CLaC at BioCreative VII LitCovid Track: Independent modules for multi-label classification of Covid articles

Parsa Bagherzadeh and Sabine Bergler
CLaC Labs, Concordia University, Montreal, Canada
{p_bagher / bergler} @ cse.concordia.ca

***Abstract*—For multi-label classification of Covid articles we present a modular system where the representation learning for each class is performed by separate, independent modules that occasionally interact. To further improve the performance of the system, we also leverage specialized gazetteer lists using an additional module. All of our submitted runs outperform the competition mean and our best run performs well in terms of marco-averaged F1 score.**

***Keywords*—Multi-input RIM, modular model, multi-label classification, knowledge sources**

I. Introduction

BioCreative VII Track 5 concerns muti-label classification of abstracts from Covid-related articles into 7 classes, namely: *Treatment*, *Mechanism*, *Prevention*, *Case Report*, *Diagnosis*, *Transmission*, and *Epidemic Forecasting* (1) and (2).

In a multi-label classification setting, a base network is responsible for embedding the input for all classes, often followed by a classification layer that uses this unified representation. Although the classes might be related, different classes often require focus on different parts of the input. We suggest that decoupling fine tuning is effective.

To address the multi-label classification of Covid articles, we use the multi-input RIM (mi-RIM) architecture (3), which comprises M independent, yet interacting recurrent modules. A cap can be set for the number of modules allowed to be active at each time step, which leads to competition among modules. As argued by (4), in competition, modules focus on specific parts of the input and subsequently, on simpler subproblems.

Here we model a multi-label classification problem with C labels as C independent binary classification tasks. The representation learning for each task is performed by an independent module in the mi-RIM architecture. Moreover, we leverage external knowledge sources (specialized gazetteer lists) injected in an additional independent module.

II. Overview of mi-RIM

Multi-input Recurrent Independent Mechanisms (mi-RIM) is a modular architecture that models a dynamic system by dividing it into M recurrent units. The units are selective, i.e they chose to use or ignore their input, and are able to communicate with one another (3).

A. Input selection

Each module R_m augments the token input x_t^m to $X_t^m = x_t^m \oplus \mathbf{0}$, where $\mathbf{0}$ is an all-zero vector and $\oplus$ denotes row-level concatenation. Then, using an attention mechanism, unit R_m selects input:

$$A_t^m = softmax\left(\frac{h_{t-1}^m W_m^{query}(K_m)^T}{\sqrt{d_h}}\right)V_m \tag{1}$$

where $h_{t-1}^m W_m^{query}$ is the *query*, $K_m = X_t^m W_m^{key}$ is the *key*, and $V_m = X_t W_m^{val}$ is the *value* in the attention mechanism. If the input x_t is considered relevant to the task, the attention mechanism in Equation 1 assigns more weight to it (selects it), otherwise more weight will be assigned to the null input. The $softmax$ values of Equation 1 determine a ranking for the modules and a subset S_t of the k highest ranked units. Among M units, those with the least attention on the null input are the active units. The selected input A_t^m determines a temporary hidden state $\tilde{h}_t^m$ for the active units:

$$\tilde{h}_t^m = R_m(h_{t-1}^m, A_t^m) \quad m \in S_t \tag{2}$$

where $R_m(h_{t-1}^m, A_t^m)$ denotes one iteration of updating the recurrent unit R_m based on the previous state h_{t-1}^m and the current input A_t^m. The hidden states of the inactive units R_m ($m \notin S_t$) remain unchanged ($h_t^m = h_{t-1}^m \; m \notin S_t$).

B. Interaction

To obtain the actual hidden states h_t^m, the active units communicate using an attention mechanism:

$$h_t^m = softmax\left(\frac{Q_{t,m}(K_{t,:})^T}{\sqrt{d_h}}\right)V_{t,:} + \tilde{h}_t^m \quad m \in S_t \tag{3}$$

where

$$Q_{t,m} = \tilde{h}_t^m \tilde{W}_m^{query}$$

$$K_{t,:} = [\tilde{h}_t^1 \tilde{W}_1^{key} \oplus \ldots \oplus \tilde{h}_t^M \tilde{W}_M^{key}]$$

$$V_{t,:} = [\tilde{h}_t^1 \tilde{W}_1^{val} \oplus \ldots \oplus \tilde{h}_t^M \tilde{W}_M^{val}]$$

Both the key $K_{t,:}$ and the value $V_{t,:}$ depend on the temporary hidden states of all units, therefore h_t^m in Equation 3 is determined by attending to all units.

