Taming the Sigmoid Bottleneck: Provably Argmaxable Sparse Multi-Label Classification

Andreas Grivas, Antonio Vergari and Adam Lopez

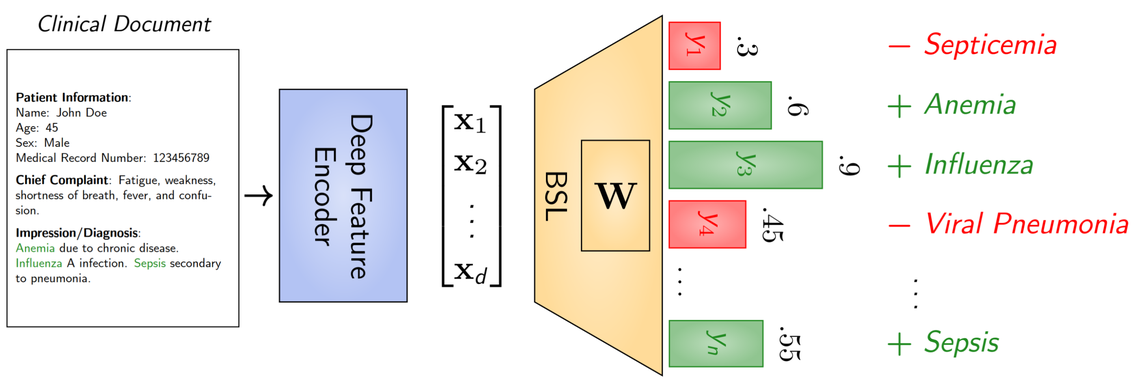
Institute for Language, Cognition, and Computation, University of Edinburgh

TL;DR;

Problem: Bottlenecked multi-label classifiers have outputs that cannot be predicted.

Bottlenecked Sigmoid Layers (BSL)

$$P(\mathbf{y}_i \mid \mathbf{x}) = \sigma\left(\mathbf{w}^{(i)\top}\mathbf{x}\right)$$

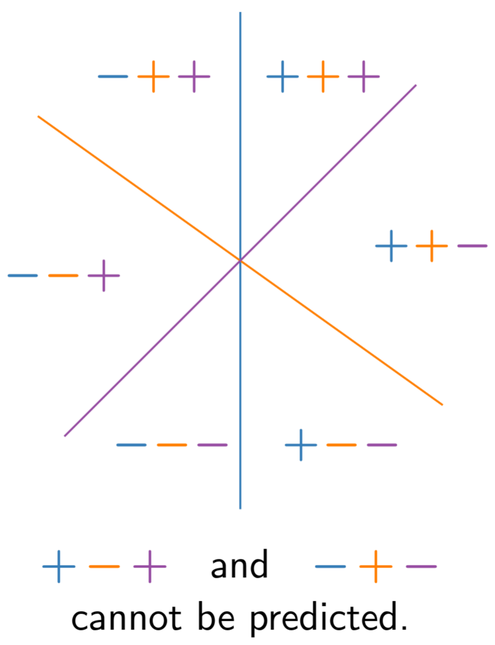


BSL: $n > d$. A linear sigmoid output layer is bottlenecked when its parametrisation, $\mathbf{W}$, is low-rank: the number of input features, d, is less that the number of output labels, n.

Un·argmax·able **Adjective**

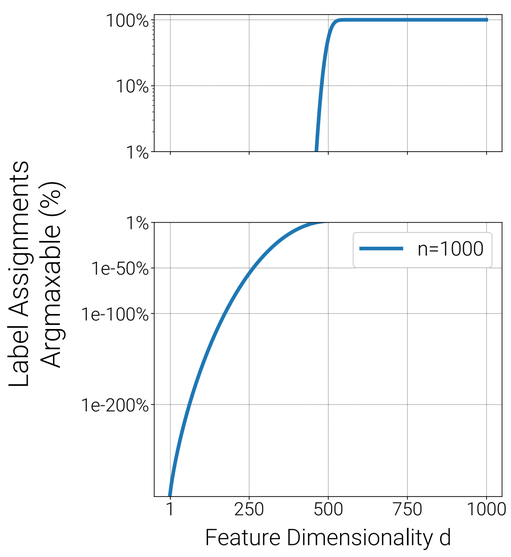
An output that is impossible to predict irrespective of input.

BSLs must have unargmaxable label assignments!



$+-+$ and $-+-$ cannot be predicted.

In a $d = 2$ feature space with $n = 3$ classification hyperplanes, only $6/8$ of label assignments can be predicted.



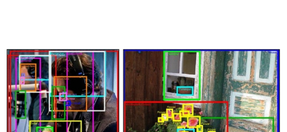
Log-plot of what percentage of the 2^{1000} label assignments is argmaxable for a BSL with $n = 1000$. The percentage drops exponentially as we shrink d.

But... Datasets are sparse (k-active y)

k = Number of $+$ in $\mathbf{y}$, e.g. $-+--+--$ = 2-active

$n \approx 9000$
$k \leq 80$

$n = 20000$
$k \leq 50$



$n \approx 9000$
$k \leq 50$

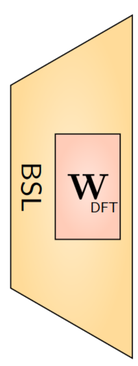
Can we guarantee all k-active label assignments are argmaxable?

