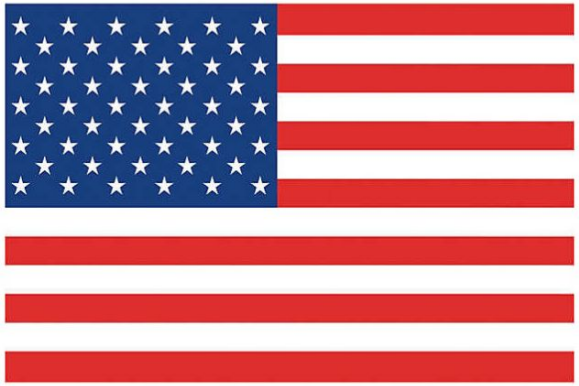
Sparse Multi-label Patent Classification with Deep Learning

W266 Final Project (Fall 2019)
Alexander Mueller & Kevin Stone

Berkeley
UNIVERSITY OF CALIFORNIA

Cooperative Patent Classification (CPC) Codes



CPC Label Hierarchy

Section (A – H, Y)

Class (Two Digits)

Subclass (One Letter)

Group (One to Three Digits)

Example - B60W 20/00

Section B
Performing operations; transporting

Class B60
Vehicles in general

Subclass B60W
Conjoint control of vehicle sub-units

Group B60W 20/00
Control systems specially adapted for hybrid vehicles

Example - B60W 20/00

Section B
Performing operations; transporting

Class B60
Vehicles in general

Subclass B60W
Conjoint control of vehicle sub-units

Group B60W 20/00
Control systems specially adapted for hybrid vehicles

3.3M patents filed in 2018

Berkeley
UNIVERSITY OF CALIFORNIA

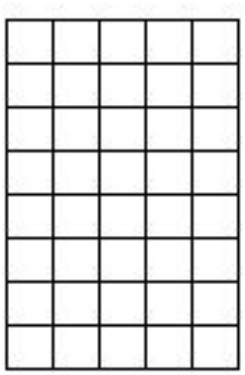
Challenges of multi-label classification

- Sparsity
  - 634 subclass labels in US 2M PTO Dataset
  - 1.3 average, up to 18 per patent
- Unbalanced dataset
- Precision vs. Recall Trade-off