C. Limit on active modules

A limit on active modules at each time step can be imposed. Limiting the number of active modules at each time step does not set an upper bound for the number of predicted labels and the model can make positive predictions for all of the labels.

III. Muli-input multi-label RIM

Independent modules can effectively inject external knowledge into neural nets (3). The configuration has to be matched to the task. Here we use an independent module for each class and one for injecting information from DrugBank and MeSH.

A. Gazetteer lists

We compile gazetteer lists from two expert curated on-line resources: DrugBank[1] (5) and MeSH[2] (6).

a) Drug: names of drugs compiled from DrugBank

Example 1: Inhibition of IL-1 by Anakinra (ANK) is potentially life-saving for severe CSS cases.

b) Therap: list of therapeutics compiled from subtree E02 of MeSH

Example 2: ...to undergo Continuous Positive Airway Pressure (CPAP) or Non-Invasive Positive Pressure Ventilation (NIPPV) due to ...

c) Diag: list of diagnostic techniques and procedures from subtree E01.370 of MaSH

Example 3: ...including plain radiography, computed tomography and magnetic resonance imaging, were performed ...

d) Prev: list of terms relating to prevention and public health practice from node N06.850.780 of MeSH

Example 4: ...and mass screening are the most common non-pharmaceutical PHIs to cope with the epidemic

e) Trans: terms relating to disease transmission from node N06.850.335 of MeSH

Example 5: This raises concern that health workers could act as silent disease vectors

An embedding layer $E_{gaz} \in \mathbb{R}^{(5+1)\times 20}$ encodes gazetteer annotations in 20 dimensions. Each row in E_{gaz} embeds one of the 5 gazetteer lists and an additional row encodes no gazetteer matches.

B. Architecture

We use a mi-RIM with two sets of recurrent modules, namely seven class modules and one gazetteer module.

a) Class modules: $\mathcal{M}_c = \{R_1, \ldots, R_7\}$, where each module $R_m \in \mathcal{M}_c$ is devised to fine tune for its corresponding class. This enables differential embeddings of the same input (e.g. focusing on different lexical triggers) for different classes. Since the modules are independent, they can develop their own expertise. We use the token representations $\langle x_1, \ldots, x_T \rangle$ provided by ClinicalBERT[3] (7) as input to all class modules.

b) Gazetteer module: Recurrent module R_8 is responsible to inject gazetteer annotations. The embedding layer E_{gaz} provides input to this module.

A class module R_m interacts with other class modules as well as the gazetteer module R_8 and the hidden states $h_1^m, \ldots, h_T^m$ are obtained by attending to the hidden states of all other modules.

For each class module R_m, $m \in 1...7$ we consider a dedicated binary classifier f_m. The input to the classifier is obtained by applying attention to hidden states h_t^m, $t = 1, \ldots, T$ of module R_m:

$$H^m = \text{softmax}(w_{att} K_m) V_m \tag{4}$$

where $K_m = V_m = [h_1^m \oplus \ldots \oplus h_T^m]$ and w_{att} is a learnable vector. Each module with a positive prediction adds it's class label to the overall multi-label prediction of the model. As an example, if $\hat{y}_1 = 1$ and $\hat{y}_7 = 1$, the overall prediction is [*Treatment, Epidemic Forecasting*].

