Deep Multi-Instance Transfer Learning

Dimitrios Kotzias[1,2] Misha Denil[2] Phil Blunsom[2,3] Nando de Freitas[2,3,4]

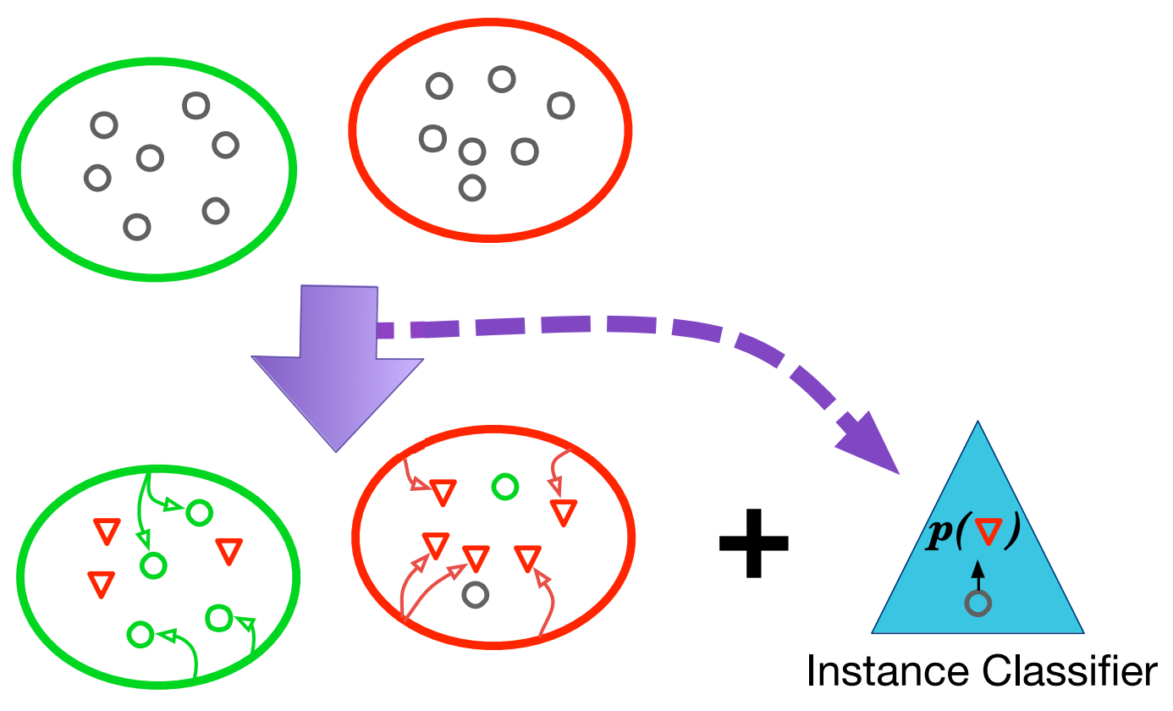
[1]University of California, Irvine [2]University of Oxford [3]Google Deep Mind [4]CIFAR

dkotzias@ics.uci.edu {misha.denil,phil.blunsom,nando}@cs.ox.ac.uk

Motivation

We often have labels for *groups of instances*, but not each individual one. We present a method to estimate the labels of the individual instances.

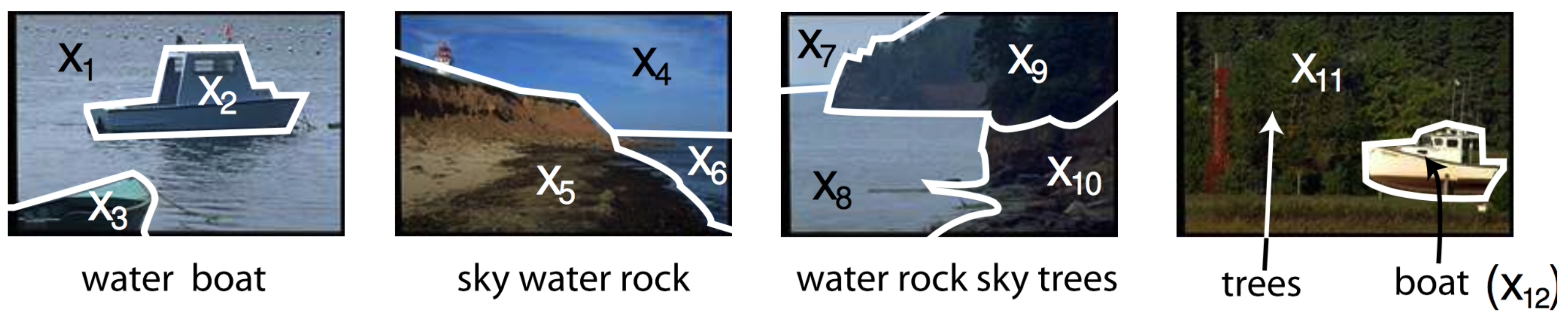
- Propose a loss function to simultaneously transfer knowledge to instances; and learn the parameters of an instance classifier
- We demonstrate this idea by inferring the ratings of sentences (individuals) from ratings of reviews (groups).
- This reduces the need for precise labelling in sentiment analysis.



