Online attention for interpretable conflict estimation in political debates

Anonymous FG 2018 submission
Paper ID ****

Abstract— **Conflict arises naturally in dyadic interactions when involved individuals act on incompatible goals, interests, or actions. In this paper, the problem of conflict intensity estimation from audiovisual recordings is addressed. To this end, we propose an online attention-based neural network in order to learn a mapping from a sequence of audiovisual features to time-series describing conflict intensity. The proposed method is evaluated by conducting experiments in conflict intensity estimation by employing the CONFER dataset. Experimental results indicate the superiority of the proposed model compared to the state of the art. Furthermore, we demonstrate that by incorporating sparsity in the model, the origin of conflict can be traced back to specific key frames facilitating the interpretation of conflict escalation.**

I. Introduction

Humans are predominantly social beings, expressing social behaviours such as agreement, conflict, politeness, empathy etc as temporal patterns of non-verbal behavioural cues [1] [2]. The importance of developing tools for an automatic analysis and prediction of human social behaviours from audiovisual recordings is profound, facilitating both basic research in cognitive and social sciences and drastically improving the state of the art in human-computer interaction.

In this paper, we focus on estimating conflict escalation and resolution in dyadic interactions from audiovisual recordings. Conflict is used to label a range of human experiences, from disagreement to stress and anger, occurring when involved individuals act on incompatible goals, interests, or actions. Various research studies in human sciences argue that a disagreement does not have to result in a conflict; conflict describes a high level of disagreement, or escalation of disagreement, where at least one of the involved interlocutors feels emotionally offended.

Prior work on automatic (dis)agreement and conflict detection and estimation is rather limited. Concretely, statistical model of verbal and acoustic features have been applied for disagreement detection [3][4][5], while in [6] [7] the task is addressed by employing a sequential discriminative model. Kim et al. [8] [9] employ audio features for conflict detection while methods for estimation of continuous-valued conflict intensity have been proposed in [10] [11] [8]. However, the aforementioned methods ignore or oversimplify the temporal dimension which is of utmost importance to the problems of conflict and (dis)agreement estimation.

In this work, motivated by the success of attention models in tasks such as neural machine translation [12], [13], caption generation [14], speech recognition [15], and lip reading [16], we investigate how attention models can be employed in the task of continuous social behaviour estimation and in particular in conflict estimation. To this end, we propose an online attention-based neural network that is learned end-to-end from facial and vocal features. We evaluate the proposed model in conflict estimation in political debates by conducting experiments on the CONFER dataset [17]. Significant improvements over the state of the art are reported. This is attributed to the fact that distinct from previous work on conflict estimation, the proposed attention-based model focuses heavily on the temporal aspect of conflict while being able to handle noisy data inherent in naturalistic real world conditions. It is also worth mentioning that the proposed model differs from previous approaches to modeling attention in that it is entirely online, suitable for interactions of arbitrary length with no increase in computational cost. Finally, we illustrate why attention is especially interesting when applied to the task under study by extracting useful insights from a learned model. To aid in this endeavour we propose a novel method to induce sparsity in the learned attention. The resulting learned attention is hard and sparse, allowing us to pinpoint key frames that indicate an arising conflict.

II. Model Architecture

We make use of the encoder-decoder architecture. It has the advantage of being entirely asymmetric; in general the decoder has access to all the encoded outputs and does not need to be of the same size as the inputs. As we work with videos and we are interested in the conflict for each frame, however, input length and output length will be the same.

A. Encoder

The video encoder operates on extracted features which we will discuss in section IV-B. We denote these image features as x_t. The encoder takes in such a sequence x_t, feeds these into a small fully-connected (FC) network that decreases in size, then passes these features through a Long Short-Term Memory (LSTM) layer [18] to produce an encoded series o^e of the same length as the input sequence, as follows:

$$\begin{aligned} f_t &= \mathrm{FC}(x_t) \\ h_t, o^e_t &= LSTM(f_t, h_{t-1}) \end{aligned} \tag{1}$$

The initial LSTM value h_1 is learned with the rest of the network.

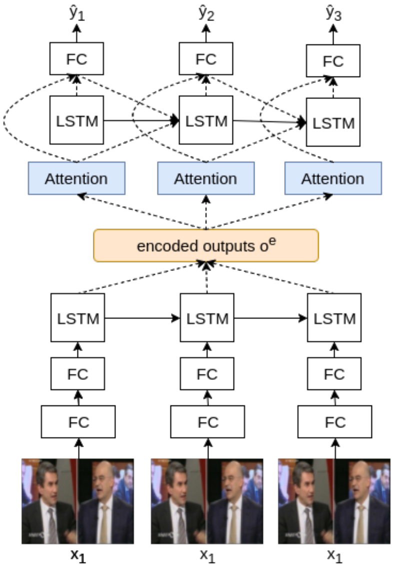


Fig. 1: At each time step, the decoder part generates an output $\hat{y}_i$. The attention mechanism is used to *attend*, i.e. to select, appropriate encoded outputs from the history.