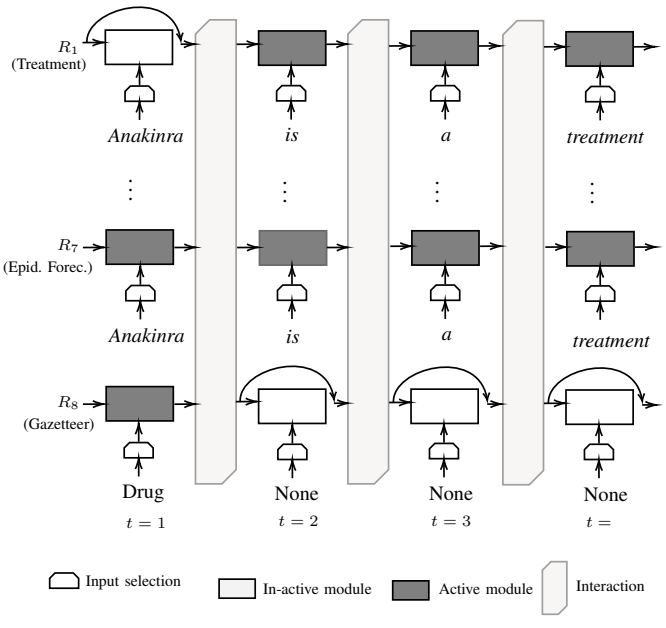

Fig. 1. An overview of the proposed system. Each class module ($R_1 - - R_7$) is responsible to learn representations for its corresponding class

Figure 1 provides an overview over the system. We use LSTMs for the class modules and a simple RNN for the gazetter module. The model is implemented using PyTorch (8)

[1]https://go.drugbank.com/

[2]https://meshb.nlm.nih.gov/treeView

[3]HuggingFace: https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT

System	k	Treatment			Mechanism			Prevention			Case Rep.			Diagnosis			Transmi.			Epid. Forc.			macro			micro		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
ClinBERT	-	.86	.85	.85	.87	.82	.85	.87	.91	.89	.88	.86	.87	.79	.87	.83	.51	.73	.62	.72	.71	.71	.78	.82	.80	.83	.86	.84
$\mathcal{M}_c$	7	.90	.86	.88	.88	.84	.86	.90	.95	.92	.89	.87	.88	.83	.89	.86	.59	.69	.64	.76	.69	.72	.82	.82	.82	.87	.88	.87
	3	.87	.90	.89	.84	.87	.85	.94	.93	.94	.93	.86	.89	.81	.91	.86	.66	.68	.67	.73	.76	.75	.82	.84	.83	.87	.89	.88
$\mathcal{M}_c + \mathcal{M}_g$	8	.90	.91	.90	.86	.87	.87	.95	.94	.95	.92	.88	.90	.84	.90	.87	.68	.68	.68	.74	.77	.76	.84	.85	.84	.89	.90	.89
	3	**.92**	**.91**	**.91**	**.90**	**.89**	**.89**	**.95**	**.95**	**.95**	**.91**	**.92**	**.92**	**.90**	**.92**	**.91**	**.68**	**.70**	**.70**	**.77**	**.77**	**.77**	**.86**	**.86**	**.86**	**.91**	**.91**	**.91**

Fig. 2. Ablation on development set. k indicates the number of active modules.

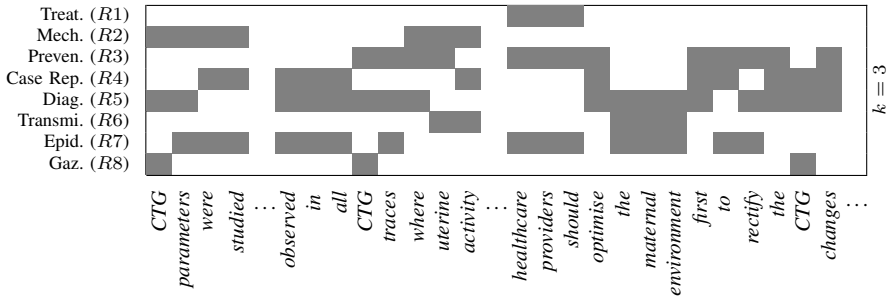

Fig. 3. Gold=[*Diagnosis*, *Treatment*], Prediction[*Diagnosis*]

and optimized with the Adam optimizer (9) with $lr = 0.5e-5$ for 5 epochs. Note that the overall classification loss is the sum of the losses of all classifiers.

