Action Anticipation

Feiyu Zhu

1 Introduction

There has been a substantial amount of work in enabling machines to understand the world. In the past decade models are developed to detect objects in images[1], extract contextualized representation from natural languages[2], recognize action from videos[3], and build connections across different modalities[4]. There are considerable advancements in these domains and some of them even achieves super-human level. However, if we want to build a system that interacts with human in reality, for example an assistive household robot, knowing what the world *is* at present is not enough, it is also crucial for the system to know what the world *will be*.

The ability to predict what is going to happen, or specifically what the human's are going to do is important. Lane segmentation model is capable of keeping an autonomous vehicle cruise on its own, but only with the ability to predict the action of the surrounding human drivers can the autonomous vehicle switch lane or enter the highway[5]. Being able to predict what a surgeon will do next is also useful to know which equipment to prepare and how should it be handed to the surgeon, and this could save precious time in a surgery. It is demonstrated that anticipating human actions can facilitate human robot interaction[6].

Action anticipation is the task where the model takes in the past observation of a person, commonly in the form of video sequences, and output a probability distribution over a set of possible actions that predicts what the person will do next[7] in a bounded time interval. It is evaluated by comparing what the person actually does next and the top candidates provided by the model. To be proficient in this task, the model should be able to reason the correlation between adjacent actions, and have some notion of human conventions (i.e. what order do humans typically follow for certain tasks).

2 Related Work

2.1 Action Anticipation

Many recent works in action anticipation follow the encoder-decoder paradigm. [8] used a encoding LSTM to generate a feature representation of the past observation, and a decoding LSTM to infer the future. They also designed a modality attention mechanism to fuse RGB, optical flow and object based features. [9] proposed a similar approach, and their improvement mainly attributes to making use of low frame rate and high frame rate pathways to disentangle spatial semantics and motion[10]. [11] and [12] replaced the LSTM with transformer and employed different attention mechanisms.

Some other formulate the task as a reinforcement problem. [13] viewed visual sequences as a Markov Decision Process and thus can make use of imitation learning to predict the future frames. They argued that inverse reinforcement learning can be directly applied to pixel level prediction and can therefore anticipate future actions based on the raw pixel values. [14] combined the encoder-decoder paradigm with reinforcement learning by using a reinforcement module to supervise the LSTM encoder and decoders, which is shown to be better than mere cross entropy loss.

Earlier work in action anticipation exploited non deep learning based methods such as hierarchical representation of actions[15], modeling human states as Hidden Markov Model [16], representing human activities using temporal stochastic grammar[17].

2.2 Early Action Recognition

Early action recognition share many similarity with action anticipation, where the most notable difference is that in early action recognition the model has access to the first few frames of the action to be predicted. Many action anticipation work claimed that their method can be applied to early action recognition without modification[8].

But a more popular path is to make modifications to the widely investigated action recognition models so that they are better at partial action recognition.

[18] designed a new loss function such the first few frames are weighted more, and thus making the LSTM model capable of recognizing actions with only the beginning of an action. [19] constructed two LSTM where one of them is trained regularly for action recognition while the other is passed in with latency so that it can learning to predict the feature representation of the regular LSTM with only partial video.

Despite more work has been done in early recognition domain, it is essentially a different task compare to action anticipation. Achieving great performance in early action recognition only requires the model to be robust at single action recognition, and does not require the model to reason the relationship between subsequent actions.

2.3 Related Dataset

In theory, any video data with consistent action labels could be used to train action anticipation models. Some data set are gathered with action anticipation task in mind. [7] is an egocentric video data set specific to the kitchen settings. Beside video sequence and action labels, it also prepares auxiliary information such as optical flow and object detection. [20] is a smaller first person data set in cooking but it also provides the eye gaze of the user. [21] extends beyond kitchen to household activities, but it is in third person view unlike the other two.

There are other instruction dataset mainly designed for weakly supervised text/video representation training. [22] contains instructional videos with natural language narrations of common actions in real life.

