Vid2Int: Detecting Implicit Intention from Long Dialog Videos – Supplementary Material –

Xiaoli Xu[1*] Yao Lu[1*] Zhiwu Lu[1,2†] Tao Xiang[3]
[1]Gaoling School of Artificial Intelligence, Renmin University of China
[2]Beijing Key Laboratory of Big Data Management and Analysis Methods
[3]University of Surrey, United Kingdom
lmnhsp@gmail.com luyao777@ruc.edu.cn luzhiwu@ruc.edu.cn

In this document, we provide more supporting materials in addition to our main paper. Firstly, we present the set of 20 candidate questions designed for constructing our own dataset. Secondly, we show the results obtained by our Vid2Int model using both RGB frames and optical flow. Thirdly, we make comparison to state-of-the-art video analysis models on our Vid2Int-Deception dataset.

1. 20 Candidate Questions for Data Collection

1. What (e.g. character, good looking, height, or family condition) is the most important when you choose a boyfriend/girlfriend? Please also explain why.
2. Have your parents done any disappointing thing to you? What is the most disappointing?
3. What will you do when you have a conflict with your roommate?
4. When you find that your experiment results are wrong, what would you do?
5. What will you do if the first author of your paper was robbed?
6. When you see someone else submit a paper to multiple conferences, will you do the same? why?
7. Do you agree with the behavior of companionship learning at Shandong University? what is the reason?
8. What will you do when you see someone cheating?
9. Will you accept homosexuality in your life? why?
10. Do you think programming is a happy experience? What is the reason?
11. What will you do when you hear someone saying bad things behind you?
12. What will you do when your adviser assigns you a complex task?
13. When your friends invite you to watch movies you don't like, what will you do?
14. Do you love your research direction now? Why?
15. Please describe your home.
16. Please describe the jobs of your parents.
17. Please describe your height, weight, looks.
18. Please describe your major and research direction.
19. Do you like your parents' careers? why?
20. If your favorite gift was given to others by your mother, What will you do?

Method	Answer-ACC	Dialog-ACC
MGRM+AT-LSTM(RGB only)	72.69	88.46
MGRM+AT-LSTM(optical flow only)	76.53	**92.31**
MGRM+AT-LSTM(optical flow + RGB)	**76.92**	**92.31**

Table 1. The results obtained by our Vid2Int model (i.e. MGRM+AT-LSTM) using both RGB frames and optical flow on our Vid2Int-Deception dataset.

2. Fusing RGB Frames and Optical Flow

Table 1 shows the results obtained by our Vid2Int model using both RGB frames and optical flow. We also present the results by using only RGB frames and by using only optical flow. It can be seen that: (1) Optical flow clearly yields better results than RGB frames for implicit intention detection from long dialog videos. (2) Adding RGB frames to optical flow leads to slight improvements. Therefore, for computational efficiency, we only exploit optical flow for implicit intention detection in this work.

*Equal Contribution
†Corresponding Author

Method	Answer-ACC	Dialog-ACC
I3D [1]	66.92	78.85
TSM [4]	67.69	80.77
SlowFast [2]	73.08	80.77
I3D [1] + LSTM [3]	70.77	84.62
TSM [4] + LSTM [3]	73.85	86.54
SlowFast [2] + LSTM [3]	75.38	88.46
Vid2Int (ours)	**76.92**	**92.31**

Table 2. Comparative accuracies (%) for IID on our Vid2Int-Deception dataset.

3. Comparison to the State-of-the-Arts

We further make comparison to state-of-the-art video analysis models [1, 2, 4] on our Vid2Int-Deception dataset. Note that I3D and TSM fuse both RGB frames and optical flow, while SlowFast does not exploit optical flow. The comparative results in Table 2 show that: (1) Our Vid2Int model yields much better results than all competitors, validating the effectiveness of our multi-grain representation (based on optical flow) for implicit intention detection. (2) Interestingly, our Vid2Int model significantly outperforms I3D and TSM, although they also use optical flow. This provides further evidence that our multi-grain representation is crucial for implicit intention detection.

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017.

[2] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. *arXiv preprint arXiv:1812.03982*, 2018.

[3] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[4] Ji Lin, Chuang Gan, and Song Han. TSM: Temporal shift module for efficient video understanding. In *ICCV*, pages 7083–7093, 2019.