Multi-Instance Learning Applications

Many problems can be considered in a multi-instance learning setting

- Image-caption correspondence
- Translation alignment
- Text analysis
- Privacy breaking



The objective is to discover which image segment matches the label boat (Kueck et. al 2004)

	Image 1	Image 2	Image 3	Image 4
Input $\mathbf{x}$	$\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_3$	$\mathbf{x}_4$ $\mathbf{x}_5$ $\mathbf{x}_6$	$\mathbf{x}_7$ $\mathbf{x}_8$ $\mathbf{x}_9$ $\mathbf{x}_{10}$	$\mathbf{x}_{11}$ $\mathbf{x}_{12}$
Labels $\mathbf{y}$ (Kueck '04)	? ? ?	0 0 0	0 0 0 0	0 1
MIL (group labels)	1	0	0	1
Output	$\boldsymbol{\theta}, p(\text{“}boat\text{”}\,\vert\mathbf{x} =, \boldsymbol{\theta})\forall\mathbf{x}$			

Deep NLP for Feature Learning

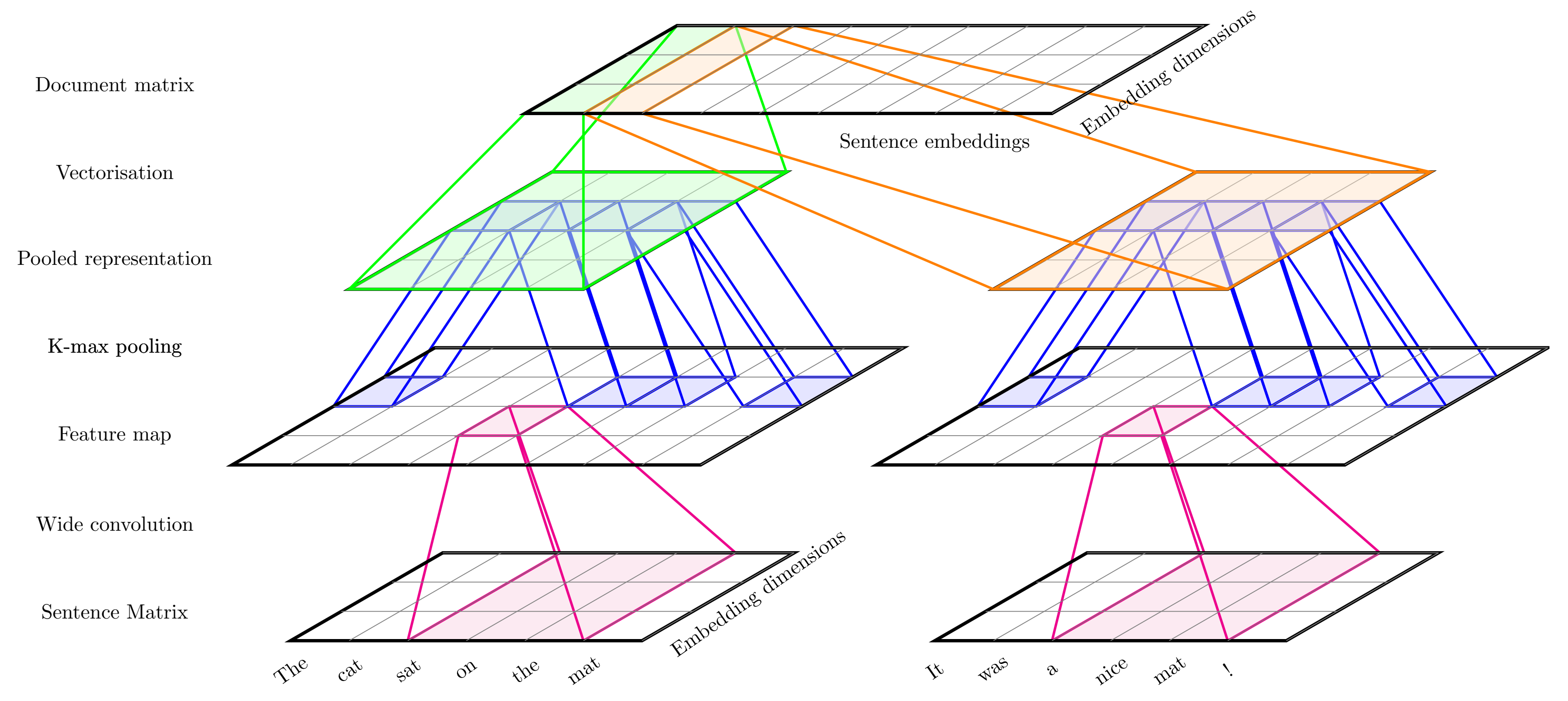
Distributed Representations of Words (Hinton et al., 1986, Bengio et al., 2003, Mikolov et al., 2013)

