AALBORG UNIVERSITET

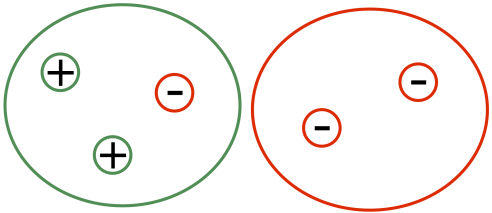
DEIS
DISTRIBUTED, EMBEDDED AND INTELLIGENT SYSTEMS

Learning and interpreting multi-multi-instance learning networks

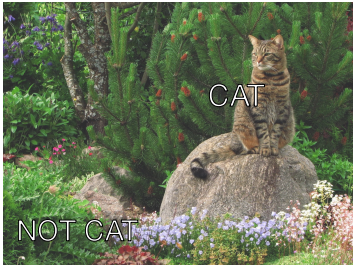
Alessandro Tibo [1] Manfred Jaeger [1] Paolo Frasconi [2]
[1]Department of Computer Science, University of Aalborg
[2]Department of Information Engineering, University of Florence

Contributions

- Multi-multi instance learning (MMIL) framework
- Interpreting MMIL models
- MMIL for Graphs



- Data consists of labeled bags of instances
- Instance labels are not observed
- A bag is positive iff it contains at least a positive instance



- Images can be decomposed into bags of regions (instances)
- Bags are labeled with either “positive” or “negative”
- An image is “positive” if it contains at least one region representing a “cat”

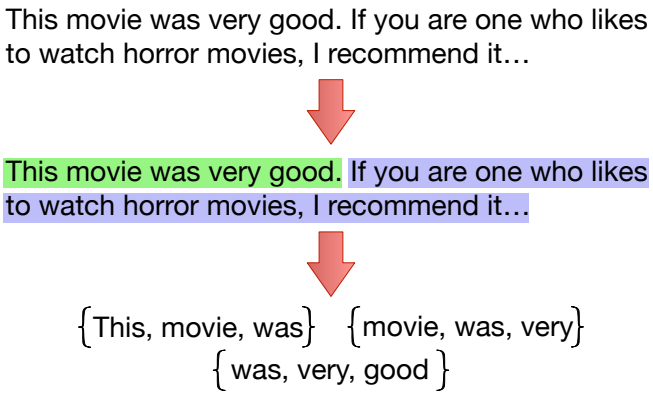
Other multi-instance applications include:

- Text categorization
- Diagnostic medical imaging

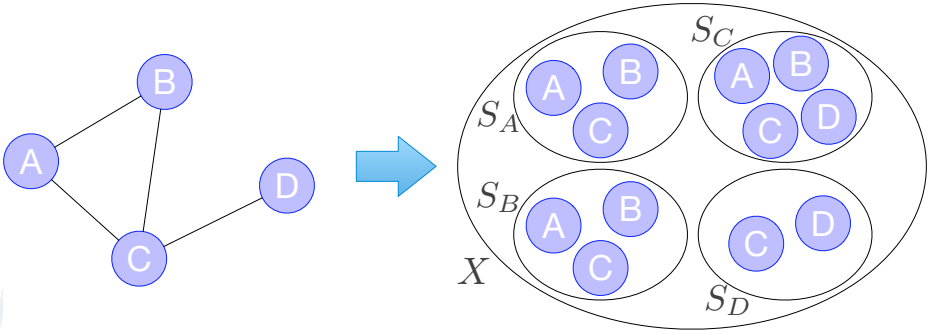
Common techiniques:

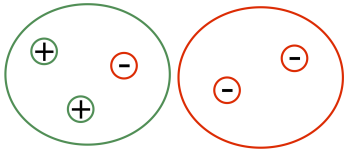
- Axis-parallel Hyper-rectangle (Dietterich et al., 1997)
- Diverse Density (Maron and Lozano-Peiérez, 1998)
- MI-SVM and mi-SVM (Andrews et al., 2003)
- Multi-instance Neural Networks (Ramon and De Raedt, 2000)

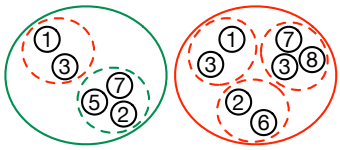
A document could be represented as a bag of sentences, which in turn are bags of trigrams



