# Two-Stream Transformer Architecture for Long-Form Video Understanding

Edward Fish, Jon Weinbren, Andrew Gilbert

A simple tokenisation method for applying transformers to long-form video tasks.

**Fusion**

Output CLS tokens are fused via concatenation and projection for downstream long video understanding tasks.

**Spatial and Temporal Transformers**

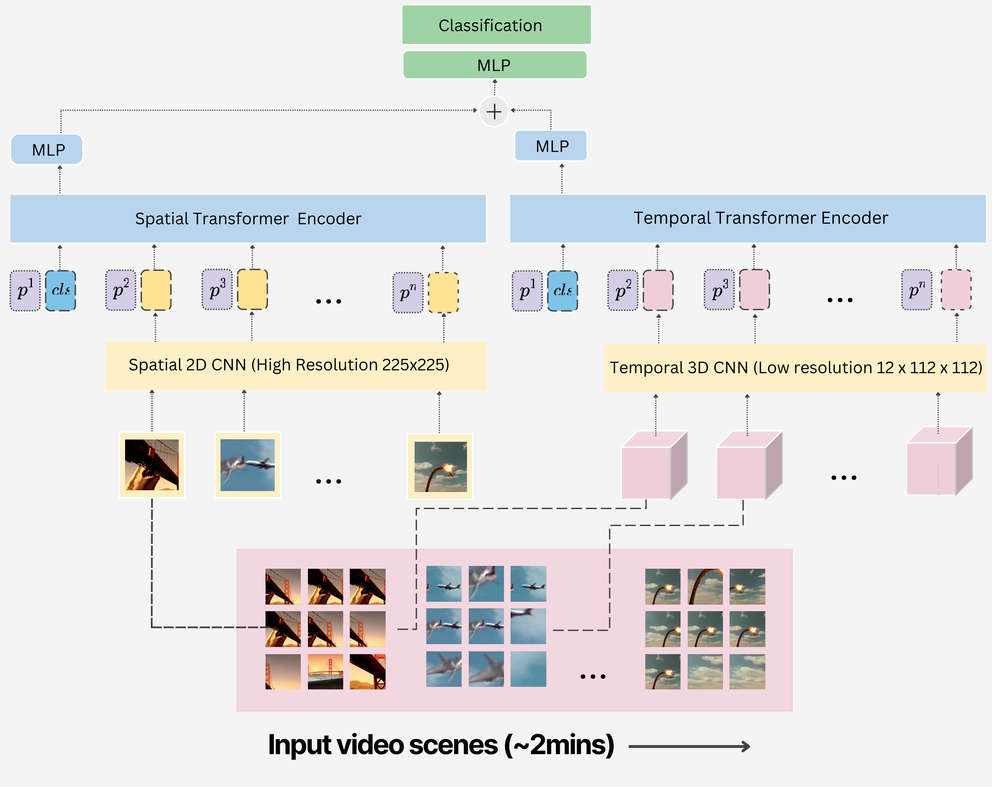
Two shallow multi-headed attention transformer encoders model dependency between spatial and temporal embeddings.

**Token Embedding**

We obtain a central frame **spatial** embedding via ResNet18 pretrained on ImageNet and a temporal **context** embedding via R(2+1)d Video ResNet pretrained on Kinetics 400.

**Frame Extraction**

We split long videos into scenes via RGB channel threshold detection and extract frames. 12 Frames are extracted from each scene.




| Method | Samples | Parameters (Millions) | mAP |
|---|---|---|---|
| Spatial with backprop | 2000 | 28 | 0.5903 |
| Temporal with backprop | 2000 | 48 | 0.6221 |
| Spatial no backprop | 2000 | 16.5 | 0.4024 |
| Temporal no backprop | 2000 | 16.5 | 0.59 |
| Distillation Network | 2000 | 44.5 | 0.6005 |
| Gated Fusion | 2000 | 45 | 0.4728 |
| STAN-Small | 2000 | 45 | 0.6151 |
| STAN-Small | 6047 | 45 | 0.6401 |
| STAN-Large | 6047 | 92.5 | 0.7506 |

