SEC2SEC CO-ATTENTION TRANSFORMER FOR VIDEO-BASED APPARENT AFFECTIVE PREDICTION

Mingwei Sun and Kunpeng Zhang
University of Maryland

ABSTRACT

Video-based apparent affect detection plays a crucial role in video understanding, as it encompasses various elements such as vision, audio, audio-visual interactions, and spatiotemporal information, which are essential for accurate video predictions. However, existing approaches often focus on extracting only a subset of these elements, resulting in the limited predictive capacity of their models. To address this limitation, we propose a novel LSTM-based network augmented with a Transformer co-attention mechanism for predicting apparent affect in videos. We demonstrate that our proposed Sec2Sec Co-attention Transformer surpasses multiple state-of-the-art methods in predicting apparent affect on two widely used datasets: LIRIS-ACCEDE and First Impressions. Notably, our model offers interpretability, allowing us to examine the contributions of different time points to the overall prediction. The implementation is available at: https://github.com/nestor-sun/sec2sec.

Index Terms— Co-attention, Transformer, Multimodal Learning, Affective Computing, Video Understanding

1. INTRODUCTION

Video plays a crucial role in the field of human-computer interaction, and understanding human responses to video content is essential for designing and optimizing these systems. A key aspect of human interpretation is the perceived affect elicited by video content, as it can influence user engagement [1], user trust in branded content [2], and user behavior [3] within these systems.

A video comprises two primary components: vision and audio. Each of these components has distinct influences on predictions. Furthermore, both visual and auditory elements can jointly impact predictions. Specifically, a good alignment between a visual component with its corresponding audio within the same time frame (e.g., within a second) creates a synergy that enhances viewers' perception of video. In the context of emotion perception, distinct combinations of audio and visual features may elicit different emotional states. For instance, a chilling image coupled with a slow audio tempo is likely to evoke a negative valence. Another crucial aspect that influences viewers' perception of a video is its sequential composition. Research indicates that people tend to recall the most recent information [4], suggesting that later clips may carry higher weights when estimating affective responses.

In recent years, there has been significant attention given to video-based affective prediction. Studies have leveraged deep learning techniques to predict affect from video. For example, one study developed a convolutional neural network (CNN) to predict emotions using only images from videos [5]. Previous research has emphasized the importance of various types of information in video-based affective prediction, including audio and visual features, temporal information, and audio-visual interactions. For instance, different attention mechanisms were explored, confirming the significance of capturing audio-visual interactions [6]. However, despite these advances, affective prediction remains a challenging research topic, as existing methods have not yet fully captured all the relevant information present in videos.

In this paper, we tackle these challenges by developing a Transformer-based second-to-second (Sec2Sec) co-attention model to predict the video-perceived affective states. Specifically, we first implement a Transformer-based co-attention network extended from the work proposed by [7] to understand the interactions between audio and vision. We further combine a Long Short-Term Memory (LSTM) module with such a co-attention network to capture the temporal information of videos at the second level. To do so, we first split each video into one-second video clips. We then feed each one-second clip into our designed co-attention network. The output of each video clip from the co-attention network is fed into an LSTM network sequentially. Lastly, we add a fully-connected feed forward (FC) for affective prediction.

Our contributions are three-fold. First, we propose a novel Sec2Sec Co-attention Transformer for the affective prediction to capture all the necessary types of information. Second, we evaluate the performance of our proposed Sec2Sec Co-attention Transformer on two real-world datasets. Our extensive experimental results demonstrate that our approach outperforms several competitive baselines in affective prediction tasks. Additionally, we conduct interpretability analyses to assess the contributions of individual one-second video segments to the final predictions.

