Agent-based Splitting of Patient-Therapist Interviews for Depression Estimation

Navneet Agarwal[1], **Gaël Dias**[1], **Sonia Dollfus**[2]
[1]Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France.
[2]CHU de Caen, Service de Psychiatrie; Normandie Univ, UNICAEN, ISTS,
GIP Cyceron; Normandie Univ, UNICAEN, UFR de Médecine, 14000 Caen, France.
navneet.agarwal@unicaen.fr, gael.dias@unicaen.fr, dollfus-s@chu-caen.fr

Abstract

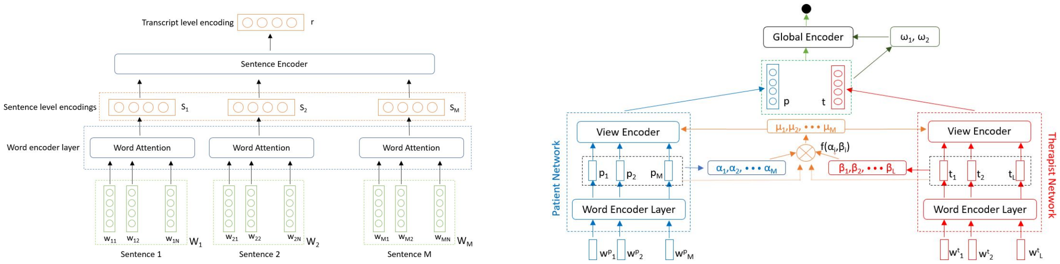
There has been considerable research in the field of automated mental health analysis. Studies based on patient-therapist interviews usually treat the dyadic discourse as a sequence of sentences, thus ignoring individual sentence types (question or answer). To avoid this situation, we design a multi-view architecture that retains the symmetric discourse structure by dividing the transcripts into patient and therapist views. Experiments on the DAIC-WOZ dataset for depression level rating show performance improvements over baselines and state-of-the-art models.

1 Introduction

Depression is a serious mental disorder that affects millions worldwide, and an increasing curve is expected as a consequence of the current health crisis [20]. Detection of depression is a challenging problem with patient-therapist interviews being the common practice to analyse a patient's mental health. Within such dialogues, the therapist looks for indicative symptoms such as loss of interest, sadness, exhaustion, sleeping and eating disorders. Complementary to these interviews, different screening tools have been defined such as the Personal Health Questionnaire depression scale, with PHQ-8 being considered a valid diagnosis and severity measure for depressive disorders [9]. Throughout the literature, different strategies have been proposed for the automatic estimation of depression, which consists of inferring the screening tool score based on the interview transcript. Multi-modal models combine inputs from different modalities [17, 16, 12]. Multi-task architectures simultaneously learn related tasks [16, 15]. Gender-aware models explore the impact of gender on depression estimation [1, 13]. Hierarchical models process transcripts at different granularity levels [10, 19]. Graph-based models encode non linear interactions within the input data [12, 7]. Attention models integrate external knowledge from mental health lexicons [19]. And feature-based solutions compute multiple multi-modal characteristics [2].

Despite this extensive list of research initiatives, ways to express the structure of an input transcript remains an unexplored research direction. Indeed, most related works either exclusively focus on the patient utterances [10, 2], or treat the overall transcript as a sequence of sentences, considering that the questions asked by the therapist contain informative content [19]. Nevertheless, this latter case disregards the type of individual sentences as questions (therapist) or answers (patient), and forces the model to understand the inter-dependencies within a sequence of unstructured utterances. In this paper, we argue that the structure of the interview plays an important role along with sentences type. For that purpose, we design multi-view architectures that separate a dialogue stream into two different views, i.e. the therapist view and the patient view. As such, the interview structure is taken into account by learning interactions (1) within the views i.e. interactions between questions only and answers only, and (2) between the two views i.e. interactions between the corresponding questions and answers. This allows the models to focus on specific structures of the transcript as

PAI4MH @ 36st Conference on Neural Information Processing Systems (NeurIPS 2022), New Orleans, USA.



(a) Hierarchical model

(b) Multi-view architecture

Figure 1: Overall models. (a) non-RNN based implementation of the hierarchical model; (b) Multi-view architecture where the intra-view information is outlined in red and blue, the inter-view linking is painted in orange, and the view fusion network is shown in green.

well as control the discourse symmetry. Experiments over the DAIC-WOZ dataset [6] show that a multi-view architecture with inter-view attention outperforms baselines models and provides new state-of-the-art results on the test split for binary depression estimation.