A graph could be decomposed in bags of node neighborhoods, where each node is an instance



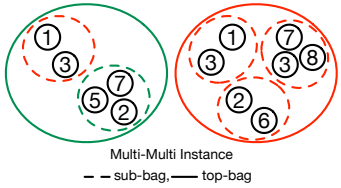




- Binary labels are attached to bags and instances
- A bag is positive iff it contains at least a positive instance

- Categorical labels are attached to instances and bags
- Many nested levels of aggregation
- ▪▪ sub-bag, ▬ top-bag

Supervised setting: only top-bag labels are observed, while sub-bag and instance labels are latent

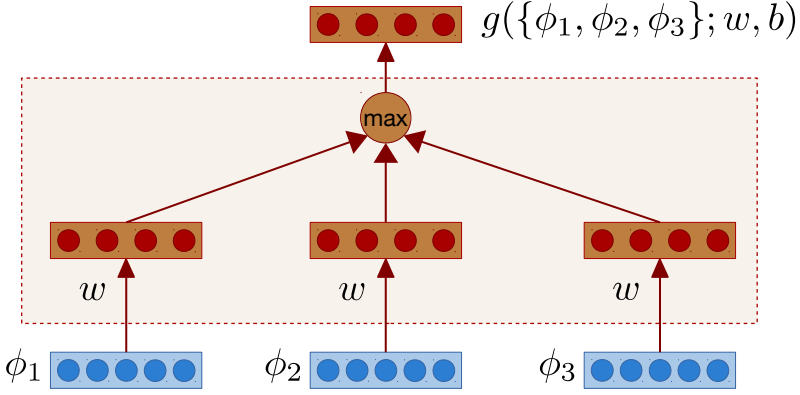


- instance labels are drawn from a distribution depending on instances
- sub-bag and top-bag labels are drawn from a distribution depending on instance and sub-bag labels
- $p(y|S^l) = p(y|y_1, \ldots, y_{|S|}),\ S^l = \{(x_1, y_1), \ldots, (x_{|S|}, y_{|S|})\}$, where S^l is a bag of labelled instances and y is the bag label

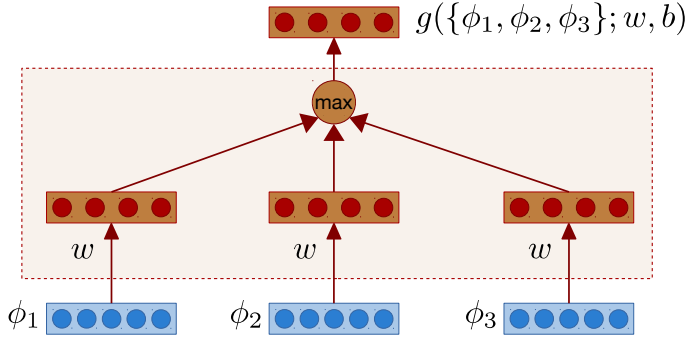
How can we learn from MMIL data using a neural network architecture?

How can we learn from MMIL data using a neural network architecture?

Using a special neural network layer called Bag-Layer



$$g(\{\phi_1, \ldots, \phi_n\}; w, b) = \max_{i=1}^{n} \alpha\left(w\phi_i + b\right)$$

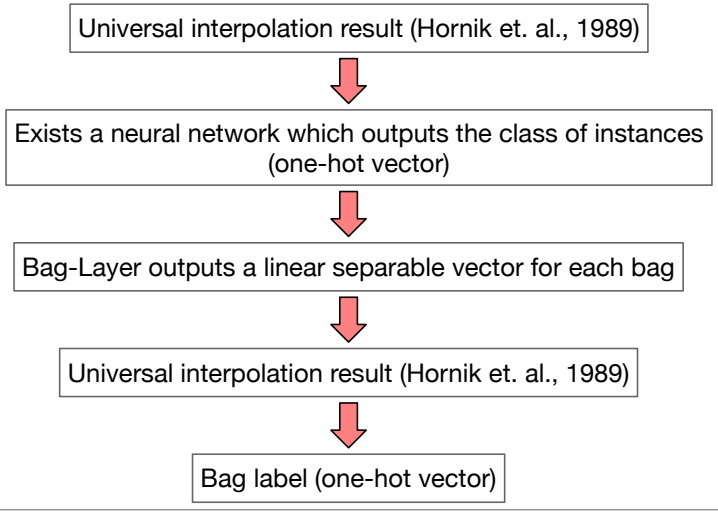


- bag-layers can be intermixed with standard neural network layers
- bag-layers aggregate internal representations of instances and bags

How expressive our model is?