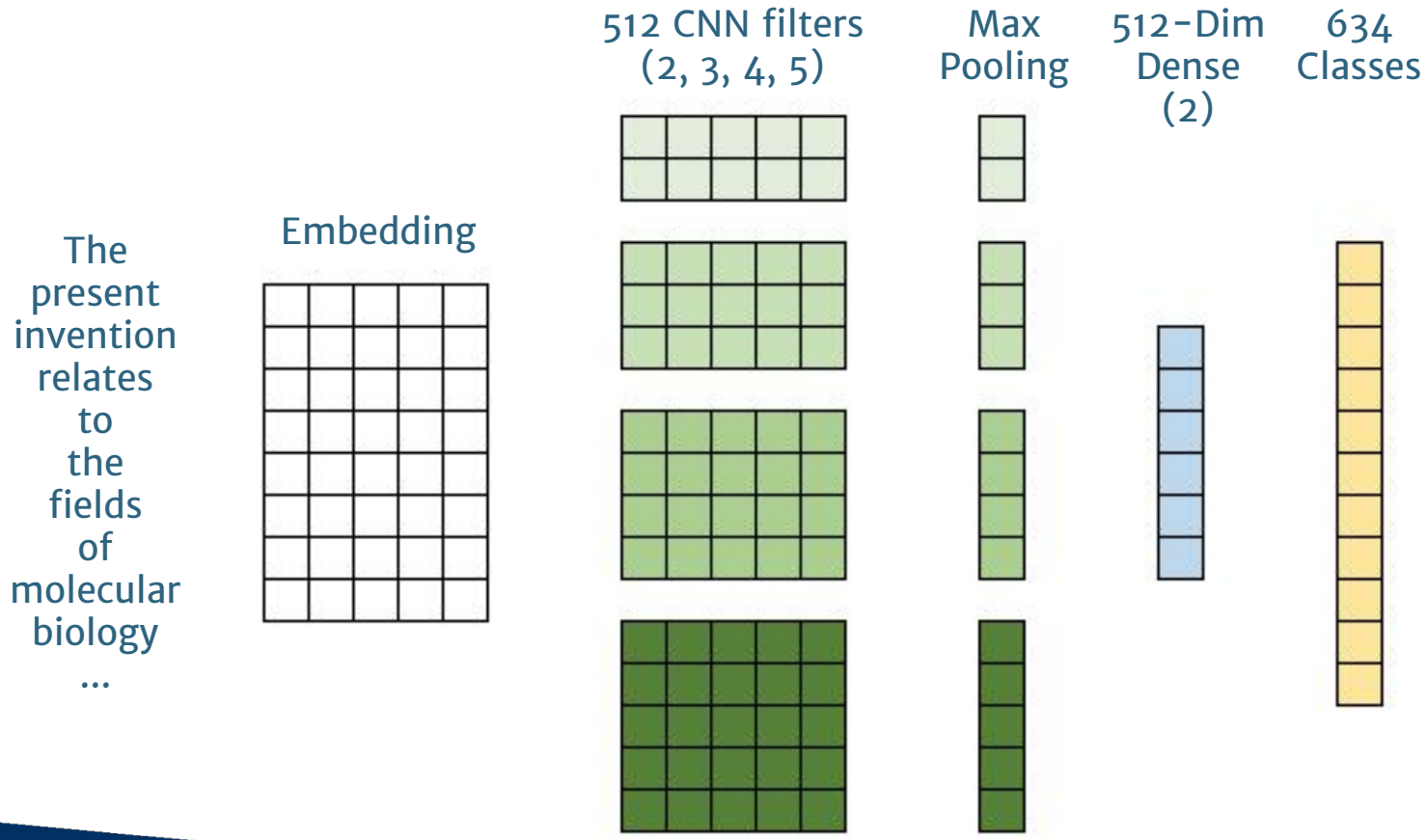
Network architecture

The present invention relates to the fields of molecular biology …

Embedding



Network architecture



Custom loss function

- Weighted binary cross-entropy loss
- Multiplicative coefficient for the positive labels: α
- Drives precision vs. recall tradeoff

$$H_{weighted}(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N} \alpha * y_i * \log(\hat{y}_i) + (1 - y_i) * \log(1 - \hat{y}_i)$$

Results (1)

Model	Loss function[1]	Corpus	Labels	Precision (%)	Recall (%)	F1
GloVe+CNN	BCE	USPTO 2M	632	74.6	45.1	56.2
GloVe+CNN	Weighted BCE	USPTO 2M	632	62.4	**56.3**	**59.2**
BERT+CNN	BCE	USPTO 2M	632	**76.1**	40.1	52.5
BERT+CNN	Weighted BCE	USPTO 2M	632	63.9	52.6	57.8
GloVe+CNN	BCE	USPTO 0.3M	624	73.3	41.1	52.7
GloVe+CNN	Weighted BCE	USPTO 0.3M	624	67.3	50.6	56.2
BERT+CNN	BCE	USPTO 0.3M	624	**74.4**	41.4	53.2
BERT+CNN	Weighted BCE	USPTO 0.3M	624	58.9	**54.8**	**56.8**

1) BCE = Binary cross-entropy

Results (2)

Model	Loss function[1]	Corpus	Labels	Precision (top-1, %)	Recall (top-5, %)
DeepPatent	BCE	USPTO 2M	632	73.9	74.0
PatentBERT	BCE	USPTO 2M	632	**80.6**	**86.1**
GloVe+CNN	BCE	USPTO 2M	632	**66.6**	83.2
GloVe+CNN	Weighted BCE	USPTO 2M	632	66.1	**83.3**
BERT+CNN	BCE	USPTO 2M	632	63.9	80.3
BERT+CNN	Weighted BCE	USPTO 2M	632	63.9	80.7
GloVe+CNN	BCE	USPTO 0.3M	624	67.6	77.5
GloVe+CNN	Weighted BCE	USPTO 0.3M	624	67.3	**77.8**
BERT+CNN	BCE	USPTO 0.3M	624	**68.6**	77.7
BERT+CNN	Weighted BCE	USPTO 0.3M	624	68.1	77.4

1) BCE = Binary cross-entropy

Interesting findings

- Predictions: up to 5 labels per patent, 20% with no labels
- Metrics correlate with number of examples (CPC section F1 scores range from 0.36 to 0.65)
- Training time vs performance tradeoff

Potential Areas of Exploration

- Use 3M+ size patent dataset with more recent patents
- Fine tuning final embedding layers
- Investigating normalization and weighting of the multi-label distribution
- Learning label co-occurrences

Thank you!

W266 Final Project (Fall 2019)
Alexander Mueller & Kevin Stone

Berkeley
UNIVERSITY OF CALIFORNIA