DirecFormer: A Directed Attention in Transformer Approach to Robust Action Recognition

Thanh-Dat Truong[1], Quoc-Huy Bui[2], Chi Nhan Duong[3],
Han-Seok Seo[4], Son Lam Phung[5], Xin Li[6], Khoa Luu[1]
[1]CVIU Lab, University of Arkansas [2]NextG, FPT Software
[3]Concordia University [4]Dep. of Food Science, University of Arkansas
[5]University of Wollongong [6]West Virginia University

This material is based upon work supported by the National Science Foundation under Award No. OIA-1946391. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

https://vizle.offnote.co

Contact us: **vizle@offnote.co**

This document was generated automatically by **Vizle**

Your **Personal Video Reader Assistant**

Learn from Videos **Faster** and **Smarter**

VIZLE **PRO / BIZ**

PDF, PPT ~~Watermarks~~

- Convert *entire* videos
- *Customize* to retain all essential content
- Include Spoken *Transcripts*
- Customer support

Visit **https://vizle.offnote.co/pricing** to learn more

VIZLE **FREE PLAN**

PDF only Watermarks

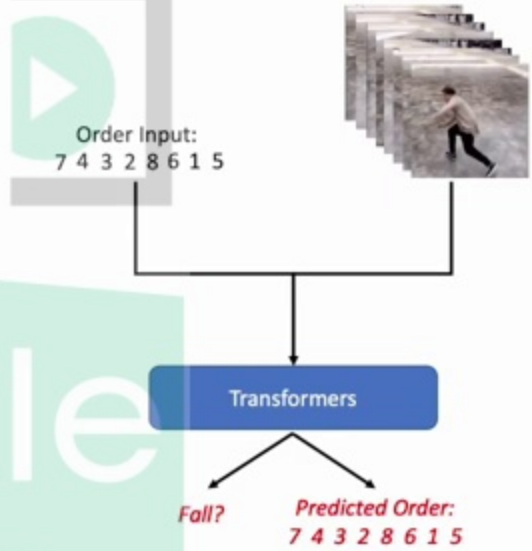
- Convert videos *partially*
- Slides may be *skipped**
- Usage restrictions
- No Customer support

Visit **https://vizle.offnote.co** to try free

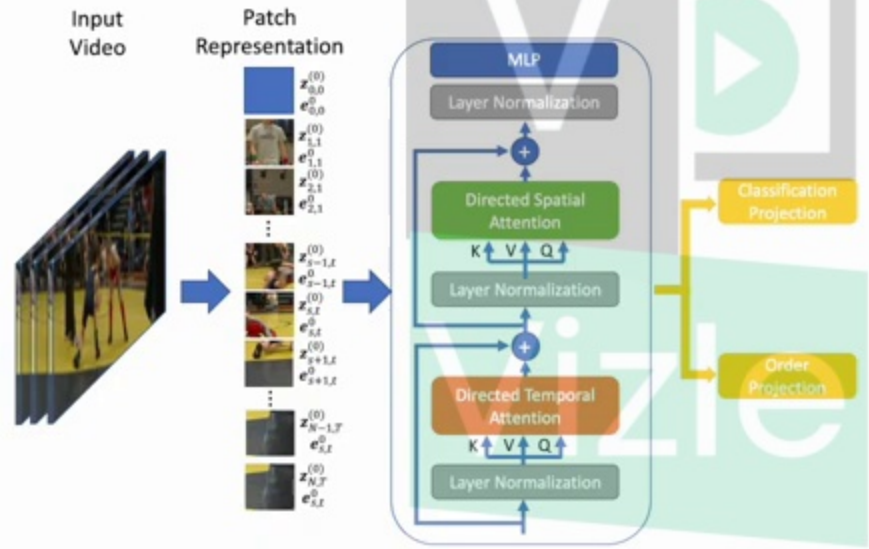
Login to Vizle to unlock more slides*

Motivation

- The goal of action recognition is to predict an action given a set of ***video frames in the correct order.***
- Given a set of ***video frames shuffled in a random order*** and different from the original one, will it be classified as ***the same label as the original recognition result?***
- Are action recognition models able to ***correct the incorrectly-ordered frames to the right ones*** and provide ***an accurate prediction?***