- **Assumptions (non-counting restricted MMIL setting)**: labels are deterministically assigned and no form of counting is involved

- **Theorem**: Given a dataset of examples generated under the non-counting restricted MMIL setting, there exist a network with two bag-layers that can correctly label all examples in the dataset


Universal interpolation result (Hornik et. al., 1989)
Exists a neural network which outputs the class of instances (one-hot vector)
Bag-Layer outputs a linear separable vector for each bag
Universal interpolation result (Hornik et. al., 1989)
Bag label (one-hot vector)

Neural networks are in general black boxes and it is hard to understand the reason of classification decisions

Local inteprability techniques for Neural Networks:

- Layer-wise Relevance Propagation (Lapuschkin et al., 2016; Samek et al., 2016)
- Local Interpretable Model-agnostic Explanations (Ribeiro et al., 2016)

The MMIL framework naturally leads to a particular form of interpretability by associating pseudo-labels to both sub-bags and instances and construct interpretable predictors

Pseudo-labels as surrogates for hypothesized actual latent labels (attached with sub-bags and instances)

Proposed interpretability framework for MMIL

1. Clustering and pseudo-label construction
2. Interpreting pseudo-labels
3. Learning interpretable rules
4. Explaining individual classifications

We explain both classifications and classification rules

Graph Classification

- Graph Neural Networks (Scarselli et al., 2009)
- Deep Graph Kernels (Yanardag and Vishwanathan, 2015)
- Patchy-SAN (Niepert et al., 2016)

Node Classification

- Graph Convolutional Network (Kipf and Welling, 2016)
- GraphSAGE (Hamilton et al., 2017)

All these methods use the message passaging strategy

The MMIL perspective can also be used to derive algorithms suitable for supervised learning over graphs by decomposing them into a MMIL data.

- Graph Classification: a graph can be seen as a top-bag. Each sub-graph induced by a node and its neighborhood is a sub-bag, and each node within a sub-bag is an instance

- Node Classification (documents): each node and its neighboord is a top-bag. Each node is a sub-bag, and the words are the instances.

Experimental Results

IMDB movie review dataset

The IMDB dataset is a sentiment analysis dataset

- binary labels
- 25,000 balanced training examples, 25,000 balanced test examples, 50,000 unlabelled examples

Positive: ... Finally, i recommend to watch this movie... And i hope You'll love it enjoy :D
Negative: This film is just plain horrible. ...

IMDB movie review dataset

MMIL Dataset

- a top-bag is a review
- a sub-bag is a sentence
- an instance is a trigram

Network structure

- 2 stacked bag-layers with linear activation followed by ReLU
- output with Sigmoid activation

Test set accuracy: 92.26%

Sub-bag pseudo-labels (4) are learnt with KMeans, while rules are learnt with Decision Trees

v_1 - 11.37%	v_2 - 41.32%	v_3 - 15.80%	v_4 - 31.51%
overrated poorly written badly acted	I highly recommend you to NOT waste your time on this movie as I have	I loved this movie and I give it an 8/ 10	It's not a total waste
It is badly written badly directed badly scored badly filmed	This movie is poorly done but that is what makes it great	Overall I give this movie an 8/ 10	horrible god awful
This movie was poorly acted poorly filmed poorly written	Although most reviews say that it isn't that bad i think that if	final rating for These Girls is an 8/ 10	Awful awful awful