B. Decoder

The decoder works asynchronously from the encoder and contains the attention mechanism:

$$\begin{aligned} h_t^d, o_t^d &= \mathrm{LSTM}(\hat{y}_{t-1}; c_{t-1}, h_{t-1}^d) \\ c_t &= \mathrm{Attention}(h_{t-1}^d, o^e) \\ \hat{y}_t &= \mathrm{FC}(c_t; o_t^e; o_t^d) \end{aligned} \tag{2}$$

At each time step t, the decoder LSTM produces h_t^d and o_t^d from last time step's context vector c_{t-1} and output $\hat{y}_{t-1}$. The context vector is produced by the attention mechanism. The output $\hat{y}_t$ is then computed from a fully-connected layer with a linear activation that takes the decoder recurrent network's output, the encoded output and the context vector. The rationale behind this is that attention is supposed to distill information from the rest of the sequence without relying on all of it.

C. Attention

The attention mechanism is inspired by work by Bahdanau, Cho, and Bengio [13] with one of the fundamental differences being that our work uses *local* instead of *global* attention. Generally speaking, an attention model works with a query and some memory to query from. Usually the query corresponds to the decoder hidden state h_t^d and the memory is simply the encoder's outputs o^e. The memory is where the *local* and *global* variants differ. Attention learns a weighting α over the memory as follows:

$$\begin{aligned} e_{ij} &= v^\top \tanh(W q_i + U m_j + b) \\ \alpha_{ij} &= \mathrm{softmax}(e_{ij}) \end{aligned} \tag{3}$$

with q_i the query at timestep i and m_j the memory at timestemp j. Weight matrices W, U and vectors v and b are learned. The weighting α is finally used to compute context vector c_i.

$$c_i = \sum_{j=L_0}^{L_0+L_w} \alpha_{ij} m_j \tag{4}$$

In the case of *global* attention, the memory would consist of the whole encoded sequence, i.e. $L_0 = 0$ and $L_w = L$, with L denoting the whole sequence length. In case of local attention a window is used, so L_0 and L_w, the window size, are set accordingly.

Windowed Attention: Since we deal with videos, global attention quickly becomes too computationally inefficient as it operates on the entire sequence. Furthermore, it requires the whole sequence to be available even before the first output is produced, as each output is dependent on the whole input sequence. To remedy this, we employ a form of windowed attention that can work entirely online. At the time the decoder produces output $\hat{y}_t$, the window of encoded outputs available to it ranges from $t - T + 1$ up to t. That is, the past T frames are available to the attention mechanism. This has the advantage of being entirely online, an absolute requirement for any real application, so predictions can be generated in real-time. Additionally, it works on sequences of arbitrary length as the computational cost stays fixed over time.

III. Training Strategy

A. Scheduled Sampling

The decoder recurrent network described in section II-B is conditioned on the previous network output so it can learn to model a coherent sequence. It is not uncommon in a scenario like this to use the ground truth values instead of predictions during training. The weakness to this approach is that the ground truth is not present during inference and the model has not learned to cope with small deviations in predictions. In practice, the deviations compound and outputs diverge dramatically, straying increasingly as the sequence progresses. This poses an increased risk to our model given that we work with long sequences, especially compared to previous work. To combat this, we employ the scheduled sampling method by Bengio et al. [19]. During training, the probability to sample from the previous prediction instead of from the previous truth changes gradually from 0 to 1 at which it stays for the rest of training. We have tried both a linear and inverse sigmoidal annealing scheme and found the more simple linear option straightforwardly effective.

B. Implementation details

The first encoder fully-connected layer contains a number of neurons equal to the input feature size whereas the second fully-connected layer contains 128 neurons. Both the encoder LSTM and decoder LSTM contain 128 cells. The final layer has a single output with a linear activation. All fully-connected layers are followed by batch normalization which we found greatly prevents overfitting of the network, and a Rectifier Linear Unit (ReLU) activation.

The model is trained with regards to the Intra-class Correlation Coefficient (ICC) [20], a correlation metric that measures ‘consistency’ or ‘agreement’. We use the common variant ICC(3,1). For training, we employ the gradient-based optimization algorithm ADAM [21] with a learning rate of 10^{-3}.