$$word \rightarrow \mathbf{x} \in \mathcal{R}^n$$

Larger blocks of text, Paragraphs, Documents (Le et al., 2014, Denil et al., 2014)

$$sentence \rightarrow \mathbf{x} \in \mathcal{R}^n$$

We train the convolutional network for documents of Denil et al (2014), which only requires labels for documents, but is able to generate features for words, sentences and the documents.



Typical Test Set Result

Paul Bettany did a great role as the tortured father whose favorite little girl dies tragically of disease.

For that, he deserves all the credit.

However, the movie was mostly about exactly that, keeping the adventures of Darwin as he gathered data for his theories as incomplete stories told to children and skipping completely the disputes regarding his ideas.

Two things bothered me terribly: the soundtrack, with its whiny sound, practically shoving sadness down the throat of the viewer, and the movie trailer, showing some beautiful sceneries, the theological musings of him and his wife and the enthusiasm of his best friends as they prepare for a battle against blind faith, thus misrepresenting the movie completely.

To put it bluntly, if one were to remove the scenes of the movie trailer from the movie, the result would be a non descript family drama about a little child dying and the hardships of her parents as a result.

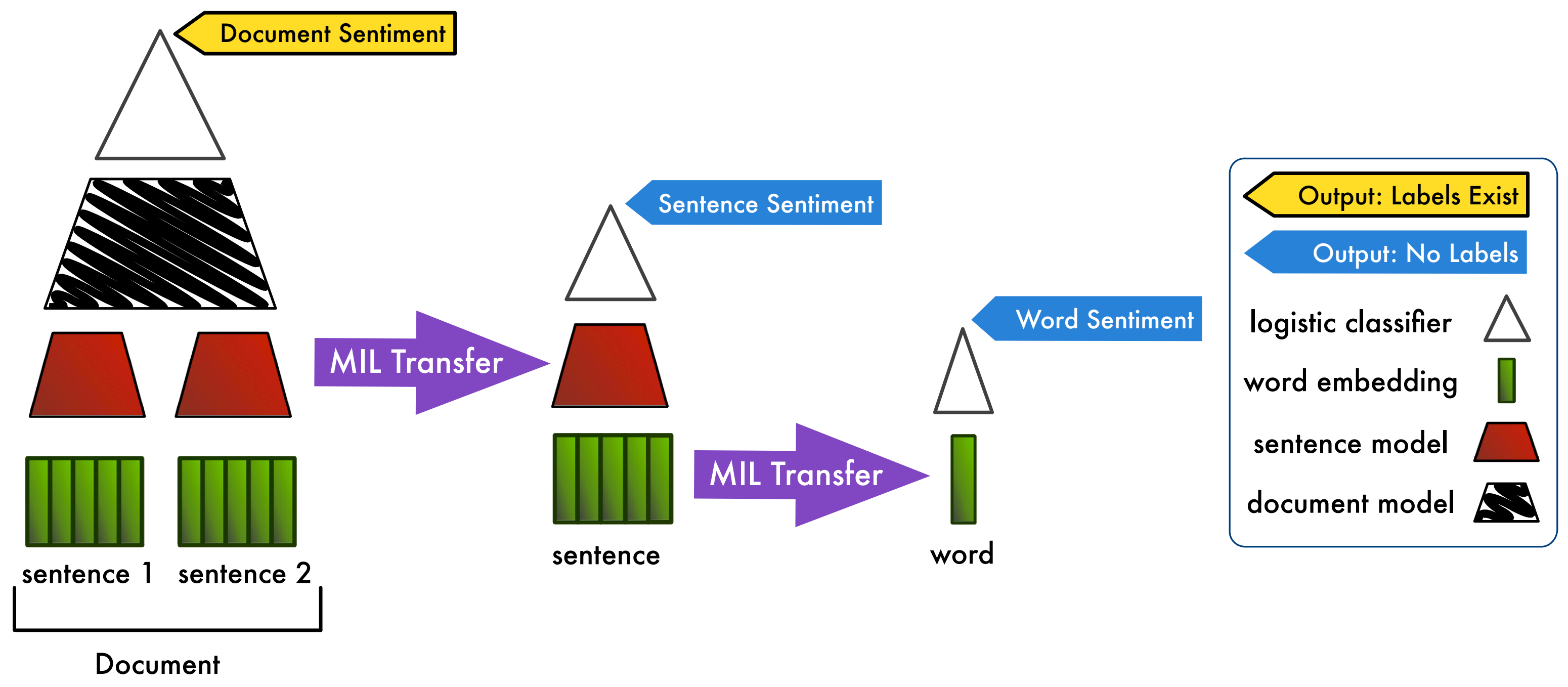
Clearly, not what I expected from a movie about Darwin, albeit the movie was beautifully interpreted.