2 Multi-view Architecture

Similarly to the patient input during an interview, the questions asked by the therapist also convey relevant information for depression estimation (cf. §3). To take into account both patient and therapist information so that discourse symmetry and structure are kept, we design a multi-view architecture that divides the input discourse into two views, i.e. the therapist view and the patient view. Figure 1b illustrates the proposed architecture. In particular, the networks corresponding to the two views, i.e. the *patient network* and the *therapist network*, are instances of a hierarchical model[1] [19] (see figure 1a), and learn transcript level view representations p (patient) and t (therapist). The *sentence encoders* from the hierarchical model are renamed as *view encoders* in the multi-view architecture with $\alpha_1, \alpha_2, ..., \alpha_M$ and $\beta_1, \beta_2, ..., \beta_L$ being the corresponding sentence attention scores, and $p_1, p_2, ..., p_M$ and $t_1, t_2, ..., t_L$ the respective sentence level encodings. Within the multi-view architecture, we define two categories of models, (1) multi-view strategies with intra-view attention, and (2) multi-view strategies with inter-view attention.

2.1 Multi-view Strategies with Intra-view Attention

Here, the two views are treated independently of each other, as represented in Figure 1b with blue and red colors. The underlying idea is that each view can be processed individually before they are fused to encode the information at transcript level containing questions and answers. This architecture includes two view-level attention layers (*view encoders*) and a global attention layer (*global encoder*). Both view-level attention layers are defined using self-attention and combine sentence level features within the respective views. The global attention layer is also defined as a self-attention model aimed at fusing transcript-level view representations p and t. Within this context, three configurations can be defined for an ablation study, where the self-attention layers are the adjustment variables.

2.2 Multi-view Strategies with Inter-view Attention

Within the context of intra-view attention models, questions and answers are treated independently, and their codependency is not tackled. However, the coherent structure of a dialogue plays an essential role for the global understanding of the message conveyed by the patient. Let's say that a patient answers "yes", which in itself does not hold any meaning or relevance. If we look at it in context of the therapist question "do you feel extremely tired?", its relevance towards the final outcome is obvious. Similarly, if the same answer is given to the following question, "can you tell me more about yourself?", this issue should not impact the final decision. As a consequence, tackling the

[1] Detailed explanation of our implementation is provided in appendix A.1.

Architectures		macro F1		UAR		Accuracy		macro Precision	
		(Dev)	Test	(Dev)	Test	(Dev)	Test	(Dev)	Test
Baseline	Patient	(0.6413)	0.6429	(0.6369)	0.6361	(0.6969)	**0.7608**	(0.6725)	0.6584
	Therapist	(0.8253)	0.5818	(0.8095)	0.5803	(0.8484)	0.6521	(0.8611)	0.6184
	Patient+Therapist	(0.7555)	0.6053	(0.7440)	0.6004	(0.7878)	0.6739	(0.7847)	0.6250
MV-Intra-Att.	View-Global Attention	(0.6944)	0.6811	(0.6845)	0.6674	(0.7575)	0.7391	(0.7870)	0.7252
	Global Attention	(0.6857)	0.7116	(0.6785)	0.7075	(0.7272)	0.7173	(0.7083)	0.6887
	View Attention	(0.6944)	0.6919	(0.6845)	0.6919	(0.7575)	0.6739	(0.7870)	0.6919
MV-Inter-Att.	Mean	(0.6857)	**0.7319**	(0.6785)	**0.7232**	(0.7272)	0.7173	(0.7083)	**0.7450**
	Learnable	(0.6434)	0.6043	(0.6428)	0.6093	(0.7272)	0.4782	(0.7571)	0.6020
	Max	(0.6616)	0.5801	(0.6845)	0.5982	(0.6666)	0.6304	(0.6709)	0.5846
	Patient	(0.5460)	0.5719	(0.5476)	0.5736	(0.6060)	0.6956	(0.5555)	0.5709
	Therapist	(0.7664)	0.5710	(0.7619)	0.5691	(0.7878)	0.6304	(0.7727)	0.5759

Table 1: Overall results over the DAIC-WOZ dataset. UAR stands for Unweighted Average Recall.

codependency between questions and answers[2] is of the utmost importance for the learning process. As a consequence, we propose to design a multi-view architecture with inter-view attention (shown with orange color in Figure 1b) that transfer attention scores from one view to another, following the cross-attention paradigm [18]. Formally, attention scores $\mu_1, \mu_2, ..., \mu_M$ are shared between the two *view encoders*, and are the result of function $\mu_i = f(\alpha_i, \beta_i)$ that combines the individual view attention scores. We propose five different instantiations of $f(.,.)$.