IV. Experiment

A. Dataset

We make use of the Conflict Escalation Resolution (CONFER) database [17], a set of recordings of televised political debates from Greek TV. It contains 73 episodes of dyadic interactions and 47 episodes of interactions among three subjects, spanning approximately 142 minutes with 54 subjects. Recordings are *real-world* and *in-the-wild*, i.e. they are not set in artificial conditions. Lighting tends to vary, poses are not always front-facing, abrupt hand gestures can appear and speech can overlap. The conflict annotation sequence is derived from 10 annotators by employing Canonical Correlation Analysis (CCA) to extract maximally correlated subspaces for the annotations and corresponding audiovisual feature sets. These conflict intensity values range between 0 and 1.

All our experiments follow the 5-fold cross-validation experimental protocol as proposed by [17] to represent a fair and representative view over the whole dataset. This means that for each fold the model is trained on 3 out of 5 parts of the dataset, validation is performed using a 4th and test metrics are calculated over a 5th. We report our test metrics averaged over all 5 folds.

B. Image Features

The CONFER database contains 68 tracked facial points per interactant. We transform these with an affine transformation based on 5 stable points: those corresponding to the corners of the eyes and the tip of the nose. Then we extract the following features:

1) Expr.+SIFT: We reuse the best-performing feature set reported by [17] to allow for a fair comparison. We only consider the Expression and SIFT feature sets in conjunction (and not separate), as [17] reports this makes up their best-performing visual feature set.

Expression: Principal Component Analysis is applied on the *Points* set by projecting the facial landmarks onto the subspace spanned by the ‘eigenshapes’ of a pre-trained Active Shape Model (ASM), following [17]. 18 coefficients are kept accounting for 95% of the total variance. Of these, the last 12 are kept that are deemed related to facial expression, i.e. face deformation.

SIFT: Appearance-based descriptors named *Scale-Invariant Feature Transform* (SIFT) features are derived over both video streams, as described in [17]. After CCA, keeping features that account for 95% of total variability, 75 features are kept for both faces combined.

2) Points: For each interactant we have 68 facial points. These facial points are zero-centered, normalized to unit variance and their coordinates kept as features.

3) VGG-Face: To study the efficacy of derived features, we also use deep-learned features that are learned separate from the task at hand. To this end, we use the architecture dubbed *VGG-Face* [22], pre-trained on a face recognition task. The features we keep are the outputs of its last max-pooling layer which leaves us with a feature vector of size 512 per face. As input to the network we use the raw image, aligned with the same transformation we use for facial points.

4) Audio: As [17] reported extensively on their results with audio features and their best visual and audio features combined, we briefly consider the same set of audio features. Audio features are extracted with the openSMILE feature extractor [23] to obtain 65 low-level descriptors (4 relating to energy, 55 spectral and 6 voicing-related). The audio features are sampled at 25 Hz so fusion with visual features is a simple concatenation.

C. Results

Table I reports results for our attention-enabled model on the CONFER dataset. Although our work focuses mainly on visual features we also report results for audio and audiovisual features to allow for a fair comparison with existing work. This means we only compare for the audiovisual features consisting of a fusion of *Expression*, *SIFT* and *Audio* features following the work of [17], even though we have found visual features that outperform *Expression+SIFT*.

Our model improves on the state of the art for each feature set with regards to ICC. It outperforms for *Expr.+SIFT* and audiovisual features with regards to Pearson correlation (COR) and is on par for audio features considered separately. Interestingly, while [17] obtained high ICC scores for audio and audiovisual only, we achieve consistently high ICC scores for each feature set. While we found ICC is a good correlation-based metric to train on, we do notice that Pearson correlation improves more slowly, making it a more critical and perhaps more apt test metric. This can be seen very clearly from the differences between visual and audiovisual results.

Our biggest gain in performance is on the audiovisual feature set. While we do well for both audio and video separate, audiovisual results far outstrip either with a correlation of 0.553. This indicates that both modalities carry information highly complementary to the other with regard to conflict estimation.

Visual Features: Table II contains presents results of the various visual features we have used with our attention model. We found that the best results are achieved using *Points* features, which can be considered less processed than the originally proposed *Expr.+SIFT* features. Using more

TABLE I: Results on the CONFER dataset.