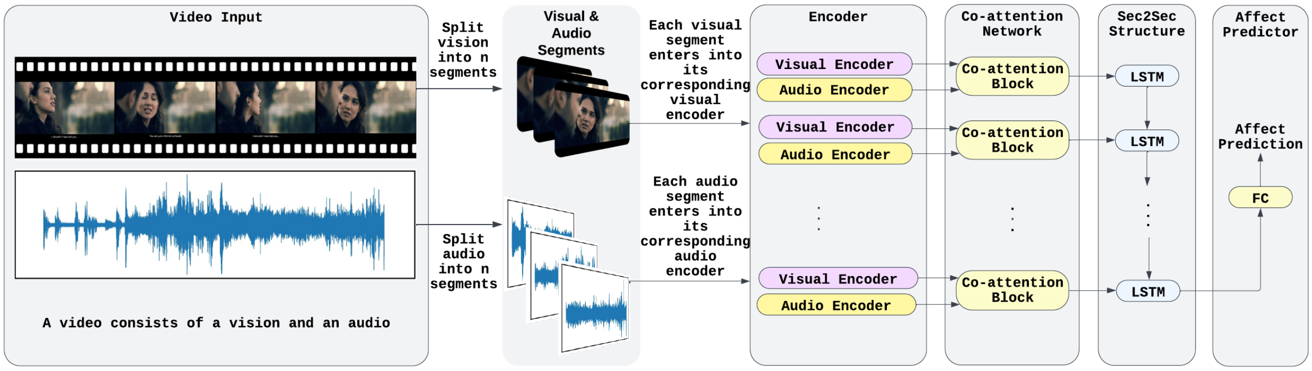

Video Input
A video consists of a vision and an audio
Split vision into n segments
Split audio into n segments
Visual & Audio Segments
Each visual segment enters into its corresponding visual encoder
Each audio segment enters into its corresponding audio encoder
Encoder
Visual Encoder
Audio Encoder
Co-attention Network
Co-attention Block
Sec2Sec Structure
LSTM
Affect Predictor
Affect Prediction
FC

Fig. 1: An overview of our proposed model: a Sec2Sec Co-attention Transformer.

2. RELATED WORK

Our work is related to two streams of literature: audio-video representation learning and applications of Transformers.

Audio-Visual Cross-Modal Learning. The mainstream of audio-visual representation learning research is to predict the synchronization or correspondence of audio and visual streams in videos. Arandjelovic and Zisserman trained an audio-visual cross-modal network from scratch to predict video correspondence [8]. Alwassel et al. used one clustered modality as a supervisory signal for another modality, and predicted correspondence between two modalities [9]. Cheng et al. further developed three self-supervised co-attention-based networks to discriminate visual events related to audio events [7]. In addition, Kuhnke et al. proposed a two-stream aural-visual model to predict facial expressions in videos [10].

Transformers in Computer Vision. Recently, Transformers have been increasingly applied to computer vision (CV) tasks as an alternative to CNN. ViT applies a Transformer model to linearly projected sequences of image patches to classify full images [11]. Swin Transformer improves ViT by introducing a hierarchical Transformer architecture and a shifted window scheme [12]. In order to classify video tasks, ViViT extends ViT by proposing two methods for embedding video samples: uniform frames sampling and tubelet embedding [13]. Video Swin Transformer further extends Swin Transformer by introducing a 3D-shifted window-based multi-head self-attention module and a locality inductive bias to the self-attention module [14]. However, all these video-based analyses do not separate vision and audio and explicitly learn the joint effect on subsequent tasks, which is our focus in this study.

3. METHOD: SEC2SEC CO-ATTENTION TRANSFORMER

In this section, we introduce the proposed model. As depicted in Figure 1, the proposed model consists of five steps: **Video Segmentation**: We first split each video into n segments. **Encoder network**: The encoder network comprises a visual encoder and an audio encoder that extract visual and audio features using pre-trained ResNet networks. [15]. **Co-attention block**: The co-attention block leverages Transformer [16] to model the interactions between visual and audio features. **Sec2Sec structure**: It captures the temporal information via an LSTM network. **Predictor**: The output from the LSTM network is fed into an FC network to make apparent predictions.