Classifier prediction for every sentence. Average score predicted is positive. Actual reviewer score : 8/10

Overall Approach

To reduce the cost of human labelling as *groups* << *instances*, we do the following:

- Train the model of Denil et al (2014) for movie reviews to learn sentence embeddings (features) $sentence -> \mathbf{x} \in R^n$
- Since sentence labels are unknown, we introduce a special loss function to learn the parameters $\boldsymbol{\theta}$ of an instance classifier y.
- Can predict the sentiment of any sentence embedding q with $y(q, \theta)$



Proposed Cost Function

$$J(\boldsymbol{\theta}) = \sum_{i,j \in I} W(\mathbf{x}_i, \mathbf{x}_j)\left(y(\mathbf{x}_i, \boldsymbol{\theta}) - y(\mathbf{x}_j, \boldsymbol{\theta})\right)^2 + \alpha \frac{|I|^2}{|G|} \sum_{g \in G} \left(\frac{1}{|g|} \sum_{i \in g} y(\mathbf{x}_i, \boldsymbol{\theta}) - s_g \right)^2$$

- $W(\mathbf{x}_i, \mathbf{x}_j)$ measures the similarity of the instances. This loss states that similar instances should have similar predictions
- This average-matching loss prevents the trivial solution $y = 0$
- Normalising scalar α balances the two objectives

Evaluation

Dataset: imdb.com movie reviews from Maas et. al., 2011

Transferring and Training Results at the Document level

- Average the score of the sentences to create a document score
- Training set accuracy (measures transfer) **94.21%**
- Testing set accuracy (measures classifier generalisation) **88.47%**

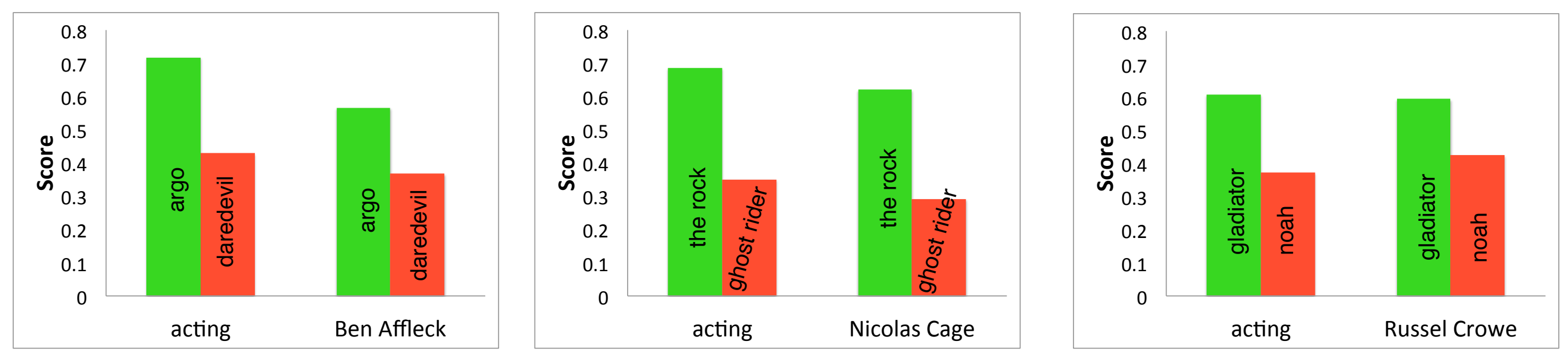
Sentiment Analysis at a Sentence Level

- Manually Tagged 2000 sentences from the dataset
- For each sentence, $sentiment = \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i)$
- Introduce bound b to create neutral class, neutral if $sentiment \in (0.5 - b, 0.5 + b)$

