Semi-supervised Learning and Active Learning

Soumyadeep Roy

CNeRG Retreat 2018

Talk overview

- SSL - Basic concepts and terminology
- Active Learning
- SSL - Needs before DL era / non-DL settings
- SSL - Taxonomy post-DL era / DL setting
- Recent papers (4)
- Works from CNeRG (2)
- Conclusion

What is Semi-Supervised Learning?

Learning from both labeled and unlabeled data. Examples:

- Semi-supervised classification: training data l labeled instances $\{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$ and u unlabeled instances $\{\mathbf{x}_j\}_{j=l+1}^{l+u}$, often $u \gg l$. **Goal**: better classifier f than from labeled data alone.
- Constrained clustering: unlabeled instances $\{x_i\}_{j=1}^{n}$, and "supervised information", e.g., must-links, cannot-links. **Goal**: better clustering than from unlabeled data alone.

Motivations

Machine learning

Promise: better performance for free...

- labeled data can be hard to get
  - labels may require human experts
  - labels may require special devices
- unlabeled data is often cheap in large quantity

Cognitive science

Computational model of how humans learn from labeled and unlabeled data.

- concept learning in children: $\mathbf{x}$=animal, y=concept (e.g., dog)
- Daddy points to a brown animal and says "dog!"
- Children also observe animals by themselves

Semi-supervised vs. transductive learning

- Inductive semi-supervised learning: Given $\{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$, $\{\mathbf{x}_j\}_{j=l+1}^{l+u}$, learn $f : \mathcal{X} \mapsto \mathcal{Y}$ so that f is expected to be a good predictor on future data, beyond $\{\mathbf{x}_j\}_{j=l+1}^{l+u}$.
- Transductive learning: Given $\{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$, $\{\mathbf{x}_j\}_{j=l+1}^{l+u}$, learn $f : \mathcal{X}^{l+u} \mapsto \mathcal{Y}^{l+u}$ so that f is expected to be a good predictor on the unlabeled data $\{\mathbf{x}_j\}_{j=l+1}^{l+u}$. Note f is defined only on the given training sample, and is not required to make predictions outside them.

Active Learning

- *The key idea behind active learning is that a machine learning algorithm can achieve greater accuracy with fewer training labels if it is allowed to choose the data from which it learns. An active learner may pose queries, usually in the form of unlabeled data instances to be labeled by an oracle (e.g., a human annotator).*

[Settles, 2012]

- *"what is the optimal way to choose data points to label such that the highest accuracy can be obtained given a fixed labeling budget."*

[Sener & Savarese, 2018]

Milestone papers/ideas - Prior DL era

- Transductive Support machine (Vapnic 1995)
  - Introduced TSVMs which helps improve the generalization accuracy of SVMs
  - *TSVMs, like SVMs, learn a large margin hyperplane classifier using labeled training data, but simultaneously force this hyperplane to be far away from the unlabeled data.*
- Extended (Joachims 1999)
  - *Proposes a combinatorial approach, known as SVMLight-TSVM, that is practical for a few thousand examples*
- Large-scale Transductive SVM (JMLR, Joachims 2006)
  - *Introduces a large scale training method for TSVMs using the concave-convex procedure*
- Transductive Learning via Spectral Graph Clustering (Joachims 2003)
- Online learning on graphs (Pontill 2005)
  - Learning a function defined on a graph from a set of labeled vertices
  - Discusses how to extend in order to allow active learning on a graph

Motivation changed now : Post-DL era

- Training a deep learning model requires a large amount of labeled examples
  - Not readily available in most cases
- We have a limited budget
  - **Batch-mode Active Learning** - optimal way to select data points to label
  - **Transfer Learning** - Pre-training the weights by a classifier of a related task and thus require much less amount of labeled data
- Settings where labels are difficult to obtain
  - **Autonomous learning system** - operates under no supervision
  - **Crowdsourcing** - Multiple workers provide answers to questions for which the correct answer is unknown

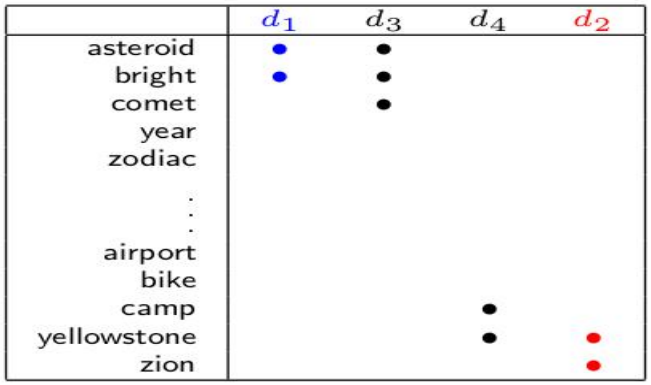
Taxonomy - Post-DL era

- Self-labeled methods : Automated annotation of unlabeled points
  - Self-training, cotraining, Label propagation
- Domain transfer : Pretraining the weights based on a similar task and training on our task with much less amount of data
  - Transfer Learning
- Semi-supervised network generation model
- Performance evaluation purpose : Using unlabeled data to estimate accuracy, when true labels/ground truth not available
- Recommender Systems : Using rich context or knowledge graphs

