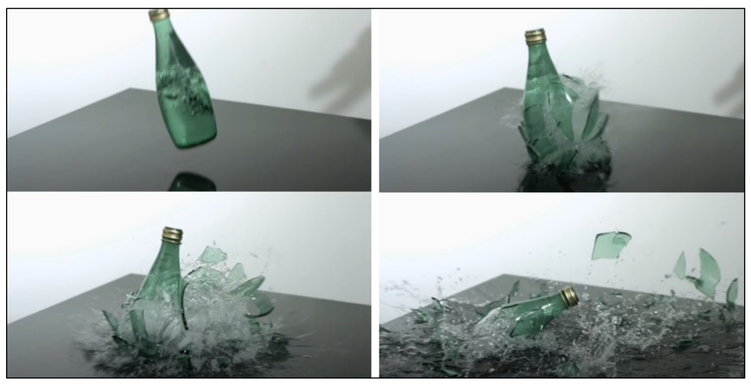
Can ModelScopeT2V Understand Arrow of Time?

Paanugoban Kugathasan
Masters In Applied Computing (MScAC), Dept. Computer Science, University of Toronto

Motivation

- Understand if Text to Video Models are capable of generating videos that follow the **Arrow of Time (AoT). AoT** is the irreversible direction of time flow, events such as a bottle breaking would not happen in reverse
- **ModelScopeT2V** [2] was released in 2023 by Alibaba group as a text to video model (Available on Hugging Face)
- Implications of improper **AoT** include poor communication of instructions, worsening of social media trust, poor media production, unable to increase FPS as poor **AoT** (DLSS like technologies for future upscaling)



[1]

Related Work

- In **2014**, **"Seeing the Arrow of Time"** [3] explored 125 forward videos and 25 reverse videos, predominantly featuring physics related content such as gravity, friction and entropy. Achieved 90%, 77% and 75% on test sets
- In **2018, "Learning and Using the Arrow of Time"** [4] took broad range of video datasets normalized FPS, eliminated artificial elements (black bars), and stabilized camera motion. Classifier used 2D convnets and optical flows. Saw accuracy of 76% on Flickr, 72% on Kinetrics datasets with human accuracy of 80%
- In **2021, "ArrowGAN: Learning to Generate Videos by Learning Arrow of Time"** [5] introduced an AoT discriminator called Arrow-D to generative adversarial networks for video generation. Discriminator used 3D convolutional nets

References

[1] *T. L. Directory, "Breaking slow motion hd a green glass mineral water bottle dropping shattering in slow mo," Online video, Feb. 2013, available from: https://www.youtube.com/watch? v=t2NSxiFo1go. [Online]. Available: https://www.youtube.com/ watch?v=t2NSxiFo1go*

[2] *J. Wang, H. Yuan, D. Chen, Y. Zhang, X. Wang, and S. Zhang, "Modelscope text-to-video technical report," 2023.*

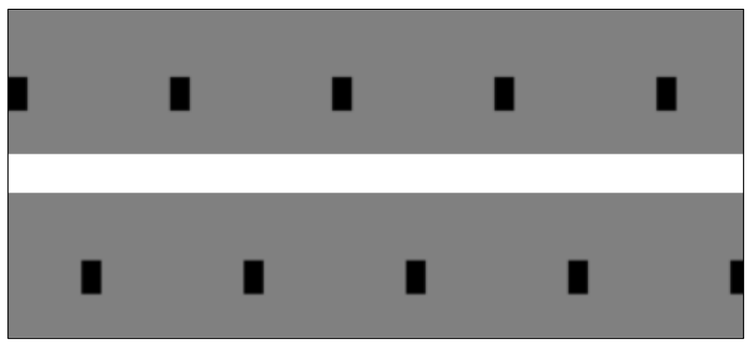
[3] *L. C. Pickup, Z. Pan, D. Wei, Y. Shih, C. Zhang, A. Zisserman, B. Scholkopf, and W. T. Freeman, "Seeing the arrow of time," in 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2043–2050.*

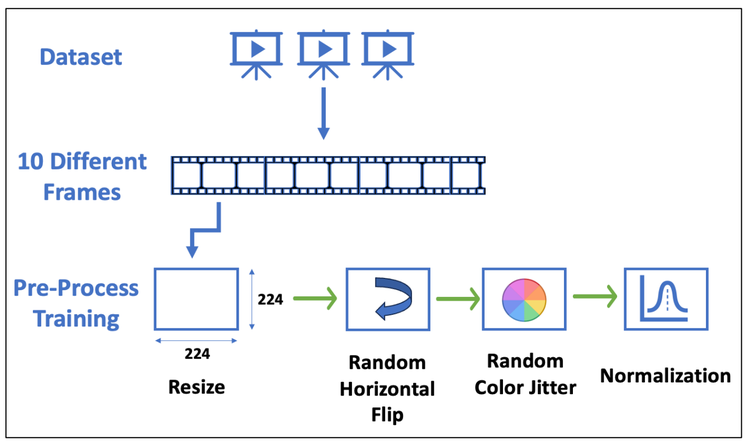
[4] *D. Wei, J. Lim, A. Zisserman, and W. T. Freeman, "Learning and using the arrow of time," in2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 8052–8060.*