IV. Results

We compare our system without the gazetteer module to a baseline of ClinicalBert (7) and to the full system that also includes the gazetteer module in Figure 2.

a) ClinBERT: For our baseline, we use the [CLS] token of ClinicalBERT as the input to a classification layer with 7 neurons with Sigmoid activation. This is common practice for multi-label classification problems

b) $\mathcal{M}_c$: mi-RIMs for class modules $\mathcal{M}_c$ only

c) $\mathcal{M}_c + \mathcal{M}_g$: class modules and gazetteer module (the full architecture)

A. Development phase

a) Numerical results: Table 2 reports performance on the development set provided by the organizers. Although for most of the classes, ClinicalBERT yields a relatively high F1 score, the classes *Transmission* and *Epidemic Forecasting* show inferior performances. In fact, the two classes are the two least frequent classes in the training data.

All variants of the proposed system outperform the baseline significantly and consistently, confirming previous observations made by (3) and (4). Moreover, limiting the number of active modules (forcing the modules into competition) yields further improvements.

Forcing the modules into competition focuses the modules on different parts of the input and consequently specializes them for their corresponding class. The gazetteer module $\mathcal{M}_g$ further improves the performance of the models, leading to the best performance across all runs.

b) Activation patterns: Figures 3–5 show activation patterns of the modules for the best performing configuration $\mathcal{M}_c + \mathcal{M}_g$ ($k = 3$) for three different abstracts. As argued by (10), the activation patterns can provide some insight into the functioning of the model and explain why a prediction was made, which can be used for error analysis.

Figure 3 shows the activation patterns for different snippets of an abstract. The model makes a true positive prediction for *Diagnosis* and a false negative for *Treatment*. The gazetteer module R_8 is active for the mentions of *CTG* (Cardiotocography, a diagnostic technique), which supports the label *Diagnosis*. Module R_5 is active for parts of the input such as *CTG parameter*, *observed in all CTG traces*, and *the CTG changes*, all related to class $Diagnosis$. This shows that the module has specialized for classification of its corresponding class. The module R_1 (responsible for class *Treatment*) however fails to focus on the phrases *optimise the maternal environment* and *rectify CTG changes* both indicating *Treatment*, explaining the false negative for *Treatment*.

Another example of activation patterns is provided in Figure 4. For this example, the model predicts the label *Treatment*, which is a true positive prediction. Module R_8 (gazetteer module) is active for a mention of drug (*Tocilizumab*) supporting a prediction for the *Treatment* class. Moreover, module R_1 is active for phrases "*monoclonal antibody against*" and "*clinical benefits in*", both further evidence for the *Treatment* class.

Figure 5 illustrates another activation pattern. In this example, the model makes two true positive predictions (*Transmission* and *Prevention*) and one false positive prediction (*Case Report*). The module R_3 is active for the phrase "*hand hygiene, frequent cleaning and disinfecting*" which is a strong indicator

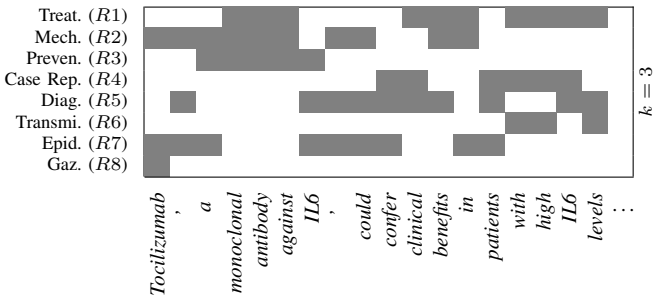

Fig. 4. Gold=[*Treatment*] Prediction=[*Treatment*]

for the *Prevention* class, and consequently the model makes a true positive classification. Moreover, the module R_6 is active for "*Transmission was observed*" which explains the true positive prediction for class *Transmission*. Interestingly, the module R_4 (corresponding to the *Case Report* class) focuses on the phrase "*Transmission was observed from two of three children*". This phrase could support the *Case Report* class, and the model reports the corresponding label as one of the predictions, nevertheless, the annotation does not report a label for this class, and this prediction is considered as a false positive.