The Proposed Approach



Ablation Study

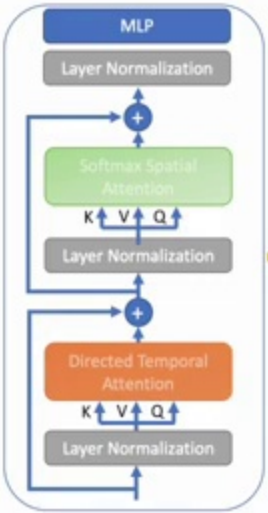


Table 1. **Ablation Study On Jester.** $X - Y$ denotes for the attention types of temporal and spatial dimension, respectively. X (and Y) could be either S: Softmax or C: Cosine.

Models	Attention Time-Space	$\mathcal{L}_{ord}$	$\mathcal{L}_{self}$	Top 1	Top 5
DirecFormer	$S - C$			94.52	99.26
DirecFormer	$S - C$	✓		94.65	99.25
DirecFormer	$C - S$			95.52	99.20

Experimental Results

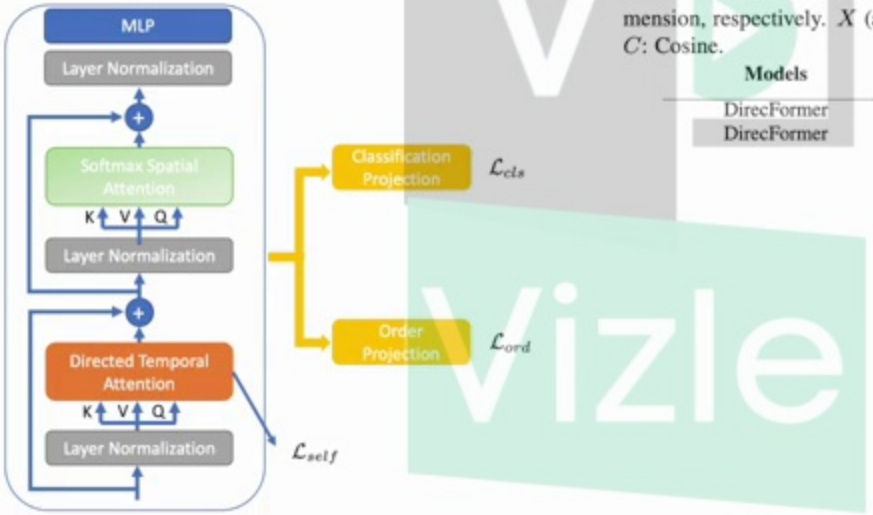


Table 4. **Comparison with the SOTA methods on Kinetics 400.** $X - Y$ denotes for the attention types of temporal and spatial dimension, respectively. X (and Y) could be either S: Softmax or C: Cosine.

Models	Attention Time-Space	Top 1	Top 5
DirecFormer	$S - C$	80.16	94.55
DirecFormer	$C - S$	81.69	94.62

https://vizle.offnote.co

Contact us: **vizle@offnote.co**

This document was generated automatically by **Vizle**

Your **Personal Video Reader Assistant**

Learn from Videos **Faster** and **Smarter**

VIZLE **PRO / BIZ**

PDF, PPT ~~Watermarks~~

- Convert *entire* videos
- *Customize* to retain all essential content
- Include Spoken *Transcripts*
- Customer support

Visit **https://vizle.offnote.co/pricing** to learn more

VIZLE **FREE PLAN**

PDF only Watermarks

- Convert videos *partially*
- Slides may be *skipped**
- Usage restrictions
- No Customer support

Visit **https://vizle.offnote.co** to try free

Login to Vizle to unlock more slides*