Visual encoder. To extract visual features, we first sample m frames per segment. Each frame is represented by a color image with the Red-Green-Blue (RGB) channels. Like prior studies, we apply pre-processing to images, such as resizing, center cropping, and normalization. Thus, each visual part is represented in a 4-dimensional space (i.e., 3-dimensional RGB plus m frames), which is fed into a pre-trained R(2+1)D ResNet model [17].

Audio Encoder. To extract audio features, we compute 2-dimensional Mel-Frequency Cepstral Coefficients (MFCCs) [18], MFCC's first-order (delta coefficients), and second-order frame-to-frame time derivatives (delta-delta coefficients) from each audio clip. Therefore, the audio feature can be represented by combining three-channel MFCC features in which each channel is one type of coefficients. The three-channel MFCC features, treated as a special type of ”image”, are fed into a pre-trained ResNet [15].

Co-attention block. The extracted visual and audio features for each segment enter into two symmetric co-attention sub-blocks, visual and audio sub-blocks, to learn guided audio and visual representations. Each sub-block is built by combining a standard multi-head self-attention module with a multi-head co-attention module. A normalization layer (Norm), a residual connection and an FC network are applied.

In the visual sub-block, the extracted visual embedding from the visual encoder is first fed into a multi-head self-attention module to get the intermediate visual representation, I_v, embedding important visual information. Similarly, we can get the intermediate audio representation, I_a, in the audio sub-block. Specifically, I_v and I_a can be computed as follows:

$$\begin{aligned} I_{iv} &= FC(Norm(MultiHead(z_{iv}, z_{iv}, z_{iv})) + z_{iv}) \\ I_{ia} &= FC(Norm(MultiHead(z_{ia}, z_{ia}, z_{ia})) + z_{ia}) \end{aligned} \tag{1}$$

where z_{iv} and z_{ia} denote the output features from the visual encoder and the audio encoder for segment i, respectively.

Next, in the visual sub-block, I_{ia} as key and value and I_{iv} as query are passed into the multi-head co-attention module. In this way, we can enforce the visual sub-block to focus on the information related to audio. Similarly, in the audio sub-block, we feed I_{iv} as key and value and I_{ia} as query into the multi-head co-attention layer. Thus, the audio sub-block tends to focus on the information corresponding to vision. Hence, the final output features of vision and audio, F_{iv} and F_{ia}, can be computed as:

$$\begin{aligned} F_{iv} &= FC(Norm(MultiHead(I_{iv}, I_{ia}, I_{ia})) + I_{iv}) \\ F_{ia} &= FC(Norm(MultiHead(I_{ia}, I_{iv}, I_{iv})) + I_{ia}) \end{aligned} \tag{2}$$

Consequently, two sub-blocks focus on important information about themselves, as well as their relationships. Using such a mechanism, we capture the interaction between visual and audio components. Finally, we combine the guided visual and audio representation by applying an FC layer, computed as $F_i = FC(concat(F_{iv}, F_{ia}))$, which is the joint representation of vision and audio for each segment i.

Sec2Sec Structure. To capture the temporal information in the video clip sequence, we feed the joint representation of each segment, F_i, from the co-attention block to an LSTM network. The LSTM network is defined as follows:

$$\begin{aligned} u_i &= \sigma(W_{Fu}F_i + W_{hu}h_{i-1} + b_u) \\ f_i &= \sigma(W_{Ff}F_i + W_{hf}h_{i-1} + b_f) \\ o_i &= \sigma(W_{Fo}F_i + W_{ho}h_{i-1} + b_o) \\ \tilde{c}_i &= tanh(W_{Fc}F_i + W_{hc}h_{i-1} + b_c) \\ c_i &= f_i \odot c_{i-1} + u_i \odot \tilde{c}_i \\ h_i &= o_i \odot tanh(c_i) \end{aligned} \tag{3}$$

where $\sigma(\cdot)$ is an activation function. $\odot$ denotes the Hadamard product. W and b are weights and biases to be learned. h_i denotes the hidden state at step i. u_i, f_i, o_i and c_i denote the update gate, forget gate, output gate and cell gate, respectively.