Yes. There are $\mathbf{W} \in \mathbb{R}^{n \times (2k+1)}$ such that all k-active labels are argmaxable (see Thm 4 in paper). E.g.: DFT Matrix.

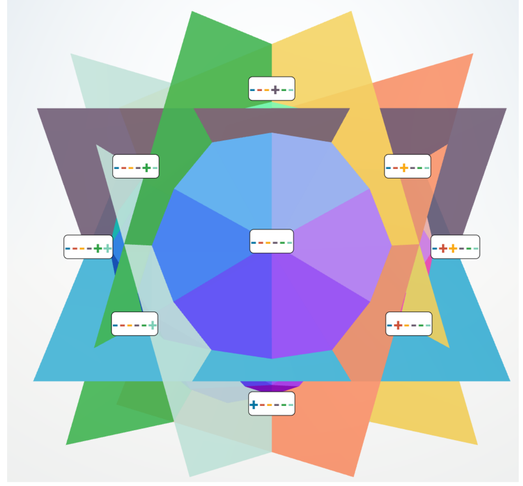
TL;DR;

Solution: We design a classifier which guarantees that sparse outputs can be predicted.

Discrete Fourier Transform (DFT) Layer

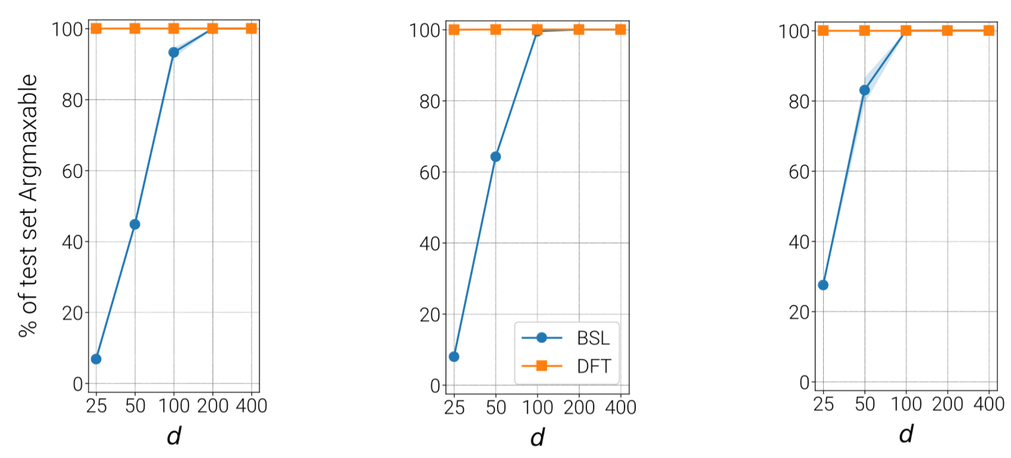


BSL $\rightarrow$ DFT Layer: Replace $\mathbf{W}$ by $\mathbf{W}_{DFT}$.



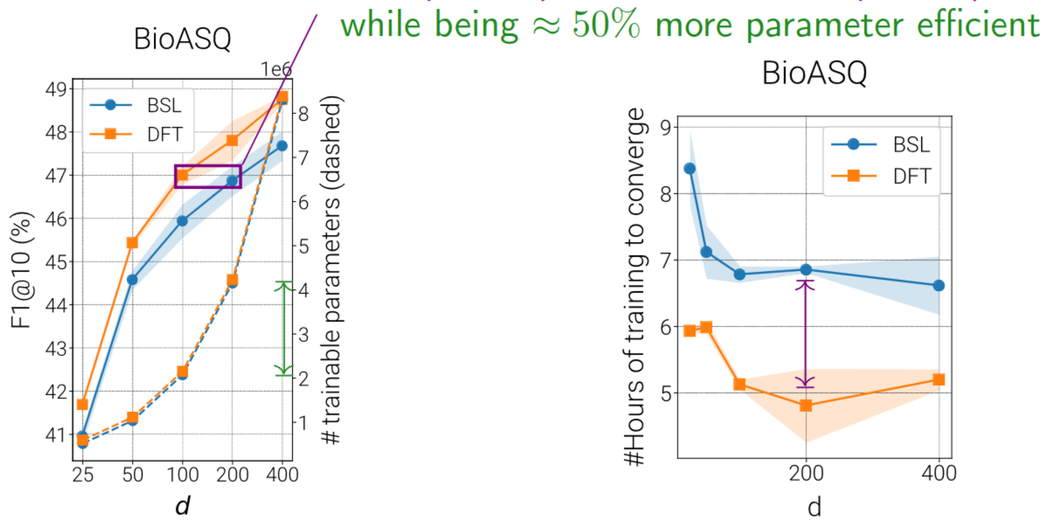
Result: $d = 3$ feature space with $n = 6$ classification hyperplanes formed by $\mathbf{W}_{DFT}$. All 1-active label assignments are argmaxable. See footer for 3D vis.

DFT Layer has argmaxability guarantees



DFT Layer is more efficient

DFT (d=100) outperforms BSL (d=200) while being $\approx 50\%$ more parameter efficient



Conclusion

BSLs must have unargmaxable label assignments. However, since our datasets are often sparse, we can use a DFT layer to guarantee the outputs of interest are argmaxable.

Visualisation

Paper

EPSRC
Engineering and Physical Sciences Research Council

Edinburgh NLP
University of Edinburgh Natural Language Processing