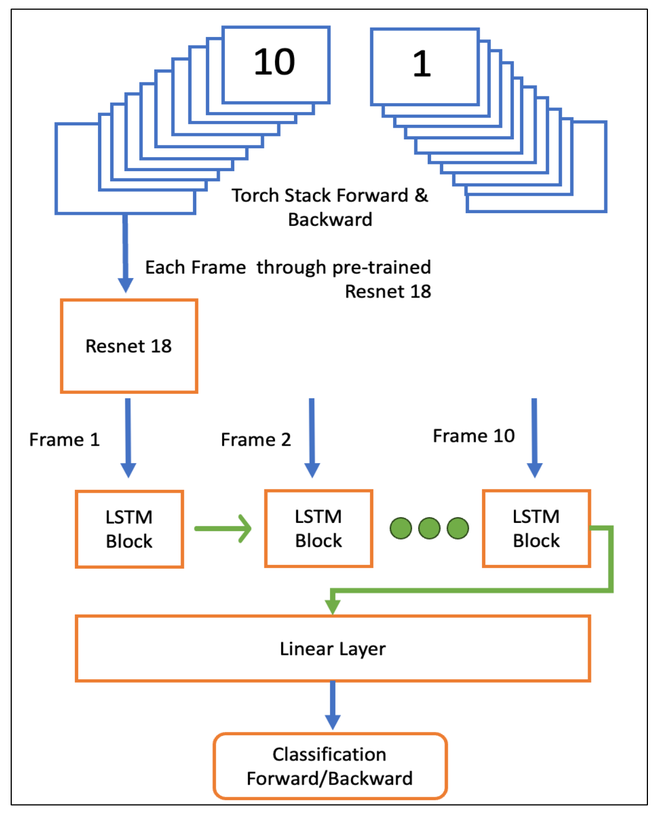
[5] *K. Hong, Y. Uh, and H. Byun, "Arrowgan : Learning to generate videos by learning arrow of time," 2021.*

New Technique

- **Used basic sequences** to trial different architectures to see which solution held promise
- Created a new architecture to classify forward and reverse videos to avoid pre-processing for camera stabilization, artificial cues, or optical flows
- Selected **UCF-101** as the best dataset to minimize camera stabilization







- Custom classifier training achieved **86%** on training set and **78%** on test set of **UCF-101** dataset
- Videos critically in test set have a **balanced prediction** between forward and backward videos. Forward and backward videos are correctly and incorrectly predicted in **equal proportion**

Experimental Results

- **Experiment 1:** Generate 4 Videos (2 forward + 2 backward) for each of the 95 categories in UCF-101
- **Experiment 2:** Select 10 categories (5 Top + 5 Bottom) from Experiment 1, and generate 10 videos for 5 different prompts for each direction
- **Experiment 3:** Select 15 categories (5 Top + 5 Chance + 5 Bottom) from Experiment 1, and generate 10 different videos for 10 different prompts for 5 different length of prompts
- **Fine-Tune Model** model with videos generated from Experiment 3, to see if forward videos follow temporal direction. **74% on forward test set, indicates temporal understanding**

49% Experiment 3/ Reg Model

59% Experiment 3/ Fine Tuned Model

Categories	Forward Accuracy %	Backward Accuracy %
SalsaSpin	73%	37%
ThrowDiscuss	66%	38%
BabyCrawling	66%	41%
HorseRiding	80%	50%
PoleVault	60%	38%
TaiChi	80%	42%
BasketballDunking	80%	50%
Kayaking	53%	44%
Fencing	60%	40%
Surfing	100%	46%
DogWalking	93%	50%
Diving	86%	66%
Punching	60%	38%
MilitaryParade	80%	42%
CuttingInKitchen	73%	25%

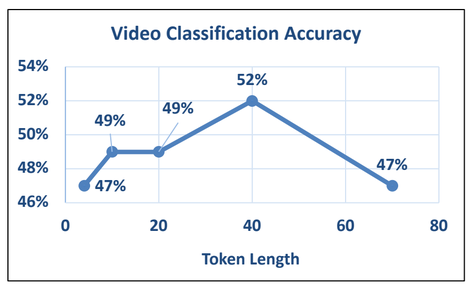
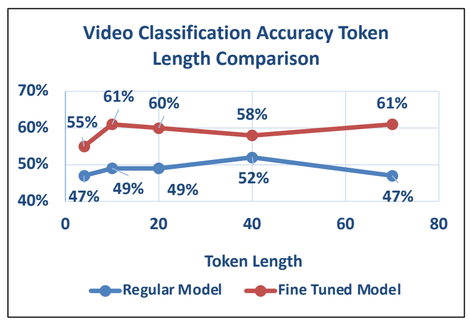
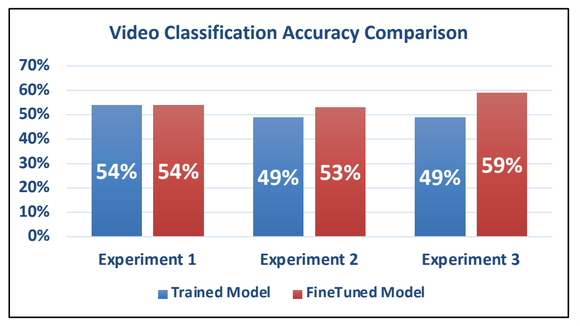


TABLE 6: Experiment 3 Results

Actual	Predicted: Backward	Predicted: Forward	Metric
Backward	233	517	Precision for Backward: 0.4844
Forward	248	502	Recall for Backward: 0.3106
F1 for Backward:			0.3785
Precision for Forward:			0.4926
Recall for Forward:			0.6693
F1 for Forward:			0.5675
Total Videos:			1500
Accuracy:			0.49





- Human Accuracy was **63%**, notably getting **5/15** Backward correct vs **14/15** Forward

Prompt: "Traversing the landscape under a setting sun, the horse and rider's serene journey rewinds, the peaceful evening light receding gently

Prompt: "Horse riding adventure, exploring trails with enthusiasm and grace"

Conclusion

Our experiments suggest forward generated videos by ModelScopeT2V follow AoT, but reverse prompts are struggling to create realistic videos. More experiments should be conducted for validation, including improving accuracy of classifier, larger datasets and increased training on reverse prompts.