Predictor. We apply an FC along with a sigmoid function to the output from the LSTM network at the last step to make affective predictions.

4. EXPERIMENTS

4.1. Dataset and Evaluation

To evaluate the effectiveness of our proposed model, we utilize two publicly available datasets: LIRIS-ACCEDE [19] and First Impressions [20]. LIRIS-ACCEDE contains $9,800$ videos extracted from 160 films. In this paper, we adopt a binary classification approach based on the existing literature [21]. For evaluation, we adopt two standard metrics for emotion classification tasks (valence and arousal), accuracy and $F1$ score. First Impressions is widely utilized in the field of apparent personality analysis consisting of 10,000 labeled clips extracted from over 3,000 YouTube videos. Since the personality traits are continuous between 0 and 1, we use the mean accuracy as the evaluation metric, computed as, $MeanAccuracy_t = \frac{1}{N}\sum_{i=1}^{N}(1 - |y_{it} - \hat{y}_{it}|)$, where y_{it} is the ground truth value for the ith video sample and tth personality trait, and $\hat{y}_{it}$ is the predicted value for the same video sample and trait. N is the total number of predicted videos.

4.2. Implementation Details

We train all models on an NVIDIA GeForce 3090 24GB GPU with 250 epochs. We set up an early stopping mechanism, where the training stops if the validation loss increases for 5 consecutive epochs. We use the grid search strategy to find a relatively optimal set of hyperparameters. In each experiment, we use the model with the best validation accuracy to report results on the holdout testing set.

4.3. Baselines

We evaluate the performance of our proposed model (called **Sec2Sec SA-CA**) with several state-of-the-art methods. **CMA** [7], **AVM** [10], **ViT** [11], **ViViT** [13], **ViT-ViViT** [11, 13]: We implement a bi-modal (audio and vision) network by combining ViT and ViViT to extract audio and visual features, respectively. We also add a co-attention network as another baseline and 3 variants of our model for comparison to understand the role of each design in our model (e.g., uni vs. bi-modal, co-attention): **Co-attention** (bi-modal), **Sec2Sec Vision** (uni-modal), **Sec2Sec Audio** (uni-modal) and **Sec2Sec SA-SA** (bi-modal with self-attention).

4.4. Results

Overall performance. Table 1 presents the experimental results for arousal and valence. Our proposed Sec2Sec models achieve the best performance on two evaluation metrics for arousal prediction. They surpass three bi-modal baselines (audio and vision) methods and the co-attention approach in terms of accuracy, F1 score and efficiency, demonstrating the benefit of incorporating LSTM (Sec2Sec) into the video understanding framework. It also outperforms Sec2Sec Audio and Sec2Sec Vision, indicating that using both audio and visual components is more effective than using either modality alone. Moreover, Sec2Sec SA-SA and Sec2Sec SA-CA obtain comparable results, suggesting that the interactions between audio and visual features is not essential for predicting arousal. It is noteworthy that ViT-ViViT performs worse than ViT and ViViT, indicating that a single FC layer fails to adequately capture the interaction between audio embeddings and visual embeddings. We have similar observations for valence prediction, in terms of performance comparison with baselines. However, Sec2Sec SA-CA outperforms Sec2Sec

Table 1: Performance comparison of our model with baselines for arousal and valence prediction.