Mean. $f(\alpha_i, \beta_i) = (\alpha_i + \beta_i)/2, 1 \leq i \leq M.$

Max. $f(\alpha_i, \beta_i) = max(\alpha_i, \beta_i), 1 \leq i \leq M.$

Patient. Focusing on the patient side.$f(\alpha_i, \beta_i) = \alpha_i, 1 \leq i \leq M.$

Therapist. Focusing on the therapist side.$f(\alpha_i, \beta_i) = \beta_i, 1 \leq i \leq M.$

Learnable. $f(.,.)$ is defined as self-attention acting on inputs $h_i = (p_i \oplus t_i), 1 \leq i \leq M.$

3 Analysis of the Results

Experiments were conducted on the DAIC-WOZ dataset [6] that contains interviews between patients and a virtual therapist as a wizard-of-oz[3]. The best model is chosen based on macro F1 over the development set to evaluate performance on the test set. Baseline models are trained with different input configurations (patient only, therapist only and patient plus therapist sentences) based on our implementation of hierarchical models for fair comparison[4]. Table 1 gives the detailed results and figures show that multi-view architectures provide a better way of combining inputs from patient-therapist interviews. In particular, the multi-view model with inter-view attention coupled with the mean function (*MV-Inter-Attention Mean*) evidences best performing results for 3 out of 4 evaluation metrics. Namely, improvements of 13.84% on macro F1 score, 13.69% on Unweighted Average Recall, and 13.15% on macro Precision are obtained compared to the best baselines.

We verify the existence of relevant information in therapist questions from the results obtained by the *Baseline Therapist* model, where only questions, as a sequence of sentences, are taken into account. Additionally, comparing results for different baseline models, we can argue that combining questions and answers as a sequence of sentences does not provide significant improvements over using just the patient's answers as input (*Baseline Patient+Therapist* vs. *Baseline Patient*). We believe that the lack of structural information within this input configuration plays an important role in restricting the learning ability of the baseline models, which is dealt with by the multi-view architecture.

From the results obtained by the multi-view strategies with intra-view attention (*MV-Intra-Attention*) compared to the ones with baseline hierarchical models, we can assess that multi-view architectures are a better alternative to process question-answer based interviews. Indeed, all multi-view architectures with intra-view attention provide significant performance improvements over the *Baseline Patient+Therapist* model for 4 out of 4 evaluation metrics, highlighting the significance of retaining

[2] Note also that a question that might not seem to be important, but for which the answer is meaningful, should definitely be highlighted by the learning model.

[3] Details of the dataset are given in appendix A.2

[4] Learning setups are given in appendix A.3

Architectures	Modality	macro F1		UAR	
		(Dev)	Test	(Dev)	Test
Raw Audio [1]	A	(0.66)	-	-	-
SVM:m-M&S [2]	T+V+A	(0.96)	0.67	-	-
HCAG [12]	T+A	(0.92)	-	(0.92)	-
HCAN [10]	T	(0.51)	0.63	(0.54)	0.66
HLGAN [10]	T	(0.60)	0.35	(0.60)	0.33
HAN [19]	T	(0.46)	0.62	(0.48)	0.63
HAN+L [19]	T	(0.62)	0.70	(0.63)	0.70
HCAG+T [12]	T	(0.77)	-	(0.82)	-
MV-IA-Mean	T	(0.69)	**0.73**	(0.68)	**0.72**

Table 2: SOTA results on DAIC-WOZ. T, V and A stand for Text, Visual and Audio modalities.

structural information of a dialogue. In particular, multi-view architectures utilize the interview semantic structure to limit the number of inter-sentence interactions learned by the model, thus reducing the amount of noisy interactions and allowing the model to focus on relevant information.