3 Preliminary Results

	Encoder	Decoder		Verb	Noun	Action
RULSTM[8]	TSN[23] + LSTM	LSTM	+ Linear	27.5	29.0	13.3
AVT[12]	Vision Transformer[24]	Causal Transformer	+ Linear	30.2	31.7	**14.9**
Naive (Ours)	Vision Transformer	–	+ Linear	64.8	17.1	8.0

Table 1: Class mean top 5 recall with RGB frame input on Epic-Kitchen[7]

Table 1 compares the performance of representative models. Action labels are verb-noun pairs so the result is considered correct if and only if both verb and noun label match. Despite transformers yield better performance than LSTM (as shown by comparing RULSTM and AVT), they still need to follow the encode-decode paradigm and is not capable of compressing the model into a single transformer. Using only vision transformer as encoder and a simple MLP as classification head (Naive Transformer model) will result in a close to majority class classifier. And because Epic-Kitchen has less verbs than the nouns (79 vs. 300), and the verbs concentrated on "Retrieve", "Leave", and "Clean" categories, a majority class classifier can yield good results (64.8 top 5 recall). However, as it failed to effectively distinguish the nouns, it has a poor performance on the overall action label.

4 Open Research Challenges

Although replacing LSTM with transformers yields better performance, no base model is powerful enough to be trained end-to-end. Currently both base model architectures have to follow the encode-decode paradigm which requires separate supervision on the encoder and decoder. So a reasonable next goal is to design an architecture that can generate representation of actions in the future given the current frames in a single stage. This will simplify training procedures, and may also inspire other work in multi-modality sequential data processing.

Furthermore, in addition to predicting what the next action is, the model should also learn to predict when will the next action happen. At present the models are trained to predict next action in some fix interval, that is, the model is told the length of interval between the current action and the next action. This information is not necessarily accessible during inference is the real world. And even within the existing data set, this interval can vary. Therefore from a practical perspective, the model should be able to estimate the interval till next action.

5 Qualitative Analysis

5.1 Correct Action

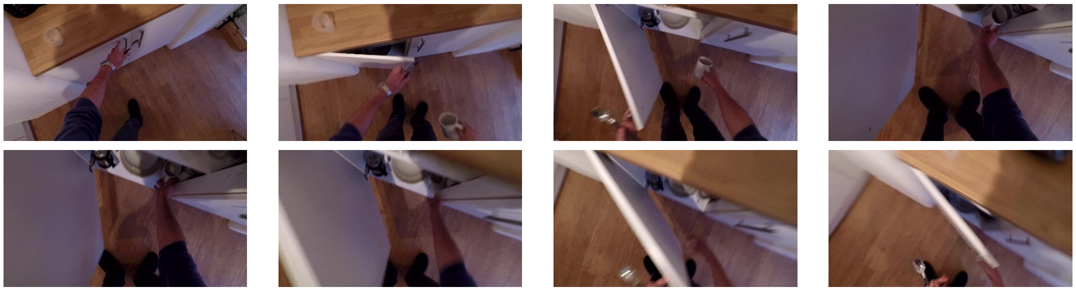
observation action frames			
ground truth	(put, knife)	predictions	(put, knife) (put, pan) (close, drawer) (wash, spoon) (turn-on, tap)
observation action frames			
ground truth	(close, cupboard)	predictions	(close, cupboard) (take, fork) (wash, pot) (take, cup) (put, cup)

Table 2: Examples for correct action inference

5.2 Correct Verb; Incorrect Noun

observation action frames			
ground truth	(put, spoon)	predictions	(put, pan) (put, spatula) (close, drawer) (put, cap) (take, cup)

observation				
action frames				
ground truth	(wash, plate)	predictions	(wash, pot) (take, fork) (put, pan) (take, cup) (wash, cloth)	

Table 3: Examples for correct verb but incorrect noun inference

5.3 Correct Noun; Incorrect Verb

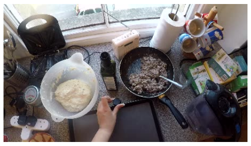
observation				
action frames				
ground truth	(take, spatula)	predictions	(put, spatula) (put, pan) (wash, top) (turn on, tap) (take, spoon)	
observation				
action frames				
ground truth	(take, oil)	predictions	(pour, oil) (close, drawer) (put, cup) (close, cupboard) (take, cup)	

Table 4: Examples for correct noun but incorrect verb inference

5.4 Observations

As shown in the correct action examples, the naive model has to some extend learned the semantic meaning of the images. In addition to having the correct verb noun pair as the top candidate, the model also rank other reasonable action in the top five, indicating some generality of the model to capture diverse user habits.

