Goal: Streaming video recognition
• It reasons about the action in every frame of a video.

➢ Previous approach: linear/quadratic w/ length
➢ Our approach:
• Reformulate x-attention through the lens of kernel
• Streaming attention with O(1) update complexity

Attention as kernels:
Attention \( \left( q_m, x_n \right|_{n=1}^{t} \) = \( \frac{\sum_{n=1}^{t} \kappa(q_m, k_n)v_n}{\sum_{n=1}^{t} \kappa(q_m, k_n)} = \frac{\phi(t)}{\psi(t)} \)

• SoftMax-Attention: \( \kappa(q_m, k_n) = \exp \left( \frac{4mkn}{\sqrt{C}} \right) \)

Streaming Attention:
• Decouple temporal and feature component

\( \kappa(q_m, k_n) \rightarrow \kappa(\lambda_m, \omega_t, f_n + \omega_n) \rightarrow K(\omega_t, \omega_n) \kappa'(\lambda_m, f_n) \)

Box kernel: \( K_B(\omega_t, \omega_n) = 1_{[t-n<N]} \)
Laplace kernel: \( K_L(\omega_t, \omega_n) = e^{-\lambda(t-n)} \)

\( \psi(t) = \psi(t-1) + \kappa'(\lambda_m, f_t) - \kappa'(\lambda_m, f_{t-N}) \)
\( \phi(t) = \phi(t-1) + \kappa'(\lambda_m, f_t)v_t - \kappa'(\lambda_m, f_{t-N})v_{t-N} \)

\( \psi(t) = e^{-\lambda} \psi(t-1) + \kappa'(\lambda_m, f_t) \)
\( \phi(t) = e^{-\lambda} \phi(t-1) + \kappa'(\lambda_m, f_t)v_t \)

Overview of TeSTra (Temporal Smoothing Transformer)

Runtime Comparison

Online action detection on THUMOS14

(a) Using ANet-pretalized feature

Method | mAP | Method | mAP
--- | --- | --- | ---
RED [17] | 45.3 | IDN [18] | 50.0
TMN [20] | 47.2 | TMN [23] | 62.1
OstTR [29] | 58.3 | LSTR [46] | 60.3
Ours' | 68.2 | Ours' | 72.3
Ours | 71.2

(b) Using Kinetics-pretalized feature

Method | mAP | Method | mAP
--- | --- | --- | ---
RED [17] | 45.3 | IDN [18] | 60.3
TMN [20] | 47.2 | TMN [23] | 62.1
OstTR [29] | 58.3 | LSTR [46] | 60.3
Ours' | 68.2 | Ours' | 72.3
Ours | 71.2

Real-time Online Video Detection with Temporal Smoothing Transformers
Yue Zhao, Philipp Krähenbühl
{yzhao, philkr}@cs.utexas.edu