1 Background: multi-label classification

As discussed in Section, the

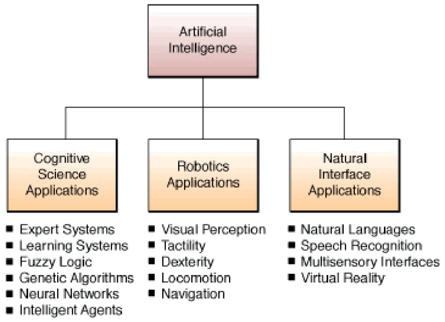

Figure 1: Overview of Artificial Intelligence

(many-shot) retrieval problem explored in this work is essentially a multi-label classification problem. Assuming that $\mathcal{Q}$ and $\mathcal{D}$ denote the query and document (or label) spaces, respectively, the underlying *query-document relevance* is defined by the distribution P where $\mathsf{P}(d|q)$ captures the true relevance for the query-document pair $(q, d) \in \mathcal{Q} \times \mathcal{D}$. In this paper, we assume that the document space is finite with a total $|\mathcal{D}| = L$ documents. Let the training data consists of N examples $\mathcal{S} := \{(q_i, \mathbf{y}_i)\}_{i \in N} \subseteq \mathcal{Q} \times \{0, 1\}^L$, where, for $j \in [L]$, $y_{i,j} = 1$ *iff* jth document in $\mathcal{D}$ is relevant for query q_i. We also denote the set of positive labels for q_i as $P_i = \{j : y_{i,j} = 1\}$

Dual-encoder models and classification networks. Note that learning a retrieval model is equivalent to learning a *scoring function* (or simply *scorer*) $s_{\boldsymbol{\theta}} : \mathcal{Q} \times \mathcal{D} \to \mathbb{R}$, which is parameterized by model parameters $\boldsymbol{\theta}$. A DE model consists of a *query encoder* $f_{\boldsymbol{\phi}} : \mathcal{Q} \to \mathbb{R}^d$ and a *document encoder* $h_{\boldsymbol{\psi}} : \mathcal{D} \to \mathbb{R}^d$ that map query and document features to d-dimensional embeddings, respectively. Accordingly, the corresponding scorer for the DE model is defined as: $s_{\boldsymbol{\theta}}(q, d) = \langle x_q, z_d \rangle = \langle f_{\boldsymbol{\phi}}(q), h_{\boldsymbol{\psi}}(d) \rangle$, where $\boldsymbol{\theta} = [\boldsymbol{\phi}; \boldsymbol{\psi}]$ and $x_q, z_d \in \mathbb{R}^d$. Note that it is common to share encoders for query and document, i.e., $\boldsymbol{\phi} = \boldsymbol{\psi}$. We also focus on this shared parameter setting in our exploration.

Different from DE models, classification networks ignore document features. Such networks consist of a query encoder $g_{\boldsymbol{\phi}} : \mathcal{Q} \to \mathbb{R}^d$ and a classification matrix $W \in \mathbb{R}^{L \times d}$, with ith row as the classification weight vector for the ith document. The scorer for the classification network then becomes $s_{\boldsymbol{\theta}}(q, d) = \langle g_{\boldsymbol{\phi}}(q), \mathbf{w}_d \rangle$, where $\mathbf{w}_d$ denotes the classification vector associated with document d and $\boldsymbol{\theta} = [\boldsymbol{\phi}; W]$. For ease of exposition, we do not always highlight the explicit dependence on trainable parameters $\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\psi}$ while discussing DE models and classification networks.

Loss functions. Given the training set $\mathcal{S}$, one learns a scorer by minimizing the following objective:

$$\mathcal{L}(s; \mathcal{S}) = \frac{1}{N} \sum\nolimits_{i \in [N]} \ell(q_i, \mathbf{y}_i; s), \tag{1}$$

where ℓ is a *surrogate loss-function* for some specific metrics of interest (e.g., Precision@k and Recall@k). Informally, $\ell(q, \mathbf{y}; s)$ serves as a proxy of the quality of scorer s for query q, as measured by that metric with $\mathbf{y}$ as ground truth labels. One of the popular loss functions for multi-label classification problem is obtained by employing one-vs-all (OvA) reduction to convert the problem into L binary classification tasks (Lapin, Hein, and Schiele, 2016). Subsequently, we can invoke a binary classification loss such as sigmoid *binary* *cross-entropy* (BCE) for L tasks, which leads to OvA-BCE:

$$\ell(q, \mathbf{y}, s) = -\sum\nolimits_{j\in[L]} \Big(y_j \cdot \log \frac{e^{s(q,d_j)}}{1 + e^{s(q,d_j)}} + (1 - y_j) \cdot \log \frac{1}{1 + e^{s(q,d_j)}} \Big). \tag{2}$$

Alternatively, one can employ a multi-label to multi-class reduction (Menon et al., 2019) and invoke softmax cross-entropy (SoftmaxCE) loss for each positive label:

$$\ell(q, \mathbf{y}, s) = -\sum\nolimits_{j\in[L]} y_j \cdot \log \Big(e^{s(q,d_j)} / \sum\nolimits_{l\in[L]} e^{s(q,d_l)} \Big). \tag{3}$$

Our empirical findings in Section reveal that DE models fail to train with BCE loss. We hypothesize that this might be due to the stringent demand of the BCE loss, which requires all negative pairs to exhibit high absolute negative similarity and all positive pairs to have a high absolute positive similarity – a feat that can be challenging for DE representations. In the section below we analyze SoftmaxCE loss (an extreme case of InfoNCE) in more detail and highlight its shortcomings.

References

Dembczyński, Krzysztof, Weiwei Cheng, and Eyke Hüllermeier (2010). “Bayes Optimal Multilabel Classification via Probabilistic Classifier Chains”. In: *Proceedings of the 27th International Conference on International Conference on Machine Learning*. ICML’10. Haifa, Israel: Omnipress, pp. 279–286. ISBN: 9781605589077.

Lapin, Maksim, Matthias Hein, and Bernt Schiele (2016). “Loss functions for top-k error: Analysis and insights”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1468–1477.

Menon, Aditya K et al. (2019). “Multilabel reductions: what is my loss optimising?” In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc. URL: `https://proceedings.neurips.cc/paper_files/paper/2019/file/da647c549dde572c2c5edc4f5bef039c-Paper.pdf`.

Reddi, Sashank J. et al. (2019). “Stochastic Negative Mining for Learning with Large Output Spaces”. In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Ed. by Kamalika Chaudhuri and Masashi Sugiyama. Vol. 89. Proceedings of Machine Learning Research. PMLR, pp. 1940–1949. URL: `https://proceedings.mlr.press/v89/reddi19a.html`.