Interesting problem areas

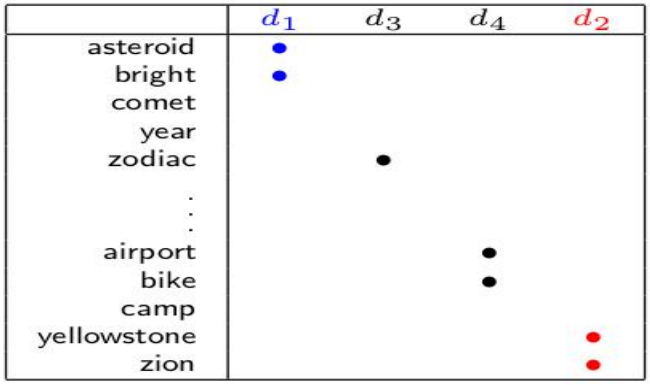
Example: text classification

- Classify astronomy vs. travel articles
- Similarity measured by content word overlap



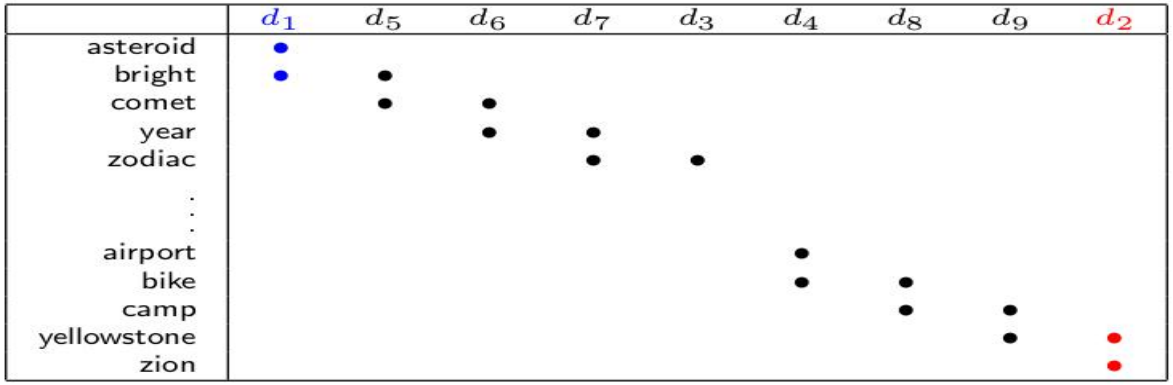
When labeled data alone fails

No overlapping words!