Further results support our argument of codependence between questions and answers, with the *MV-Inter-Attention Mean* model outperforming all other architectures including current state-of-the-art HAN+L model [19] (see Table 2). Nevertheless, this improvement does not stand for all cross-attention functions. Indeed, we observe that results obtained with non-balanced attention functions (i.e. Patient, Therapist, Max) are lower compared to (1) the balanced architectures (i.e. Mean, Learnable), and (2) all other configurations (i.e. Baselines, Multi-view with intra-view attention). Within non-balanced functions, attention scores are transferred from one view to the other one by making the hypothesis that only one of the two views drives the learning process. As such, these models represent the extreme case of cross-attention, where questions (resp. answers) importance is directly impacted by answers (resp. questions) importance, while neglecting their own attention score. Results prove that both views, questions and answers, play a role in defining their importance, and selecting either one as the sole criteria for importance can be counterproductive. Also, we expected the *MV-Inter-Attention Learnable* model to perform on par with the *MV-Inter-Attention Mean* architecture if not better. The small size of the dataset restricted its learning ability.

From Table 2, we show that our best performing model provides new state-of-the-art results over the DAIC-WOZ, successfully outperforming recent initiatives with comparable setups (HAN[19], HCAN[10]) as well as those relying on external knowledge (HAN+L[19]) or different modalities (SVM:m-M&S[2]). Note that the reported results are taken directly from the original papers, and that some related work surprisingly do not evidence results over the test split, such as HCAG and HCAG+T [12], although they highly perform on the development set.

Finally, experiments have been carried out with different dense input representations. Within the context of the DAIC-WOZ, best results so far [19, 12] have been achieved with hierarchical structures built on top of GloVe embeddings [14]. However, recent advances in long text encoding [5, 8] have focused on initializing deep hierarchical architectures with contextualized embeddings (BERT [4]). Thus, we implemented the multi-view architecture with BERT and GloVE. BERT-based models proved to be unstable at learning and could not generalize due to the small size of the dataset[5]. As a consequence, all results from Table 1 are given for GloVe.

4 Conclusions and Future Work

In this paper, we propose a multi-view architecture for automated depression estimation that treats patient-therapist interviews as a combination of two views (therapist questions and patient answers). The underlying idea it to take advantage of the information conveyed by the therapist when asking his questions, and not only the message produced by the patient. In particular, the presented multi-view approach allows to handle discourse symmetry as well as discourse structure, thus outperforming the simple encoding of the input data as a sequence of sentences. Results on the DAIC-WOZ show that the multi-view architecture steadily outperforms comparable baselines and evidences new state-of-the-art results. Based on the insightful recent research of Xezonaki et al. [19], we plan to further improve our results by incorporating external knowledge from different medical resources, such as lexicon or psychiatrist manual annotation.

[5] Details of the BERT-based implementation are given in appendix A.4

References

[1] A. Bailey and M. D. Plumbley. Gender bias in depression detection using audio features. In *29th European Signal Processing Conference (EUSIPCO)*, pages 596–600, 2021.

[2] Z. Dai, H. Zhou, Q. Ba, Y. Zhou, L. Wang, and G. Li. Improving depression prediction using a novel feature selection algorithm coupled with context-aware analysis. *Journal of Affective Disorders*, 295:1040–1048, 2021.

[3] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, et al. Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *International Conference on Autonomous Agents and Multi-agent Systems (AAMAS)*, pages 1061–1068, 2014.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, page 4171–4186, 2019.

[5] S. Gao, M. Alawad, M. Young, J. Gounley, N. Schaefferkoetter, H.-J. Yoon, X.-C. Wu, E. Durbin, J. Doherty, A. Stroup, L. Coyle, and G. Tourassi. Limitations of transformers on clinical text classification. *IEEE Journal of Biomedical and Health Informatics*, 02 2021.

[6] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, et al. The distress analysis interview corpus of human and computer interviews. In *9th International Conference on Language Resources and Evaluation (LREC)*, pages 3123–3128, 2014.

[7] S. Hong, A. G. Cohn, and D. C. Hogg. Using graph representation learning with schema encoders to measure the severity of depressive symptoms. In *10th International Conference on Learning Representations (ICLR)*, 2022.

[8] J. Kong, J. Wang, and X. Zhang. Hierarchical bert with an adaptive fine-tuning strategy for document classification. *Knowledge-Based Systems*, 238:107872, 2022.

[9] K. Kroenke. Enhancing the clinical utility of depression screening. *Canadian Medical Aassociation Journal*, 184(3):281–282, 2012.

[10] A. Mallol-Ragolta, Z. Zhao, L. Stappen, N. Cummins, and B. W. Schuller. A hierarchical attention network-based approach for depression detection from transcribed clinical interviews. In *Interspeech (INTERSPEECH)*, pages 221–225. ISCA, 2019.

[11] A. K. Mohankumar, P. Nema, S. Narasimhan, M. M. Khapra, B. V. Srinivasan, and B. Ravindran. Towards transparent and explainable attention models. In *58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4206–4216, July 2020.

