SNORKEL

YueLiu

What is Snorkel?

Related Concepts

Unsupervised Learning

Unsupervised learning doesn't need to get a collection of labelled target data. Most of them are using clustering method, it can be used to reduce dimension sometimes.

Semi-supervised Learning

A branch of machine learning between supervised learning and unsupervised learning, but it cannot get rid of structural assumptions.

Weak Supervision

Weak Supervision uses noisy, limited or imprecise sources but it can avoid structural assumptions by using subject matter experts (SMEs).

Development History

Snorkel v0.9 Teaser is developed by the Hazy Research Team from Stanford University in 2016. It is to become a modern Python library for building and managing training datasets. The developer highlights three improvements: (1) adding transformation functions and slicing functions; (2) upgrading labeling pipeline; (3) redesigning the codebase.

The URL of the new official website is: https://www.snorkel.org. The former one is moving into the new address, which is an indirect evidence that the model is becoming stable.

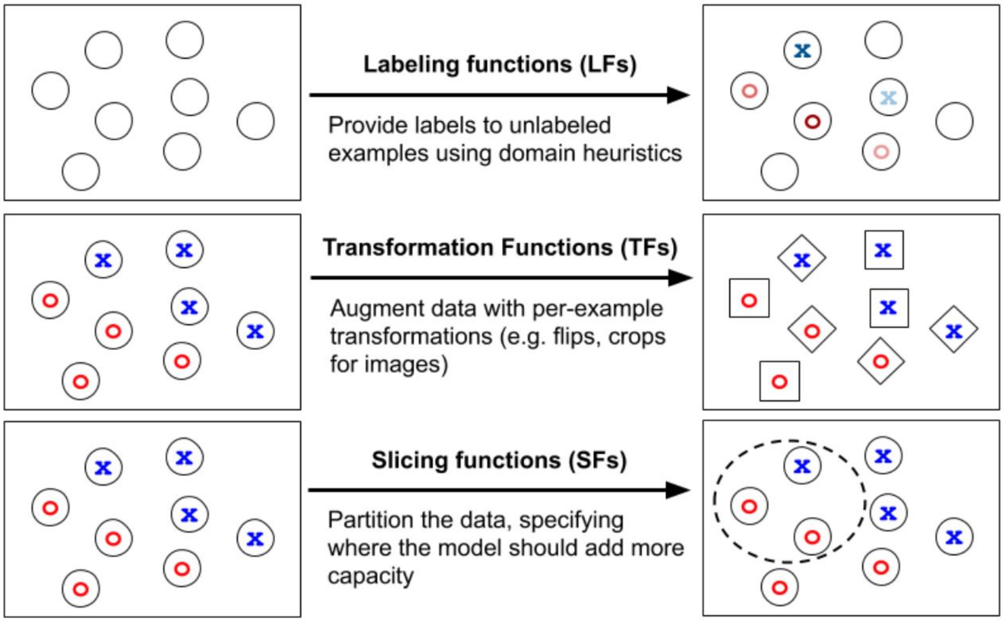
Key Functions

Functions	**Techniques**

Labeling Function (LF)	Labeling Training Data
Transformation Functions (TF)	Data Augmentation
Slicing Function (SF)	Monitoring Critical Data Subsets

When to use Snorkel?

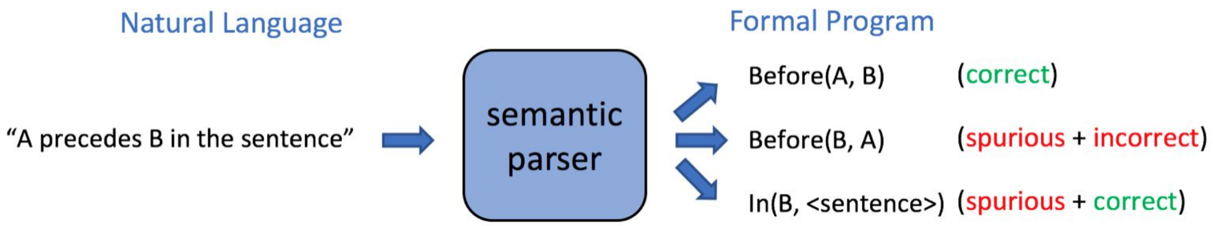
Getting Start: applying core data operations


Labeling functions (LFs)
Provide labels to unlabeled examples using domain heuristics
Transformation Functions (TFs)
Augment data with per-example transformations (e.g. flips, crops for images)
Slicing functions (SFs)
Partition the data, specifying where the model should add more capacity

I followed the steps in the spam tutorial to apply these three functions.

Text

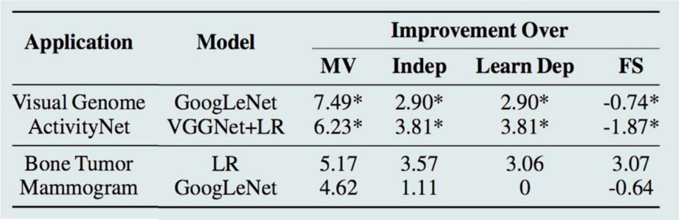
The team introduces a babble labble option with explanations. It uses a semantic parser to convert natural language explanations into “accurate enough” labelling functions. This is an implementation of subject matter experts(SMEs).



For example, in the sentence ‘Barack and Michelle visited Stanford University with their daughter for a college visit.’, we need to process the component, ‘with their daughter’.

Image

Snorkel is more helpful in analyzing the image relation, it is more likely to be solved automatically. The result in identifying and modeling correlations among heuristics is more accomplished than the fully supervised learning.



Application	Model	Improvement Over			
		MV	Indep	Learn Dep	FS
Visual Genome	GoogLeNet	7.49*	2.90*	2.90*	-0.74*
ActivityNet	VGGNet+LR	6.23*	3.81*	3.81*	-1.87*
Bone Tumor	LR	5.17	3.57	3.06	3.07
Mammogram	GoogLeNet	4.62	1.11	0	-0.64

MV: *Majority Vote across heuristic functions;* **Indep**: *not modeling dependencies;*
Learn Dep: *learning, not inferring dependencies;* **FS**: *fully supervised model*

The figure comes from the result in Visual Genome dataset.

Conclusions

In the future, we can use Snorkel in multi-task learning and superGlue.

The pros and cons

Comparison

References

[1] Marsland, S. (2014). *Machine learning: an algorithmic perspective*. Chapman and Hall/CRC.
[2] Ratne, Alex., Bach, Stephen., Varma, Paroma., Ré, Chris (2017-07-16). An Overview of Weak Supervision [Web blog post]. Retrieved from https://www.snorkel.org/blog/weak-supervisionL.
[3] Mastering Machine Learning: A Step-by-Step Guide with MATLAB
Retrieved from https://www.mathworks.com/discovery/unsupervised-learning.html.
[4] Weak Supervision(2019-11-18)[Web blog post]. Retrieved from https://en.wikipedia.org/wiki/Weak_supervision
[5]snorkel-tutorials/getting_started/getting_started.ipynb.

Retrieved from https://github.com/snorkel-team/snorkel-tutorials/blob/master/getting_started/getting_started.ipynb

[6] Hancock, Braden., Liang, Percy., Ré, Chris (2018-05-15). Training with Natural Language [Web blog post]. Retrieved from https://www.snorkel.org/blog/babble.

[7] Varma, Paroma., He, Bryan., Ré, Chris (2017-09-14). Snorkel for Image Data [Web blog post]. Retrieved from https://www.snorkel.org/blog/coral.