	Precision	Recall
Socher *et al.* 2013 (converting neutrals)	84.5%	100%
MIL Transfer	**85.5%**	100%
Socher *et al.* 2013 (ignoring neutral class)	84.7%	76.2%
MIL Transfer (ignoring neutral class $b = 0.048$)	**92.6%**	76.2%

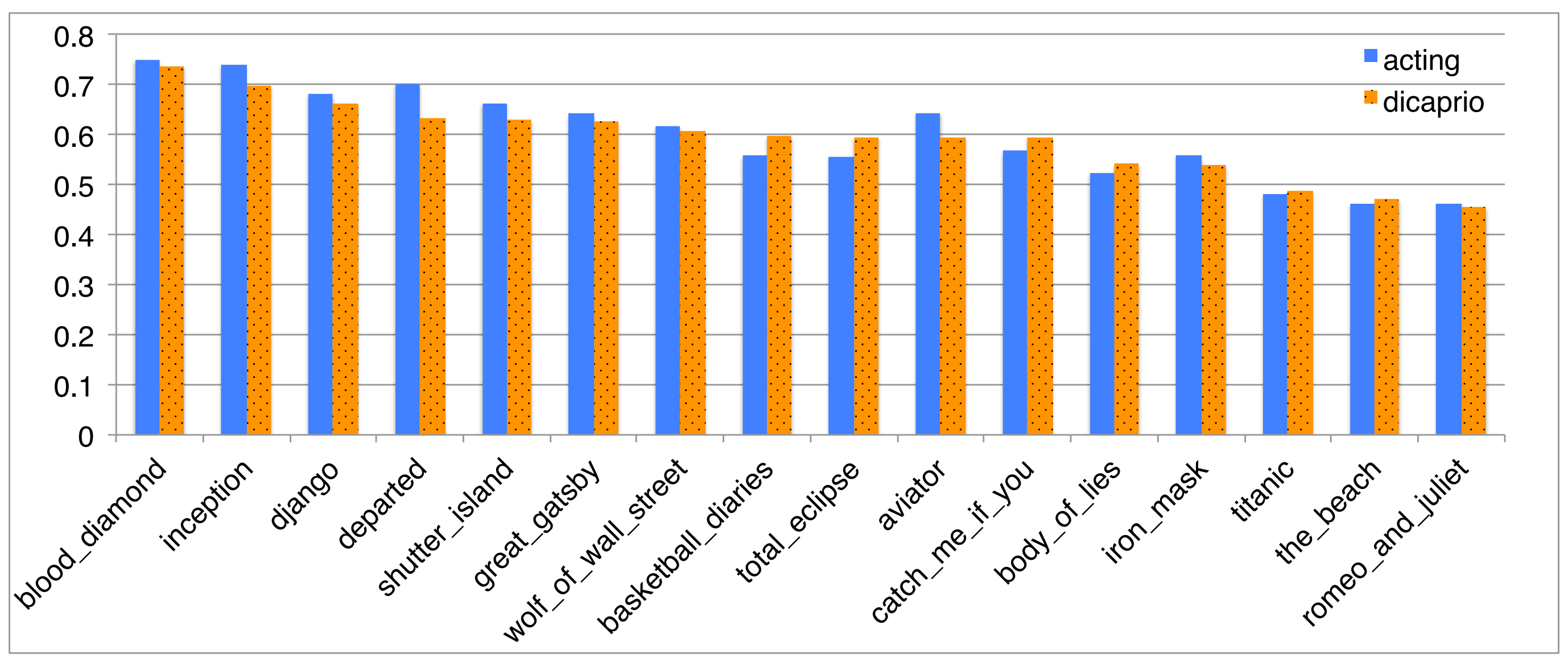
Context Aware Sentiment

- By restricting the training data to different movies, we learn classifiers that predict sentiment within a specific context
- We can use the context-specific classifier to predict the sentiment associated with an actors name or any other phrase



Ben Affleck Nicolas Cage Russel Crowe

Scores associated with the embedding of the word *acting* and the protagonist names, for each movie



Visualisation Results; movies by *Leonardo Di Caprio* sorted by the inferred sentiment for the embedding of his name