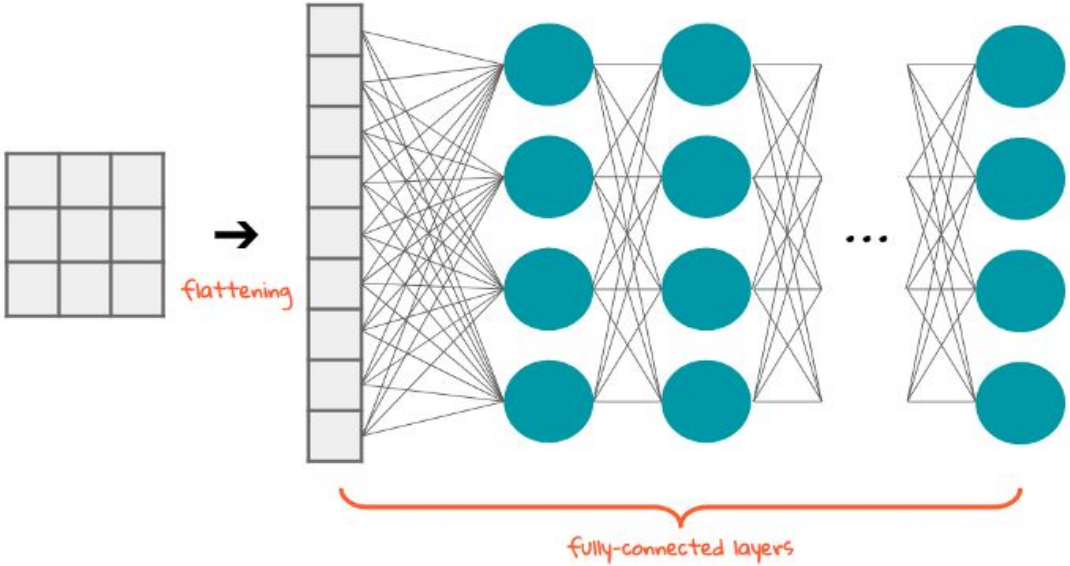
Trabalhos Relacionados

Tabela 1: Resumo dos Trabalhos Relacionados.

Autor	Base	Nº Classes	Método	Paradigma Av.	Aplicação/Notas	Acurácia
Mousavi et al., 2019	MIT	4 e 3	CNN	Intra e Inter Paciente	Classificação com redes profundas utilizando SMOTE	99,92% e 99,53%
Sannio, 2018	MIT	2	DNN	Intra-Paciente	Pré-processamento utilizando técnicas matemáticas	100%
Wu et al., 2018	MIT e DeepQ	5 e 2	CNN	Intra-Paciente	Classificação com redes profundas	93% e 94%
Li et al., 2019	MIT	5	Bi-LSTM Atenção	Intra-Paciente	Classificação com redes profundas	99,49%
Kachuee et al., 2018	MIT e PTB	5 e 2	CNN	Intra-Paciente	Classificação utilizando transf. de aprendizagem/ F: 86% e S:89% acc	93,4% e 95,9%

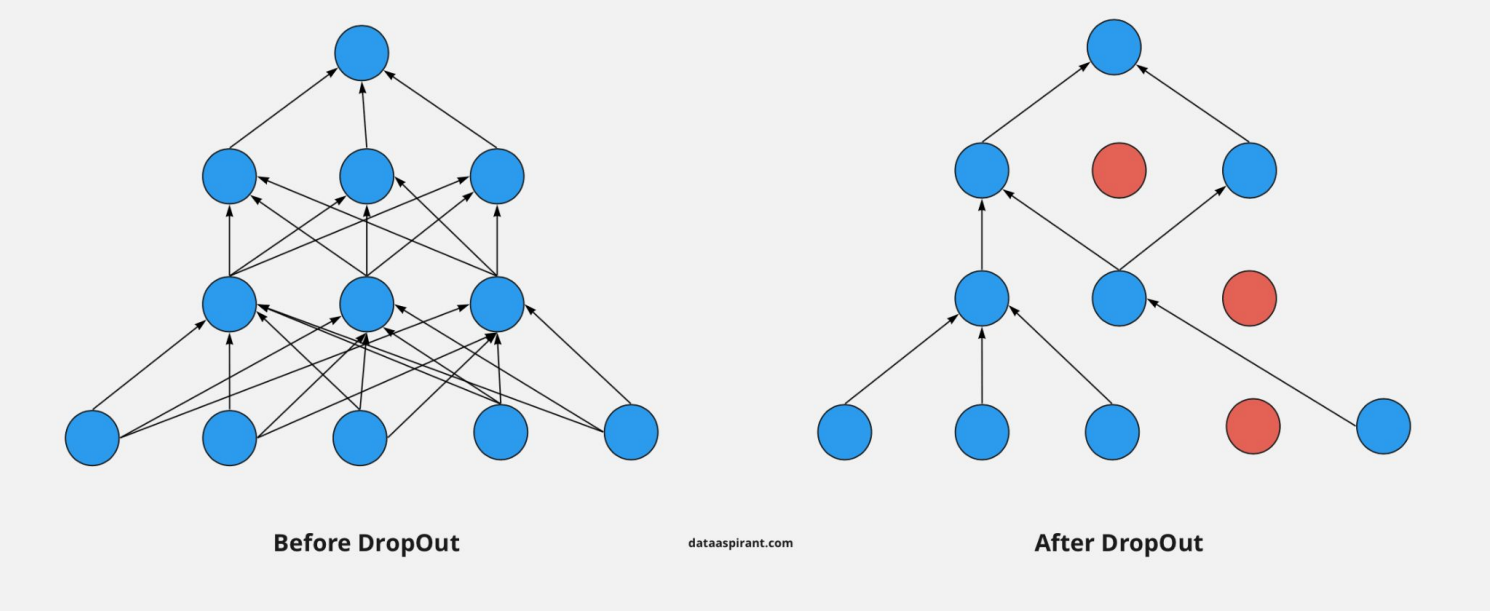
Flatem

Flattening is converting the data into a 1-dimensional array for inputting it to the next layer. We flatten the output of the convolutional layers to create a single long feature vector. And it is connected to the final classification model, which is called a ***fully-connected*** layer. In other words, we put all the pixel data in one line and make connections with the final layer.



Dropout

Dropout has the effect of making the training process noisy, forcing nodes within a layer to probabilistically take on more or less responsibility for the inputs.



Transposed Convolution

A transposed convolution is somewhat similar because it produces the same spatial resolution a hypothetical deconvolutional layer would. However, the actual mathematical operation that's being performed on the values is different. A transposed convolutional layer carries out a regular convolution but reverts its spatial transformation.

-1	0	+1
-2	0	+2
-1	0	+1

x filter

+1	+2	+1
0	0	0
-1	-2	-1

y filter