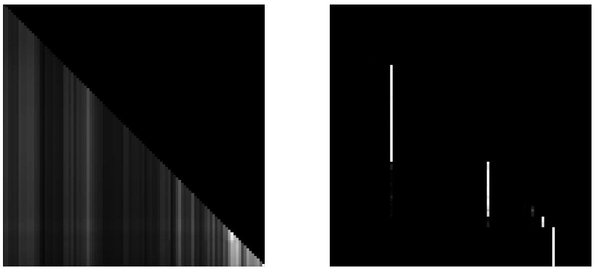
	Audio (A)		**Expr.+SIFT (V)**		**V + A**	
	COR	**ICC**	**COR**	**ICC**	**COR**	**ICC**
SVR [17]	0.233	*0.774*	*0.204*	0.174	*0.294*	0.781
BiLSTM [17]	0.232	0.178	0.126	*0.183*	0.178	0.160
CCRF [17]	**0.285**	0.160	0.026	-0.001	0.221	0.163
Attention	0.283	**0.886**	**0.303**	**0.895**	**0.553**	**0.928**

TABLE II: Visual feature comparison

Features	MSE	COR	ICC
E.+SIFT	0.101	0.303	0.895
VGG	0.105	0.323	0.895
Points	**0.085**	**0.430**	**0.931**

TABLE III: Window size exploration for *Points* features

T	MSE	COR	ICC
50	0.107	0.353	0.902
100	0.085	0.430	0.931
200	0.108	0.317	0.891



(a) Common *soft* attention

(b) Induced sparse attention

Fig. 2: Learned alignments for the same sequence both with and without induced sparsity. A white vertical line indicates that frame is attended to throughout time with the brightness corresponding to the alignment strength.

fundamental features allows the model to learn its own hypothesis instead of supplying it with interpreted features such as the *Expression* features.

Controversially, *VGG-Face* features do not outperform processed features. It is possible that the task of face recognition is too dissimilar to that of conflict estimation and that features learned while solving the former hold relatively little value for the latter.

Attention Window: Table III lists results on the CONFER dataset for different time windows for the *Points* feature set. Given that the video is sampled at 25 Hz, a time window of 50 steps corresponds to 2 seconds. We get our best results for a window of $T = 100$. While it is perhaps not surprising that this window performs better than the smallest, it is not so intuitive that the largest window would perform worst. While theoretically the latter has the most information available to it, we suspect it has a hard time learning how to deal with this abundance of information.

D. Interpretative Conflict Analysis

The attention mechanism discussed here allows the model to soft-align to past frames, assigning weights to each for each output to be predicted. To gain insights into what the model actually learned we can inspect the learned alignment weights α_{ij} for a given sequence. We found that the model has a tendency to focus strongly on an important frame the first time it is encountered, then to maintain some lingering alignment with it until it disappears from the window. Sometimes, no such crucial frames are encountered in which case a soft, weak alignment is applied over most of the history. An example of this behaviour can be seen in Fig. 2a which illustrates a soft alignment to most of the history window.

While having a model approximate the average annotator rating to the best of its capacity according to some quantitative metric is a worthwhile goal, it is not the only one

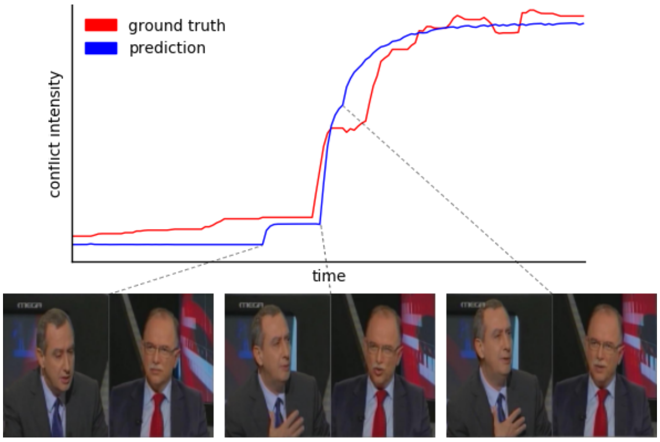


Fig. 3: Sparse attention trained exclusively on visual features. The indicated frames point to time steps where a new single input frame is attended to. The conflict can be seen arising as speakers start talking simultaneously and hand gestures come into view.

worth pursuing. Sometimes it is more interesting to only get the rough general tendencies of a conflict and trace these back to only a few key frames. This should be a more robust model, able to capture greater tendencies without falling for noise. To this end, we propose a regularization method that encourages sparsity of the learned alignments. We use the Hoyer sparsity measure [24], defined for a vector x of length n as $Sparsity(x) = \frac{\sqrt{n}-\frac{\|x\|_1}{\|x\|_2}}{\sqrt{n}-1}$. We then adjust the loss function to incorporate this sparsity measure applied to the alignments α_i. The new loss function then becomes $ICC(y,\hat{y}) + \beta * (1 - Sparsity(\alpha))$. Lower values of this loss function, for the same ICC term, correspond to an increased sparsity in α. The constant β represents the tradeoff between model accuracy in terms of ICC and the need for hard, sparse alignments that are more interpretable.