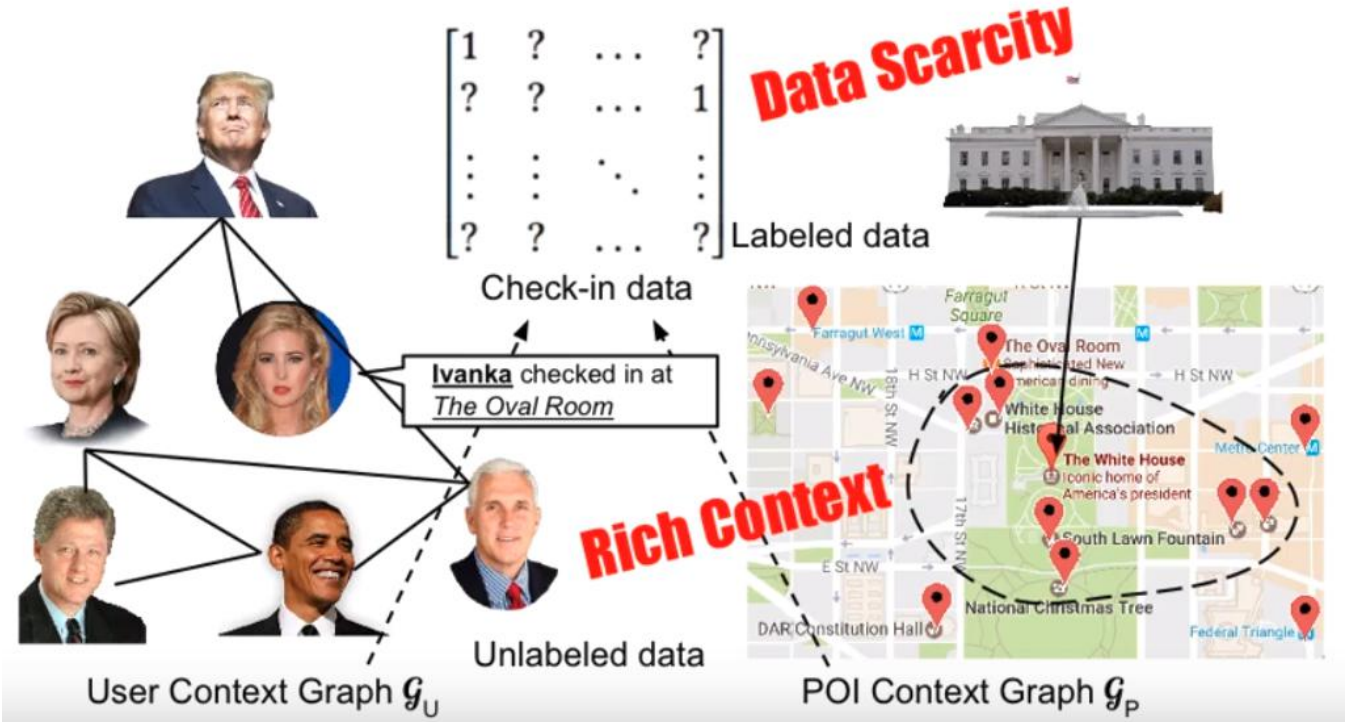


Unlabeled data as stepping stones

Labels "propagate" via similar unlabeled articles.



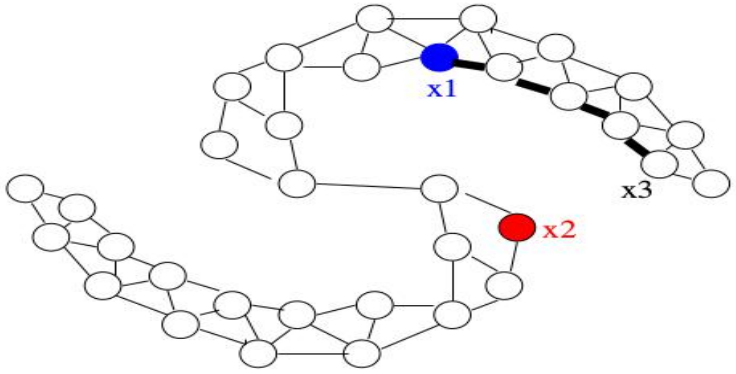
Semi-supervised techniques - A solution



Yang, C., Bai, L., Zhang, C., Yuan, Q., & Han, J. (2017, August). Bridging collaborative filtering and semi-supervised learning: a neural approach for poi recommendation. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1245-1254). ACM.

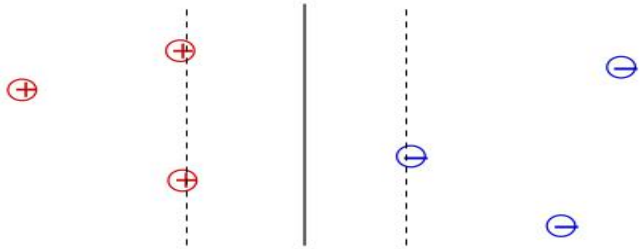
Graph-based semi-supervised learning

- Nodes: $X_l \cup X_u$
- Edges: similarity weights computed from features, e.g.,
  - k-nearest-neighbor graph, unweighted (0, 1 weights)
  - fully connected graph, weight decays with distance $w = \exp\left(-\|x_i - x_j\|^2/\sigma^2\right)$
  - ϵ-radius graph
- Assumption Instances connected by heavy edge tend to have the same label.

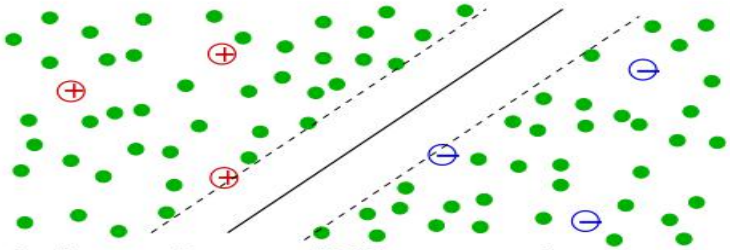


Semi-supervised Support Vector Machines

SVMs

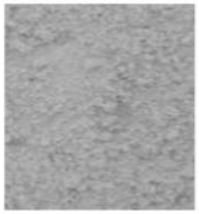
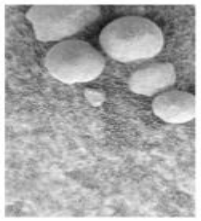


Semi-supervised SVMs (S3VMs) = Transductive SVMs (TSVMs)



Assumption: Unlabeled data from different classes are separated with large margin.

Life-long learning

x_1	x_2	...	x_{1000}	...	$x_{1000000}$	...
		...		...		...
$y_1 = 0$	-	-	$y_{1000} = 1$	...	$y_{1000000} = 0$	...