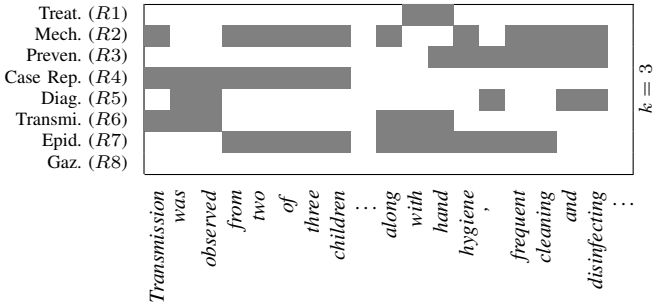

Fig. 5. Gold=[*Transmission*, *Prevention*], Prediction=[*Transmission*, *Prevention*, *Case Report*]

B. Evaluation phase

The official competition result is provided in Table 6. Our best performing run is the full system $\mathcal{M}_c + \mathcal{M}_g$, with a cap of $k = 3$ on the number of active modules.

		macro			micro		
System	k	P	R	F1	P	R	F1
$\mathcal{M}_c$	7	.82	.82	.82	.87	.86	.87
	3	.82	.83	.82	.87	.86	.87
$\mathcal{M}_c + \mathcal{M}_g$	8	.84	.85	.84	.88	.89	.88
	3	.84	**.86**	.84	.88	.89	.88
Baseline (ML-Net) (11)		.83	.73	.76	.87	.81	.84
Mean (all teams)		.86	.80	.81	.89	.86	.87
Std (all teams)		.06	.07	.07	.05	.04	.04

Fig. 6. Official test results

The full system with gazetteer module outperforms the class modules without gazetteers, further supporting the case for knowledge injection. All our runs beat the official baseline in recall and F1 and nearly tie for precision. Similarly, our best run beats the competition mean in recall and F1, but not in precision.

V. Conclusion

For the multi-label classification of abstracts from CoVID related articles, the addition of external knowledge from different MeSH nodes is effective. The injection of the external knowledge and a competitive system of seven classifiers for seven labels is implemented in the architecture of independent, interacting modules (mi-RIMs). The independence of the representation learning components decouples the different labels and allows them to specialize. Interaction additionally allows them to benefit from each other's hidden states.

The proposed system can be inspected trough visualization of activation patterns, beneficial for inspecting system behavior for individual samples, such as error cases

References

1. Chen, Q., Allot, A. & Lu, Z. LitCovid: an open database of COVID-19 literature. *Nucleic acids research* **49,** D1534–D1540 (2021).
2. Chen, Q., Allot, A. & Lu, Z. Overview of the BioCreative VII LitCovid Track: multi-label topic classification for COVID-19 literature annotation. *Proceedings of the seventh BioCreative challenge evaluation workshop* (2021).
3. Bagherzadeh, P. & Bergler, S. *Multi-input Recurrent Independent Mechanisms for leveraging knowledge sources: Case studies on sentiment analysis and health text mining* in *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures* (2021), 108–118.
4. Goyal, A. *et al.* Recurrent independent mechanisms. *arXiv preprint arXiv:1909.10893* (2019).
5. Wishart, D. S. *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic acids research* **46,** D1074–D1082 (2018).
6. Lipscomb, C. E. Medical Subject Headings (MeSH). *Bulletin of the Medical Library Association* **88** (2000).
7. Alsentzer, E. *et al. Publicly Available Clinical BERT Embeddings* in *Proceedings of the 2nd Clinical Natural Language Processing Workshop* (2019), 72–78.
8. Paszke, A. *et al. Automatic differentiation in PyTorch* in *NIPS 2017* (2017).
9. Kingma, D. P. & Ba, J. L. *Adam: A method for stochastic optimization* in *Proceedings of the 3rd International Conference on Learning Representations, ICLR'15* (2015).

10. Bagherzadeh, P. & Bergler, S. *Interacting Knowledge Sources, Inspection and Analysis: Case-studies on Biomedical text processing* in *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP* (2021).
11. Du, J. *et al.* ML-Net: multi-label classification of biomedical texts with deep neural networks. *Journal of the American Medical Informatics Association* **26,** 1279–1285 (2019).