We found this approach allows us to learn hard alignments that would often focus on just one frame for multiple steps at a time, as illustrated in Fig. 2b, albeit at a penalty to performance. There is also a computational advantage to this that we did not exploit: instead of resorting to matrix multiplication, the context can be a simple selection from history. Fig. 3 illustrates the resulting behaviour. The lines indicate when a new frame is attended to. The resulting model tends to behave smoothly for the same attended frame, then changes behaviour when a new critical frame is encountered. We found that the attended frames were especially representative of the escalating conflict.

V. Summary and Conclusion

In this work we introduced the first attention-based model for conflict estimation. Additionally, the discussed attention mechanism takes a window-based, online approach to what heretofore has only been done in an unscalable, entire-sequence-to-sequence manner. Finally, we proposed a novel method to induce sparsity resulting in hard alignments. Compelling both computationally and for interpretability, this method allows us to trace arising conflict back to a few key frames, making it invaluable for conflict analysis.

References

[1] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schroeder, "Bridging the gap between social animal and unsocial machine: A survey of social signal processing," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 69–87, 2012.

[2] M. Pantic and A. Vinciarelli, *Social Signal Processing*. Berlin: Springer, 2014, pp. 84–93.

[3] M. Galley and K. McKeown, "Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies," in *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, 2004, p. 669.

[4] W. Wang, S. Yaman, K. Precoda, and C. Richey, "Automatic identification of speaker role and agreement/disagreement in broadcast conversation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2011, pp. 5556–5559.

[5] W. Wang, K. Precoda, C. Richey, G. Raymond, S. R. I. International, and M. Park, "Identifying agreement/disagreement in conversational speech: A cross-lingual study," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011, pp. 3093–3096.

[6] K. Bousmalis, L. P. Morency, and M. Pantic, "Modeling hidden dynamics of multimodal cues for spontaneous agreement and disagreement recognition," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2011, pp. 746–752.

[7] K. Bousmalis, M. Mehu, and M. Pantic, "Spotting agreement and disagreement: A survey of nonverbal audiovisual cues and tools," in *Affective Computing and Intelligent Interaction and Workshops*, IEEE, 2009, pp. 1–9.

[8] S. Kim, F. Valente, M. F. Member, and A. V. Member, "Predicting Continuous Conflict Perception with Bayesian Gaussian Processes," *IEEE Transactions on Affective Computing*, vol. 5, no. 2, pp. 1–14, 2014.

[9] S. Kim, F. Valente, and A. Vinciarelli, "Automatic detection of conflicts in spoken conversations: Ratings and analysis of broadcast political debates," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012, pp. 5089–5092.

[10] C. Georgakis, Y. Panagakis, and M. Pantic, "Dynamic Behavior Analysis via Structured Rank Minimization," *International Journal of Computer Vision*, pp. 1–25, 2017.

[11] Y. Panagakis, M. A. Nicolaou, S. Zafeiriou, and M. Pantic, "Robust correlated and individual component analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1665–1678, 2016.

[12] M.-T. Luong, H. Pham, and C. D. Manning, "Effective Approaches to Attention-based Neural Machine Translation," in *Proceeding of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1412–1421.

[13] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation By Jointly Learning To Align and Translate," in *International Conference on Learning Representations*, 2015, pp. 1–15.

[14] K. Xu, J. Ba, R. Kiros, K. Cho, and A. Courville, "Show, attend and tell: Neural image caption generation with visual attention," *International Conference on Machine Learning*, 2015.

[15] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, Attend and Spell," *arXiv preprint arXiv:1508.01211*, pp. 1–16, 2015.

[16] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip Reading Sentences in the Wild," *arXiv preprint arXiv:1611.05358*, 2016.

[17] C. Georgakis, Y. Panagakis, S. Zafeiriou, and M. Pantic, "The Conflict Escalation Resolution (CONFER) Database," *Image and Vision Computing*, 2016.

[18] S. Hochreiter and J. Urgen Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[19] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, "Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks," pp. 1–9, 2015.

[20] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: uses in assessing rater reliability.," *Psychological bulletin*, vol. 86, no. 2, p. 420, 1979.

[21] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv preprint arXiv:1412.6980*, pp. 1–15, 2014.

[22] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep Face Recognition," *Procedings of the British Machine Vision Conference 2015*, no. Section 3, pp. 41.1–41.12, 2015.

[23] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, ACM, 2010, pp. 1459–1462.

[24] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of machine learning research*, vol. 5, pp. 1457–1469, 2004.