- $n \longrightarrow \infty$ examples arrive sequentially, cannot store them all
- most examples unlabeled
- no iid assumption, $p(x, y)$ can change over time

This is how children learn, too

x_1	x_2	$\cdots$	x_{1000}	$\cdots$	$x_{1000000}$	$\cdots$
		$\cdots$		$\cdots$		$\cdots$
$y_1 = 0$	-	-	$y_{1000} = 1$	$\cdots$	$y_{1000000} = 0$	$\cdots$

New paradigm: online semi-supervised learning

1. At time t, adversary picks $x_t \in \mathcal{X}, y_t \in \mathcal{Y}$ not necessarily iid, shows x_t
2. Learner has classifier $f_t : \mathcal{X} \mapsto \mathbb{R}$, predicts $f_t(x_t)$
3. With small probability, adversary reveals y_t; otherwise it abstains (unlabeled)
4. Learner updates to f_{t+1} based on x_t and y_t (if given). Repeat.

List of papers

- Active Learning for Convolutional Neural Networks, ICLR 2018
  - Select the optimal set of unlabeled data to annotate within a limited budget
- Cost-effective training of deep CNNs with active model adaptation, KDD 2018
  - Perform pre-training using a similar task
- When does label propagation fail? a view from a network generative model, IJCAI 2017
  - Proposes a semi-supervised network generation model based on Label propagation
- Estimating Accuracy from Unlabelled Data, NIPS 2017
  - Place constraints on target classes in multiple classification settings

Paper 1 - AL approaches for CNN

- For CNNs, traditional AL fails, so we resort to batch-mode AL
  - Since few examples for training, difficult to obtain proper feature representation
  - AL depends on assumption that feature representation remains the same, which is not true for CNNs

O. Sener, S. Sarvarese, “Active Learning for Convolutional neural Networks: A Core-Set Approach, ICLR 2018

Paper 1 - AL approaches for CNN

- For CNNs, traditional AL fails, so we resort to batch-mode AL
  - Since few examples for training, difficult to obtain proper feature representation
  - AL depends on assumption that feature representation remains the same, which is not true for CNNs

- Previous intuitions fail for batch-mode AL is due to correlation among batch samples
  - Proposes a way to select a batch in the form of a core set (selecting a subset)
  - Recent methods trade-off uncertainty with diversity

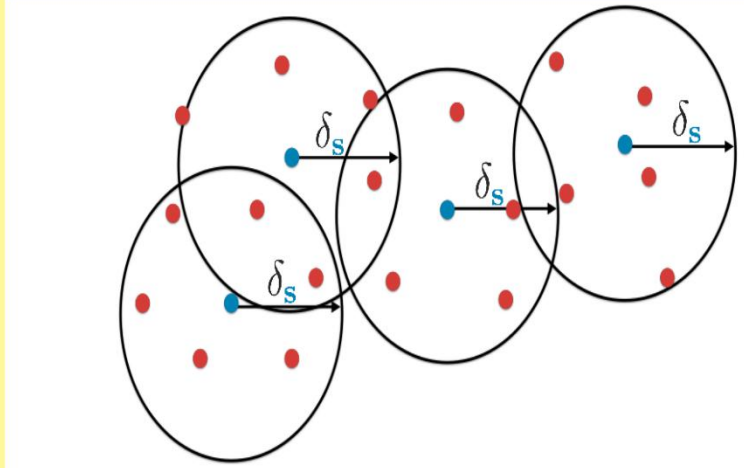


Figure 1: **Visualization of the Theorem 1** Consider the set of selected points s and the points in the remainder of the dataset $[n] \setminus s$, our results shows that if s is the δ_s cover of the dataset, $\left|\frac{1}{n}\sum_{i\in[n]} l(\mathbf{x}_i, y_i, A_s) - \frac{1}{|s|}\sum_{j\in s} l(\mathbf{x}_j, y_j; A_s)\right| \leq \mathcal{O}(\delta_s) + \mathcal{O}\left(\sqrt{\frac{1}{n}}\right)$

O. Sener, S. Sarvarese, “Active Learning for Convolutional neural Networks: A Core-Set Approach, ICLR 2018

Paper 1 - AL approaches for CNN

- A weakly supervised deep learning scheme
  - Comparable methods : Ladder networks, adversarial techniques, k-Center-Greedy algorithm
- Evaluated against baselines
  - Random, Best Empirical Uncertainty, Bayesian AL, Best Oracle uncertainty, Cost-effective AL
  - Image datasets - CIFAR-100 and Caltech-256

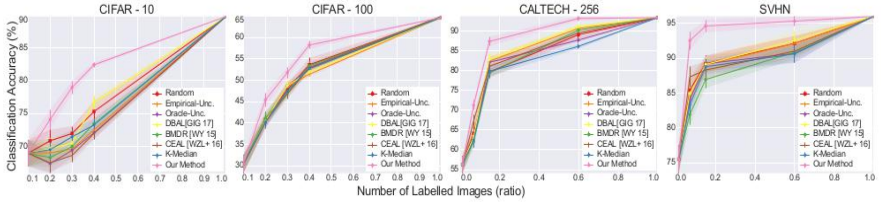


Figure 3: Results on Active Learning for Weakly-Supervised Model (error bars are std-dev)