Instance pseudo-labels (5) are learnt with KMeans, while rules are learnt with Decision Trees

u_1 - 5.73%	u_2 - 8.68%	u_3 - 28.86%	u_4 - 2.82%	u_5 - 53.91%
PAD 8/ 10	trash 2 out	had read online	it's pretty poorly	give this a
an 8/ 10	to 2 out	had read user	save this poorly	like this a
for 8/ 10	PAD 2 out	on IMDb reading	for this poorly	film is 7
HBK 8/ 10	a 2 out	I've read innumerable	just so poorly	it an 11
Score 8/ 10	3/5 2 out	who read IMDb	is so poorly	the movie an
to 8/ 10	2002 2 out	to read IMDb	were so poorly	this movie an
verdict 8/ 10	garbage 2 out	had read the	was so poorly	40 somethings an
Obscura 8/ 10	Cavern 2 out	I've read the	movie amazingly poorly	of 5 8
Rating 8/ 10	Overall 2 out	movie read the	written poorly directed	gave it a
it 8/ 10	rating 2 out	Having read the	was poorly directed	give it a
fans 8/ 10	film 2 out	to read the	is very poorly	rating it a
Hero 8/ 10	it 2 out	I read the	It's very poorly	rated it a

v_1 corresponds to negative sentences, v_2 corresponds to descriptive, neutral or ambiguous sentences, v_3 corresponds to positive sentence, and v_4 corresponds to negative sentences. The pseudo-labels colored in grey are not used for building the rules.

1	$\hat{h} =$ positive $\leftarrow f_{v_1} \leq 4.04, f_{v_3} \leq 12.63, f_{v_4} \leq 39.17.$
2	$\hat{h} =$ positive $\leftarrow f_{v_1} \leq 12.97, f_{v_3} > 12.63.$
3	$\hat{h} =$ positive $\leftarrow f_{v_1} > 12.97, f_{v_3} > 25.66.$
4	$\hat{h} =$ negative $\leftarrow f_{v_1} \leq 4.04, f_{v_3} \leq 12.63, f_{v_4} > 39.17.$
5	$\hat{h} =$ negative $\leftarrow f_{v_1} > 4.04, f_{v_3} \leq 12.63.$
6	$\hat{h} =$ negative $\leftarrow f_{v_1} > 12.97, f_{v_3} \in (12.63, 25.66].$

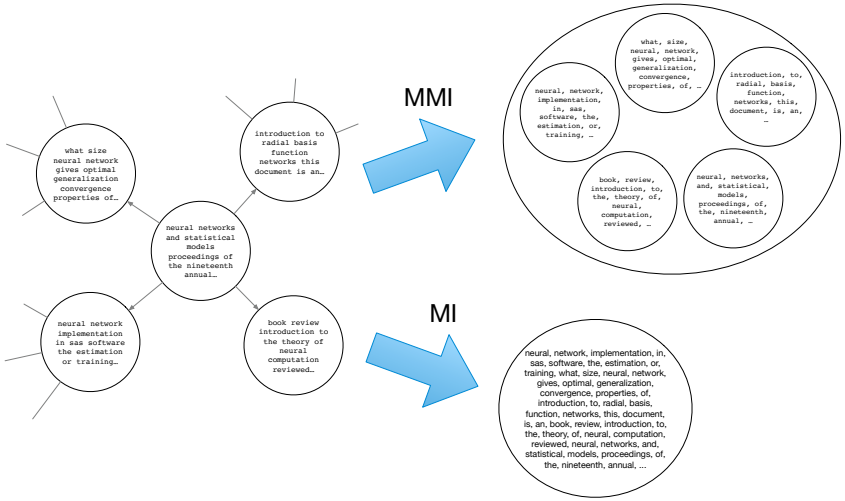
The fired rule for sentences is $f_{v_1} > 12.97, f_{v_3} > 25.66$

Story about three eclipse (maybe even Indigo, ha) children beginning their love for murder. Oh, and the people who are "hot" on their trail. [v_1] Bloody **Birthday, a pretty mediocre title**[4] for the **film, was a nice lil**[1] surprise. I was in no way expecting a film that dealt with blood-thirsty psychopath kids. [v_3] And I may say it's also **one of the best flicks**[1] I've seen with kids as the villains. By the end of the movie I seriously wanted these kids to die in horrible fashion. [v_3] **It's a really solid 80s**[1] horror flick, but how these kids are getting away with all this mayhem and murder is just something that **you can't not**[2] think about. Even the slightest bit of investigation would easily uncover these lil sh!ts as the murderers. But there seems to be only a couple police in town, well by the end, only one, and he seemed like a dimwit, so I suppose they could have gotten away with it. Haha, yeah, and I'm a Chinese jet-pilot. Nevertheless, this movie delivered some evilass kids who were more than entertaining, a lot of premarital sex and a decent amount of boobage. No kiddin! If you're put off by the less than stellar title, dash it from your mind and give this flick a shot. [v_3] It's **a very recommendable and underrated 80s**[1] horror flick.