			Arousal			Valence		
	Method	Input	Accuracy	$F1$ Score	Avg Training Time Per Epoch (min)	Accuracy	$F1$ Score	Avg Training Time Per Epoch (min)
Baselines	ViT [11]	Audio	0.7823	0.8768	1:55	0.7022	0.8154	1:52
	ViViT [13]	Vision	0.7853	0.8795	1:32	0.7002	0.8234	1:29
	CMA [7]	Audio and Vision	0.5680	0.6768	4:50	0.6078	0.7033	6:05
	AVM [10]	Audio and Vision	0.7756	0.8722	4:49	0.7205	0.8287	4:49
	ViT-ViViT [11, 13]	Audio and Vision	0.7517	0.8541	3:31	0.6901	0.8009	3:32
	Co-attention	Audio and Vision	0.5599	0.6603	4:50	0.5864	0.6688	4:50
Variants of Ours	Sec2Sec Audio	Audio	0.7832	0.8780	1:51	0.7021	0.8191	1:50
	Sec2Sec Vision	Vision	0.7766	0.8733	1:20	0.6970	0.8179	1:20
	Sec2Sec SA-SA	Audio and Vision	**0.7990**	**0.8876**	2:17	0.7047	0.8179	2:14
	Sec2Sec SA-CA	Audio and Vision	0.7949	0.8840	2:14	**0.7322**	**0.8372**	2:15

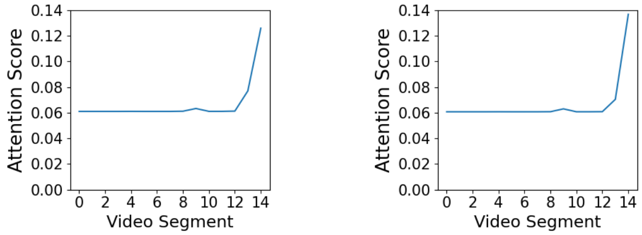
Table 2: Performance comparison of our model with baselines for personality prediction.

	Method	Modality	Agreeableness	Conscientiousness	Extraversion	Neuroticism	Openness	Avg Training Time Per Epoch (min)
Baselines	ViT [11]	Audio	0.891	0.879	0.885	0.886	0.887	1:03
	ViViT [13]	Vision	0.894	0.876	0.878	0.877	0.883	4:48
	CMA [7]	Audio and Vision	0.894	0.882	0.887	0.889	0.890	2:24
	AVM [10]	Audio and Vision	0.894	0.881	0.889	0.887	0.893	2:17
	ViT-ViViT [11, 13]	Audio and Vision	0.897	0.882	0.886	0.889	0.892	6:15
	Co-attention	Audio and Vision	0.890	0.883	0.883	0.888	0.892	1:29
Variants of Ours	Sec2Sec Audio	Audio	0.895	0.878	0.884	0.881	0.888	0:23
	Sec2Sec Vision	Vision	0.893	0.876	0.879	0.877	0.883	0:23
	Sec2Sec SA-SA	Audio and Vision	0.895	0.878	0.884	0.881	0.888	0:49
	Sec2Sec SA-CA	Audio and Vision	**0.898**	**0.891**	**0.892**	**0.892**	**0.896**	1:27

SA-SA, indicating the usefulness of the co-attention mechanism at predicting valence.

Table 2 presents the experimental results for personality predictions. The proposed Sec2Sec SA-CA method consistently outperforms three bi-modal baseline methods and the co-attention method in all five personality label predictions. This confirms the effectiveness of the Sec2Sec structure and highlights the importance of spatio-temporal information. Furthermore, CMA consistently outperforms other baselines, demonstrating the importance of visual and audio information as well as audio-visual interactions captured by cross-modal attention. Notably, Sec2Sec SA-CA requires less training time than the baselines, illustrating the efficiency of the Sec2Sec structure. Additionally, Sec2Sec SA-CA outperforms both Sec2Sec Audio and Sec2Sec Vision, indicating that bi-modal information is more effective than using either modality alone for personality predictions. Finally, the superior performance of Sec2Sec SA-CA over Sec2Sec SA-SA demonstrates the ability of the co-attention mechanism to capture rich interaction information between audio and vision.

Model Interpretability. We assess the contribution of each video segment (i.e., every one-second clip) to emotion prediction by substituting LSTM with an attention-based LSTM proposed by [22]. After training, we obtain the learned LSTM attention values. The attention values of each video segment for valence and arousal are plotted in Figure 2a and 2b, respectively. Similar patterns are observed in both figures. They showcase that the emotion prediction power reaches the highest for the last 3 seconds of the video, suggesting that emotions are mostly influenced by the last 3 seconds. Moreover, the impact of video segments increases as they approach the end of a video.