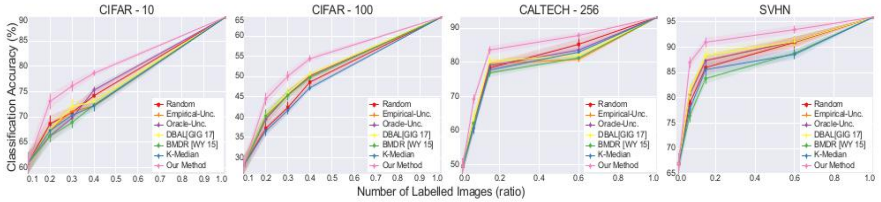
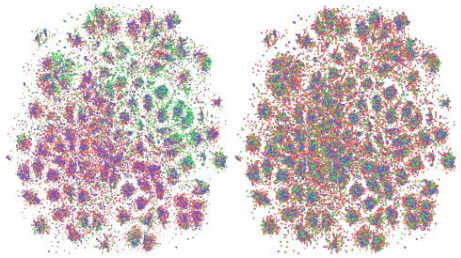


Figure 4: Results on Active Learning for Fully-Supervised Model (error bars are std-dev)



(a) Uncertainty Oracle (b) Our Method

tSNE embeddigns of the CIFAR dataset andnd behavior of uncertainty oracle and our method

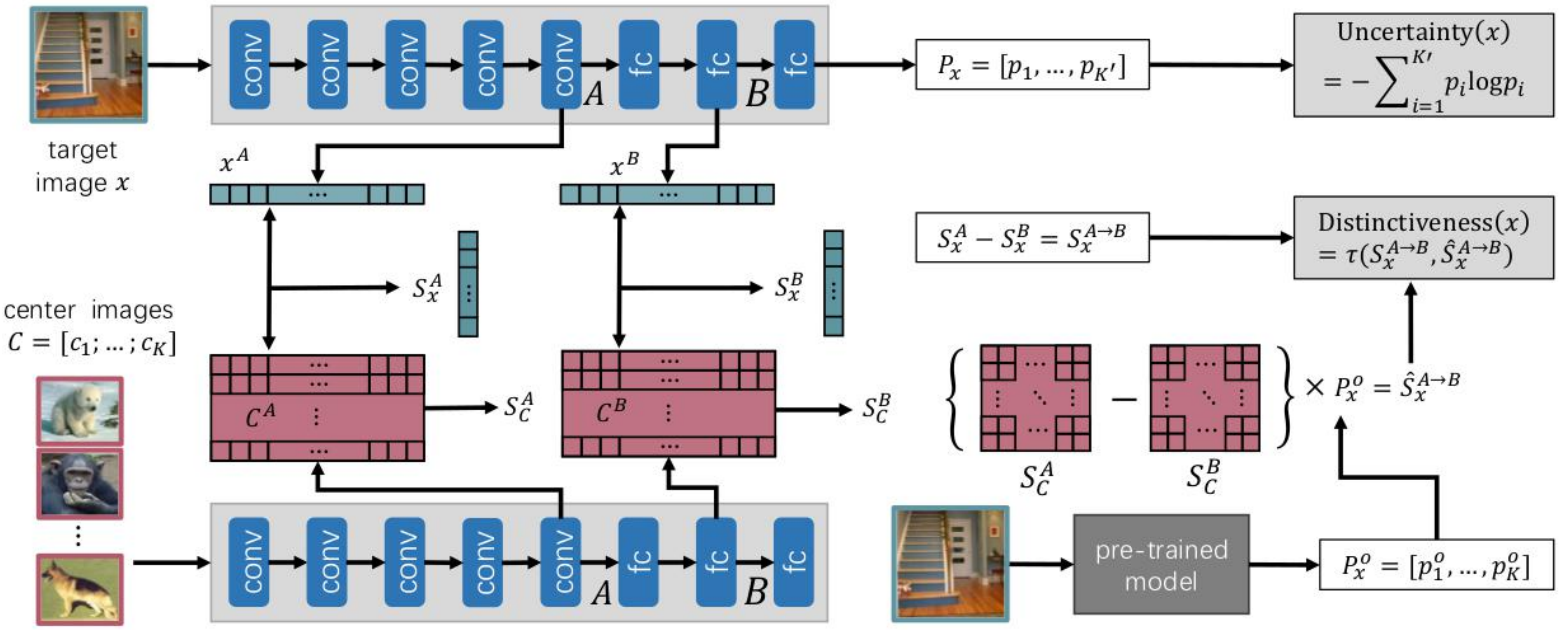
O. Sener, S. Sarvarese, “Active Learning for Convolutional neural Networks: A Core-Set Approach, ICLR 2018

2. Cost-effective training of Deep CNNs actively

- Adapting a pretrained model to a new task
- Proposes a general framework of active model adaptation for deep CNNs
- Actively querying data points to label
  - Proposes a novel criterion for selection which best optimizes the feature representation along with the classifier performance
  - Proposes an algorithm that can actively sellect instances to achieve better feature representation and label prediction.