[12] M. Niu, K. Chen, Q. Chen, and L. Yang. Hcag: A hierarchical context-aware graph attention model for depression detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4235–4239, 2021.

[13] S. A. Oureshi, G. Dias, S. Saha, and M. Hasanuzzaman. Gender-aware estimation of depression severity level in a multimodal setting. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.

[14] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[15] S. A. Qureshi, G. Dias, M. Hasanuzzaman, and S. Saha. Improving depression level estimation by concurrently learning emotion intensity. *IEEE Computational Intelligence Magazine*, 15(3):47–59, 2020.

[16] S. A. Qureshi, S. Saha, M. Hasanuzzaman, and G. Dias. Multitask representation learning for multimodal estimation of depression level. *IEEE Intelligent Systems*, 34(5):45–52, 2019.

[17] A. Ray, S. Kumar, R. Reddy, P. Mukherjee, and R. Garg. Multi-level attention network using text, audio and video for depression prediction. In *9th International on Audio/Visual Emotion Challenge and Workshop (AVEC)*, page 81–88, 2019.

[18] E. Sood, S. Tannert, P. Müller, and A. Bulling. Improving natural language processing tasks with human gaze-guided neural attention. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[19] D. Xezonaki, G. Paraskevopoulos, A. Potamianos, and S. Narayanan. Affective conditioning on hierarchical attention networks applied to depression detection from transcribed clinical interviews. In *Interspeech (INTERSPEECH)*, pages 4556–4560, 2020.

[20] Z. Şimşir, H. Koç, T. Seki, and M. D. Griffiths. The relationship between fear of covid-19 and mental health problems: A meta-analysis. *Death Studies*, 46(3):515–523, 2022.

A Appendix

A.1 Hierarchical Architectures

Hierarchical models treat a patient-therapist interview as a hierarchy of intermediate representations [10, 19], and are the current state-of-the-art for depression level estimation. In particular, we use hierarchical models defined by Xezonaki et al.[19] as our baseline with two main differences: (1) we define a non-RNN implementation of hierarchical models based on the findings of [11], who show the limits of attention mechanisms over RNN encodings; (2) we do not include context vectors in attention since lexicon-based external knowledge is not used in our work. Figure (1a) gives an overview of our hierarchical implementation, where w_{ji} represents the embedding of the j^{th} word of the i^{th} sentence, W_i represents the word encoding sequence for the i^{th} sentence, S_i is the learned representation of the i^{th} sentence, and r is the transcript level representation of the textual input. *word attention* and *sentence encoder* networks are defined as self-attention networks. Formally, let $[h_1, h_2, ..., h_N]$ be the input of the attention model. The learned representation rep is defined in Equation 1, where $g(.)$ is a learnable mapping function, and γ_i is the attention score of the i^{th} input. Note that for *word attention*, $[w_{ji}, \forall j]$ acts as the input of the self-attention mechanism giving rise to S_i, while for *sentence encoder*, $[S_i, \forall i]$ is the input sequence.

$$\alpha_i = g(h_i), \gamma_i = \frac{e^{\alpha_i}}{\sum e^{\alpha_i}}, rep = \sum \gamma_i \cdot h_i \tag{1}$$

A.2 DAIC-WOZ Dataset

The DAIC-WOZ dataset is part of a larger corpus, the Distress Analysis Interview Corpus (DAIC) [6]. The dataset contains clinical interviews aimed towards psychological evaluation of participants for detecting conditions such as anxiety, depression and post-traumatic stress disorder. These interviews were collected with the goal of developing a computer agent that interviews participants to identify verbal and non-verbal signs for mental illness [3]. This part of the dataset contains the Wizard-of-Oz interviews, conducted by virtual interviewer Ellie, controlled by a human interviewer in another room. These interviews have been transcribed and annotated for a variety of verbal and non-verbal features. Along with the transcripts, the dataset also contains corresponding visual and audio features extracted from the interview recordings. Depression severity was accessed based on PHQ-8 depression scale, and score of 10 is used as threshold to differentiate between depressed and non-depressed participants. The dataset is divided into training, development and test sets containing 107, 35 and 47 interviews respectively. The dataset is biased towards lower PHQ-8 scores (Figure 2) with almost 70% data points belonging to negative class in case of binary classification and only 6 instances with severe depression (PHQ-8 score > 17).