(a) Valence

(b) Arousal

Fig. 2: The LSTM Attentions for LIRIS-ACCEDE.

5. CONCLUSION

This study introduces an innovative Sec2Sec Co-attention Transformer model for perceived affect prediction in videos. Our approach harnesses the power of pre-trained ResNet networks, LSTM, and a co-attention mechanism to effectively encode and integrate multimodal features. The experimental results underscore the efficiency of our Sec2Sec structure and the significance of inter-modal interaction in affective prediction. In addition, we present an attention-driven LSTM technique to examine the impact of each second clip within a video on the overall affective prediction.

6. REFERENCES

[1] F. Kujur and S. Singh, "Emotions as predictor for consumer engagement in youtube advertisement," *Journal of Advances in Management Research*, 2018.

[2] C. Lou and S. Yuan, "Influencer marketing: How message value and credibility affect consumer trust of branded content on social media," *Journal of interactive advertising*, vol. 19, no. 1, pp. 58–73, 2019.

[3] T. M. DINH, T. H. L. LE, and N. L. H. NGUYEN, "Emotions and video sharing behavior on facebook of young generation," in *International Conference on Emerging Challenges: Business Transformation and Circular Economy*, pp. 407–414, 2021.

[4] B. B. Murdock Jr, "The serial position effect of free recall.," *Journal of experimental psychology*, vol. 64, no. 5, p. 482, 1962.

[5] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari, *et al.*, "Combining modality specific deep neural networks for emotion recognition in video," in *Proceedings of the 15th ACM on International conference on multimodal interaction*, pp. 543–550, 2013.

[6] H. Zhou, D. Meng, Y. Zhang, X. Peng, J. Du, K. Wang, and Y. Qiao, "Exploring emotion features and fusion strategies for audio-video emotion recognition," in *2019 International conference on multimodal interaction*, pp. 562–566, 2019.

[7] Y. Cheng, R. Wang, Z. Pan, R. Feng, and Y. Zhang, *Look, Listen, and Attend: Co-Attention Network for Self-Supervised Audio-Visual Representation Learning*, p. 3884–3892. ACM, 2020.

[8] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *Proceedings of ICCV*, pp. 609–617, 2017.

[9] H. Alwassel, D. Mahajan, B. Korbar, L. Torresani, B. Ghanem, and D. Tran, "Self-supervised learning by cross-modal audio-video clustering," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9758–9770, 2020.

[10] F. Kuhnke, L. Rumberg, and J. Ostermann, "Two-stream aural-visual affect analysis in the wild," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 600–605, 2020.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[12] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of ICCV*, pp. 10012–10022, 2021.

[13] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of ICCV*, pp. 6836–6846, 2021.

[14] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *Proceedings of CVPR*, pp. 3202–3211, 2022.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of CVPR*, pp. 770–778, 2016.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017.

[17] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE Conference on CVPR*, June 2018.

[18] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980.

[19] Y. Baveye, E. Dellandréa, C. Chamaret, and L. Chen, "Liris-accede: A video database for affective content analysis," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 43–55, 2015.

[20] V. Ponce-López, B. Chen, M. Oliu, C. Corneanu, A. Clapés, I. Guyon, X. Baró, H. J. Escalante, and S. Escalera, "Chalearn lap 2016: First round challenge on first impressions-dataset and results," in *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pp. 400–418, 2016.

[21] E. Acar, F. Hopfgartner, and S. Albayrak, "Understanding affective content of music videos through learned representations," in *MultiMedia Modeling: 20th Anniversary International Conference, MMM 2014, Dublin, Ireland, January 6-10, 2014, Proceedings, Part I 20*, pp. 303–314, 2014.

[22] Y. Wang, M. Huang, X. Zhu, and L. Zhao, "Attention-based lstm for aspect-level sentiment classification," in *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 606–615, 2016.