Huang et. al, "Cost-effective Training of Deep CNNs with Active Model Adaptation", KDD 2018

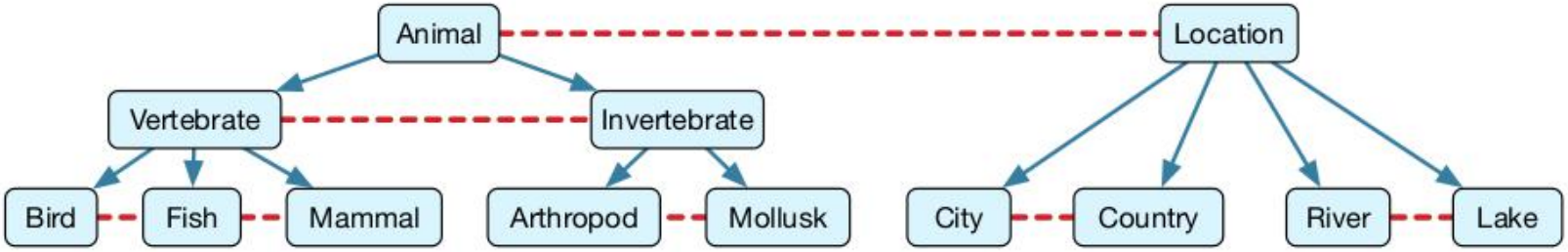
2. Cost-effective training of Deep CNNs actively



Huang et. al, "Cost-effective Training of Deep CNNs with Active Model Adaptation", KDD 2018

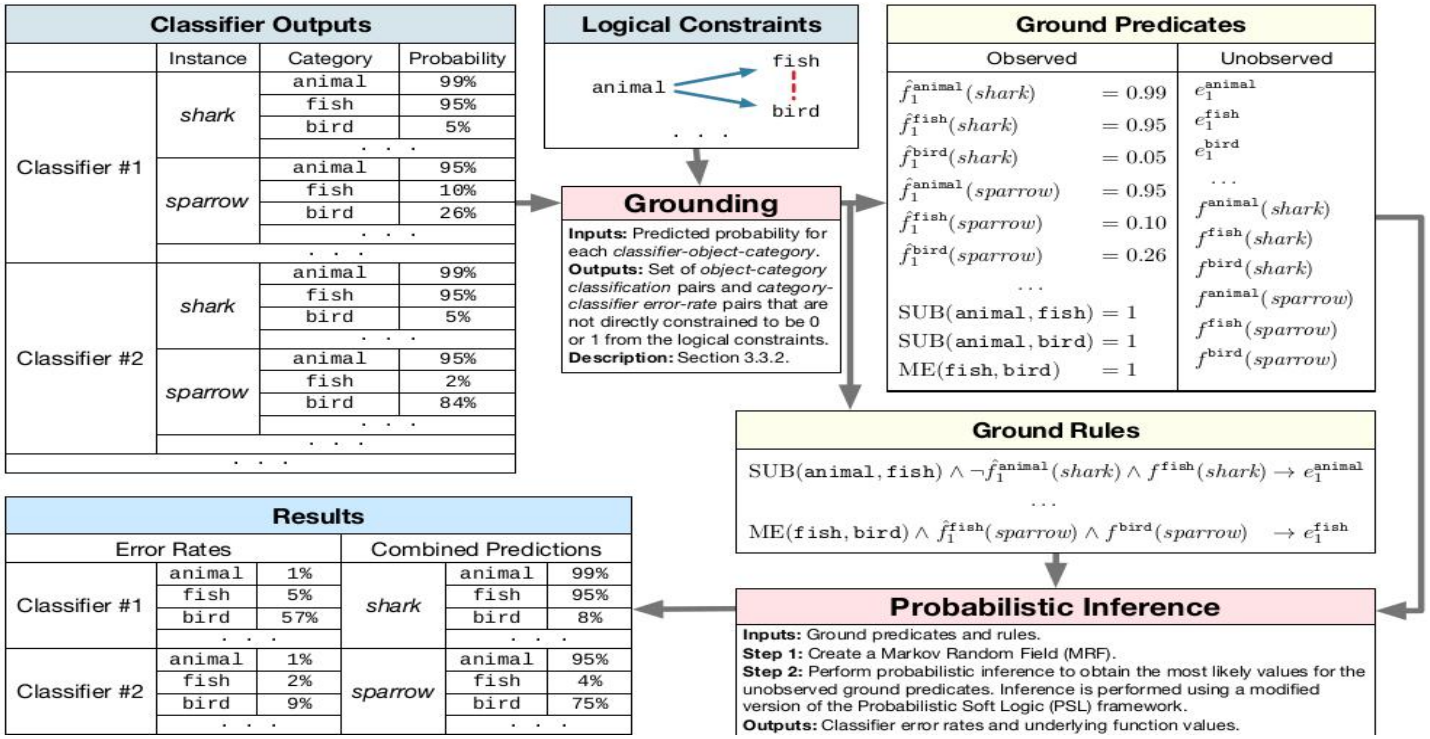
3. Estimating Accuracy from Unlabeled Data