A.3 Implementation Details

We use pre-trained GloVe embeddings (300D) for word encodings [14][6]. Adam optimizer is utilized with a learning rate of $5 * 10^{-4}$ and the binary cross-entropy (BCELoss) is the final loss function. A dropout rate of 0.4 is applied. All implementations are done using PyTorch. A more detailed explanation of our architectures is given below.

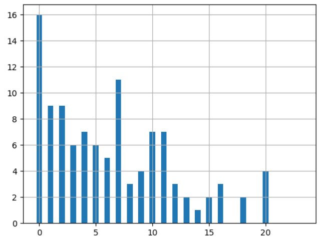
A.3.1 Baseline Models

To have a fair comparison, we implement baseline models based on our definition of hierarchical models (cf. §A.1) with three different input configurations.

Baseline Patient. Only the answers given by the patient are taken as input for the model, and therapist questions are ignored, similarly to [10].

Baseline Therapist. Only the questions asked by the therapist are taken as input for the model, ignoring the answers.

[6] https://github.com/stanfordnlp/GloVe



(a) Distribution of PHQ-8 scores.

Depression severity		Data split	
	Train	Validation	Test
No symptoms [0..5[	47	17	22
Mild [5..10[	29	6	11
Moderate [10..15[	20	5	5
Moderately severe [15..20[	7	6	7
Severe [20..24]	4	1	2
Total	107	35	47

(b) Dataset splits.

Figure 2: Details of the DAIC-WOZ dataset.

Baseline Patient+Therapist. Questions and answers are combined as a sequence of sentences neglecting their type as question or answer, similarly to [19].

A.3.2 Multi-view Strategies with Intra-view Attention

Within this category, we define three different architectures as part of an ablation study with self-attention layers as adjustment variables.

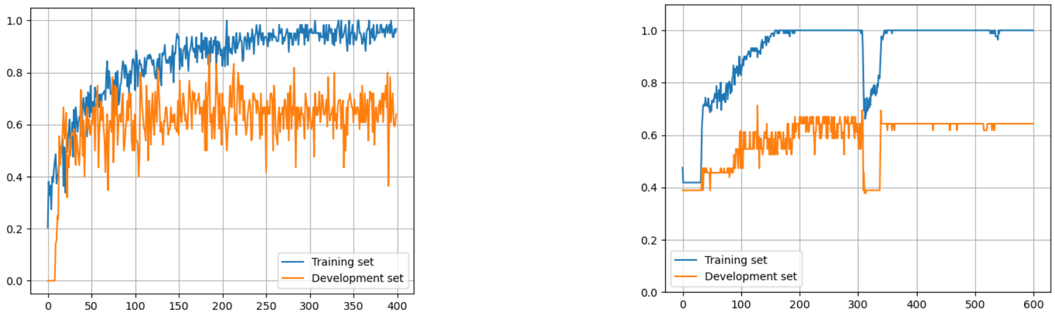
View-Global Attention. This architecture includes two view-level attention layers (*view encoders*) and a global-level attention layer (*global encoder*). Both view-level attention layers are defined using self-attention (cf. Equation 1), with corresponding sentence encodings acting as inputs. These are sentence attention layers. The global-level attention layer is also defined as a self-attention model aiming at fusing transcript-level view representations p and t. This layer can be seen as an aggregator of all the information contained in a transcript, i.e. questions and answers.

Global Attention. In this configuration, *view encoders* are replaced by a simple averaging operation instead of a self-attention layer[7], while the *global encoder* remains the same as in the View-Global Attention model.

View Attention. In this model, the *global encoder* is replaced by a simple concatenation of the patient representation p and the therapist representation t[8], while the *view encoders* remain the same as in the View-Global Attention model.

A.4 BERT-based Text Embedding

All our results are based on non-contextualized embeddings (GloVe [14]) despite the availability of contextualized embeddings such as BERT [4]. In our experiments, we found the training to be highly irregular when using BERT encodings as illustrated in Figure 3. The plots highlight an unstable learning process, thus making the predictions of the model unreliable. We believe that given the small size of the training set, the model favours simpler architectures (i.e. GloVe with a basic attention mechanism) over more complex ones (e.g. BERT with a transformer-based attention mechanism).



(a) BERT based inputs.

(b) GloVe based inputs.

Figure 3: Plots of the learning curves for the model *MV-Intra-Attention View-Global Attention* trained with BERT and GloVe input encodings.

[7] There is no attention information at sentence view level.

[8] There is no attention information at transcript level.