We trained a MIL and MMIL networks on three common citation datasets: Citeseer, Cora, and PubMed

Dataset	# Classes	# Nodes	# Edges	# Training	# Validation	# Test
Citeseer	6	3,327	4,732	1,560 ($y \leq$ '99)	779 ('99 $< y \leq$ '00)	988 ($y >$ '00)
Cora	7	2,708	5,429	1,040 ($y \leq$ '94)	447 ('94 $< y \leq$ '95)	1,221 ($y >$ '95)
PubMed	3	19,717	44,338	8,289 ($y \leq$ '97)	3,087 ('97 $< y \leq$ '01)	8,341 ($y >$ '01)

Citation Network Datasets



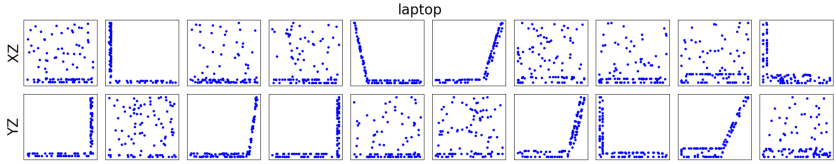
Citation Network Datasets

Model	Cora	Citeseer	PubMed
Naive Bayes (Bernoulli)	71.34%	63.77%	75.47%
Logistic Regression	74.94%	64.37%	73.67%
GCN	82.23%	66.50%	78.66%
GraphSage MeanPool	80.18%	66.19%	75.59%
GraphSage MaxPool	80.43%	67.61%	76.60%
MIL-Mean	79.93%	62.96%	81.15%
MIL-Max	81.08%	67.41%	80.22%
MMIL-Mean	82.80%	**70.75%**	**81.27%**
MMIL-Max	**84.03%**	69.64%	80.65%

We trained a MMIL model on 6 social network datasets. For each graph dataset a MMIL dataset was constructed.

Dataset	DGK	Patchy-SAN	SAEN	Our Method (S)	Our Method (MS)
COLLAB	73.09 ± 0.25	72.60 ± 2.15	78.50 ± 0.69	77.94 ± 0.62	$\mathbf{79.46 \pm 0.31}$
IMDB-BINARY	66.96 ± 0.56	71.00 ± 2.29	71.59 ± 1.20	71.99 ± 1.24	$\mathbf{72.62 \pm 1.04}$
IMDB-MULTI	44.55 ± 0.52	45.23 ± 2.84	48.53 ± 0.76	47.81 ± 0.63	$\mathbf{49.42 \pm 0.68}$
REDDIT-BINARY	78.04 ± 0.39	86.30 ± 1.58	$\mathbf{87.22 \pm 0.80}$	79.74 ± 0.48	86.54 ± 0.64
REDDIT-MULTI5K	41.27 ± 0.18	49.10 ± 0.70	$\mathbf{53.63 \pm 0.51}$	44.75 ± 0.33	53.42 ± 0.67
REDDIT-MULTI12K	32.22 ± 0.10	41.32 ± 0.42	$\mathbf{47.27 \pm 0.42}$	38.47 ± 0.57	45.25 ± 0.48

We consider a point cloud dataset, consisting of $\sim 10k$ training and $\sim 2.5k$ test objects distributed over 40 classes.



We subsequently created bags of bags by considering R equally distributed rotations, i.e. $\frac{2\pi i}{N}$, around the z-axis.

Test set accuracy: 88.10 $\pm$ 0.43% againts 85.35 $\pm$ 0.49% from Zaheer et al., 2017

- The MMIL framework handle data organized in nested bags
- Several problems can be expressed as MMIL problems
- Theoretical results show the expressivity of this type of model
- Natural interpretation of the behavior of the network

Questions?

alessandro@cs.aau.dk