- A multiple classification problem where target classes are tied together in logical constraints
  - **Intuition 1** : When classifiers agree, there are more likely to be correct
  - **Intuition 2** : When the classifiers make a prediction that violates the constraint, then at least one of the classifiers is making an error
- Constraints - Mutual exclusion rule, Subsumption rule



Platanios et. al., "Estimating Accuracy from Unlabeled Data : A Probabilistic Logic Approach", NIPS 2017

3. Estimating Accuracy from Unlabeled Data



Platanios et. al., "Estimating Accuracy from Unlabeled Data : A Probabilistic Logic Approach", NIPS 2017

When Does Label Propagation Fail? A View from a Network Generative Model

Yuto Yamaguchi and Kohei Hayashi

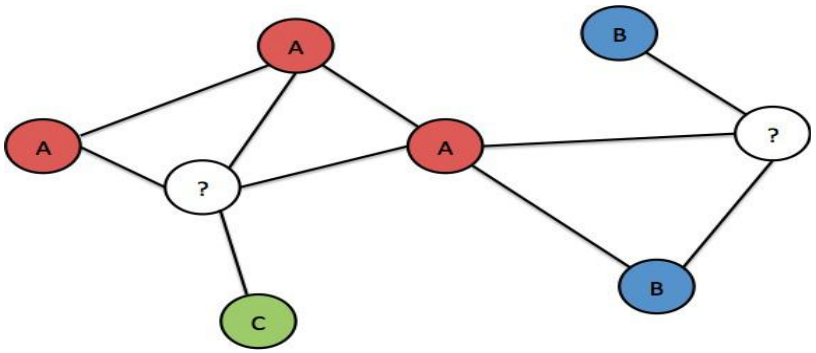
AIST

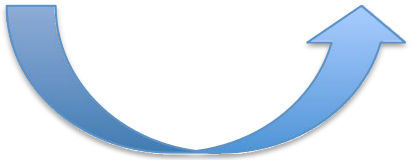
IJCAI@Melbourne

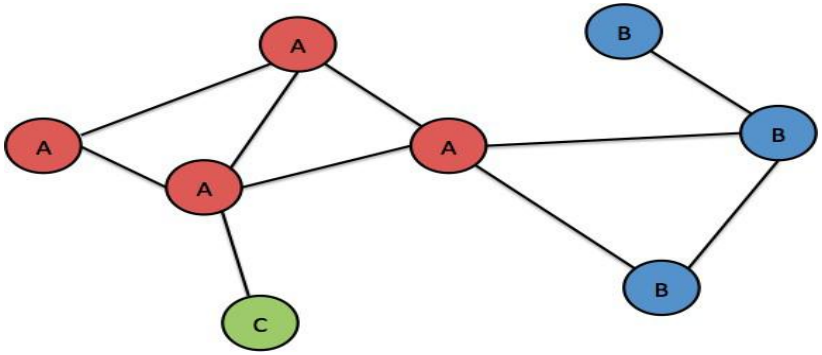
Node Classification

Given Partially labeled undirected graph

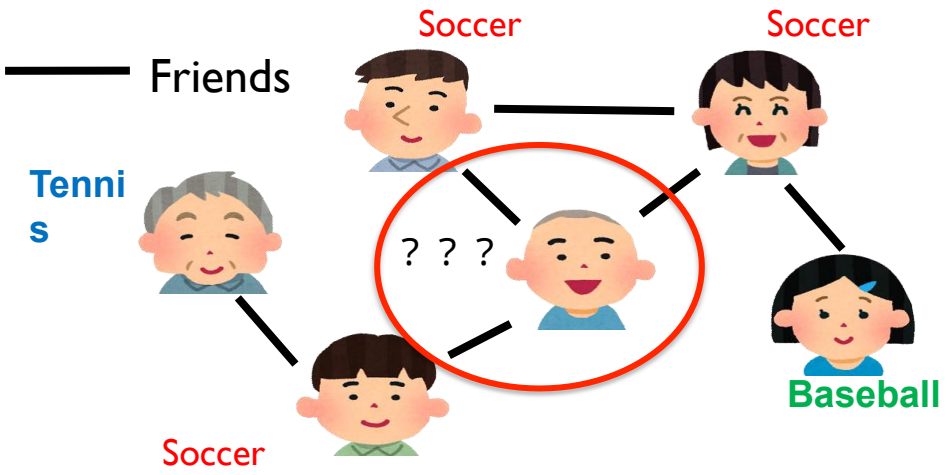
Find Labels of all nodes







Example: User profile inference



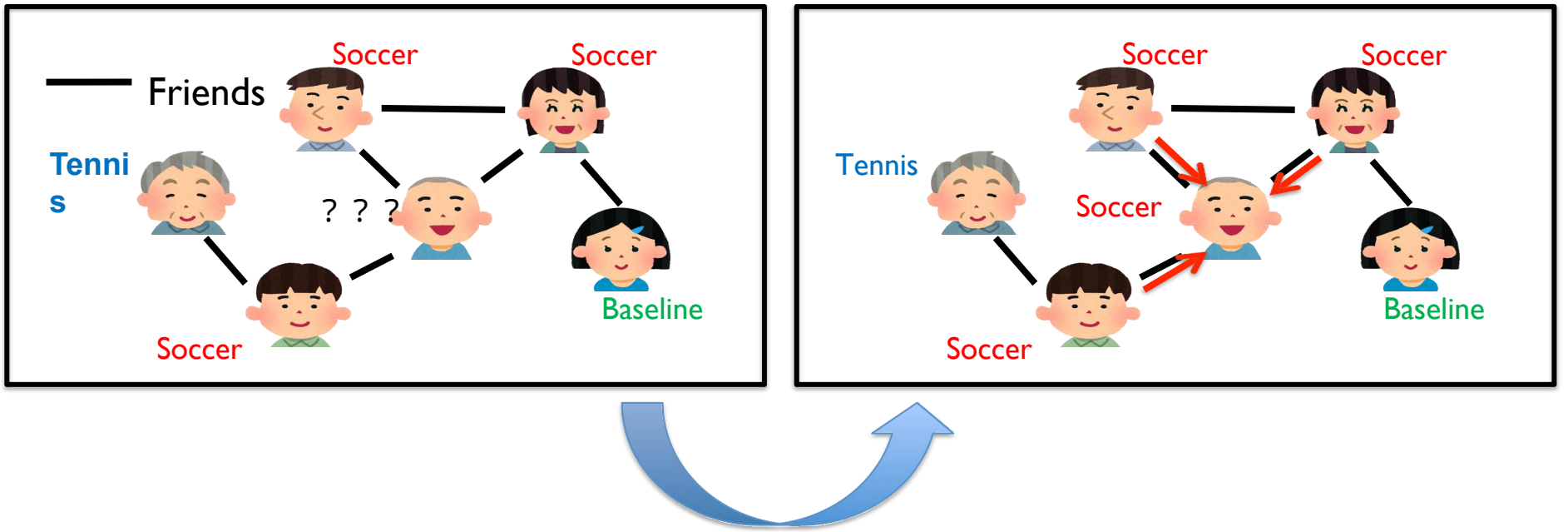
What's his hobby?

→ Node Classification

Label Propagation(LP) (1/2)

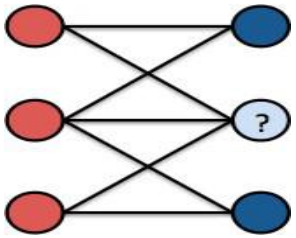
[Zhu+, 03], [Zhou+, 03], etc.

Propagate neighbors' labels

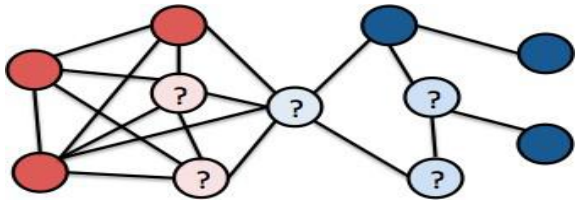