However the incorrect cases reflect some of the problems. There are actions such as "(close, drawer)" that appears in the top predictions for above examples despite there is no clear implication of it in the scene (i.e. no open drawer). This attributes to the majority class classification problem where the model given common action near constant confidence regardless of the frame contents.

6 Proposed Interventions and Future Work

Some actions such as "open cupboard" can happen at anytime: the user may open the cupboard in the middle of stirring food in order to look for sauce, or it can also happen at ingredients preparation phase and/or the final cleaning phase. Because their independence nature, it is hard for the model to learn the correlation between these actions and the visual features from the previous actions. This have a major impact on the performance as the model always give them scores close to their statistical frequency.

One possible solution is to add a post process that filter out impossible action based on logical deduction of the pre-condition and post-condition of actions. For example "close cupboard" cannot happen unless the cupboard is opened previously. Information like this help constraining the candidate action space by removing incorrect actions that are not less dependent on previous actions. Possible ways to implement such a post filtering could be crowd-sourcing and hard code a knowledge graph that encodes the relationships between different actions, or alternatively Regression Planning Networks [25] can be used to learn classic planning from video data. At present no previous work has explored incorporating external knowledge for action anticipation task.

References

[1] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*, pp. 213–229, Springer, 2020.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[3] D. Moltisanti, S. Fidler, and D. Damen, "Action recognition from single timestamp supervision in untrimmed videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9915–9924, 2019.

[4] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: an overview of methods, challenges, and prospects," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1449–1477, 2015.

[5] D. Sadigh, "Interaction-aware planning: A human-centered approach toward autonomous driving." Institute for Pure & Applied Mathematics, 2020.

[6] H. S. Koppula and A. Saxena, "Anticipating human activities using object affordances for reactive robotic response," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, pp. 14–29, 2015.

[7] D. Damen, H. Doughty, G. M. Farinella, , A. Furnari, J. Ma, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, "Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100," *International Journal of Computer Vision (IJCV)*, 2021.

[8] A. Furnari and G. M. Farinella, "Rolling-unrolling lstms for action anticipation from first-person video," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2020.

[9] N. Osman, G. Camporese, P. Coscia, and L. Ballan, "Slowfast rolling-unrolling lstms for action anticipation in egocentric videos," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 3437–3445, October 2021.

[10] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6202–6211, 2019.

[11] Q. Ke, M. Fritz, and B. Schiele, "Time-conditioned action anticipation in one shot," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9925–9934, 2019.

[12] R. Girdhar and K. Grauman, "Anticipative Video Transformer," in *ICCV*, 2021.

[13] K.-H. Zeng, W. B. Shen, D.-A. Huang, M. Sun, and J. Carlos Niebles, "Visual forecasting by imitating dynamics in natural sequences," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2999–3008, 2017.

[14] J. Gao, Z. Yang, and R. Nevatia, "Red: Reinforced encoder-decoder networks for action anticipation," *arXiv preprint arXiv:1707.04818*, 2017.

[15] T. Lan, T.-C. Chen, and S. Savarese, "A hierarchical representation for future action prediction," in *European Conference on Computer Vision*, pp. 689–704, Springer, 2014.

[16] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert, "Activity forecasting," in *European conference on computer vision*, pp. 201–214, Springer, 2012.

[17] S. Qi, S. Huang, P. Wei, and S.-C. Zhu, "Predicting human activities using stochastic grammar," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1164–1172, 2017.

[18] M. Sadegh Aliakbarian, F. Sadat Saleh, M. Salzmann, B. Fernando, L. Petersson, and L. Andersson, "Encouraging lstms to anticipate actions very early," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 280–289, 2017.

[19] R. De Geest and T. Tuytelaars, "Modeling temporal structure with lstm for online action detection," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1549–1557, IEEE, 2018.

[20] Y. Li, Z. Ye, and J. M. Rehg, "Delving into egocentric actions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 287–295, 2015.

[21] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *Proceedings of the ieee conference on computer vision and pattern recognition*, pp. 961–970, 2015.

[22] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips," in *ICCV*, 2019.

[23] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *European conference on computer vision*, pp. 20–36, Springer, 2016.

[24] 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.

[25] D. Xu, R. Martín-Martín, D.-A. Huang, Y. Zhu, S. Savarese, and L. F. Fei-Fei, "Regression planning networks," *Advances in Neural Information Processing Systems*, vol. 32, pp. 1319–1329, 2019.