**Ablation on fusion methods and impact of temporal and spatial stream.**

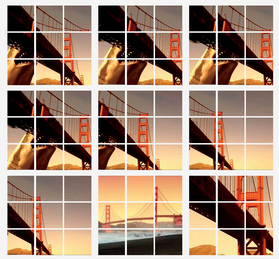
| Method | CNN Feature Extractor | Frames per Scene | Aggregation Method | $\overline{mAP}$ |
|---|---|---|---|---|
| ResNet [24] | ResNet18 | 1 | Avg Pool | 0.434 |
| SqueezeExcite [28] | SE-ResNet | 1 | Avg Pool | 0.544 |
| I3D [6] | I3D | 12 | Avg Pool | 0.487 |
| Collaborative Gating [35] | ResNet18 | 12 | Gated Unit | 0.4723 |
| S(TPN) [64] | ResNet18 | 12 | Temporal Pyramid | 0.492 |
| Fine-Grained Semantic [18] | Multi-Modal | 16 | Concatenation | 0.583 |
| LSTM [12] | ResNet18 | 1 | LSTM | 0.596 |
| Distillation [25, 49] | ResNet18 + R2+1D | 12 | Transformer | 0.601 |
| **STAN-Small** | ResNet18 + R2+1D | 12 | Transformer | **0.640** |
| **STAN-Large** | ResNet18 + R2+1D | 12 | Transformer | **0.750** |

**Results on MMX-Trailer 20 [2] for Genre Classification of long videos.**

| Method | Relation | Speaking | Scene |
|---|---|---|---|
| R101-SlowFast+NL [17] | 52.4 | 35.8 | 54.7 |
| VideoBERT [48] | 52.8 | 37.9 | 54.9 |
| Object Transformer [56] | 53.1 | 39.4 | 56.9 |
| STAN-Small (ours) | **54.2** | **40.6** | **57.3** |
| STAN-Large (ours) | **56.25** | **41.41** | **58.33** |

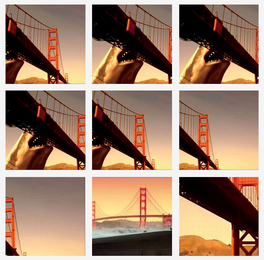
**Results on Long Form Video Understanding Dataset [1] for relation identification, speaker recognition, and scene classification.**

**ViT [3] & ViViT[4] Tokens**



**1 Scene = 9 x 9 Tokens**

**STAN Temporal Context Token**

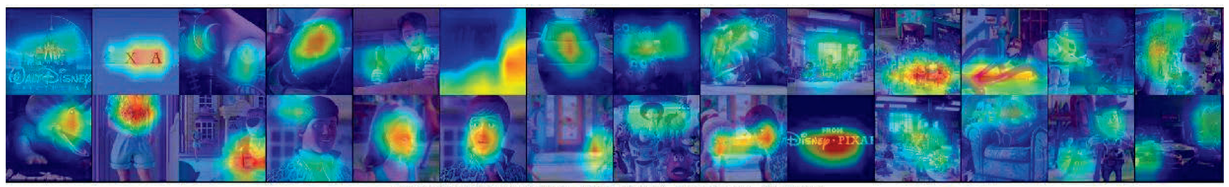



**1 Scene = 1 Token**

**STAN Spatial Token**


**Plus additional spatial token per scene**

**Class attention map for all spatial tokens of 28 scene trailer with temporal context.**



Predicted:['Animation', 'Adventure', 'Comedy', 'Family']
Target:['Animation', 'Comedy', 'Family']

Paper

edward.fish@surrey.ac.uk

[1] Wu, Chao-Yuan, and Philipp Krahenbuhl. "Towards long-form video understanding." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[2] Fish, Edward, Jon Weinbren, and Andrew Gilbert. "Rethinking movie genre classification with fine-grained semantic clustering." arXiv preprint arXiv:2012.02639 (2020).
[3] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).
[4] Arnab, Anurag, et al. "Vivit: A video vision transformer." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.