Cases when LP fails (practically known)

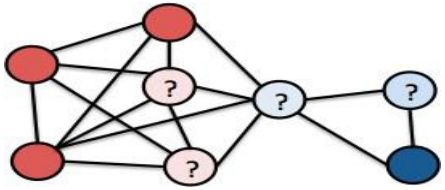
Different labels are connected



Edge probability is not uniform uniform



Label ratio is not



Q. So, do we know why LP fails in these cases?

A. No. Since it's not a probabilistic model, we don't know the assumptions behind the model.

What we do in this work

- Prove a theoretical relationship between LP and Stochastic Block Model, which is a well-studied probabilistic generative model
- Show when and why LP fails. Answers the following research questions regarding suitability for LP :
  - Type of node labels (assortative/disassortative)
  - Should the density of intra-cluster edges be uniform or non-uniform
  - Should the label distribution be uniform or non-uniform
  - Can the observed labels be allowed to be correct or incorrect.

Open questions and observations

- SSL is a suite of techniques that uses unlabeled data in some form in their learning models
  - No organised or compact study of literature is observed
- SSL is a too broad a definition
  - Ranges across classification, clustering and recommender systems
- There have been changes in its usage from traditional SSL to adapt to either :
  - Deep Learning setting - Which requires a lot of training examples
  - Other settings - Labeled data is hard / impossible to obtain