Temporal Models for Video Processing
Tasks

- Action recognition
- Action prediction
- Tracking

- Monocular 3D estimation (active sensing)
Datasets

- **HMDB-51:** 7000 videos of 51 actions
- **UFC 101:** 13,320 videos of 101 actions
- **Kinetics:** Up to 650,000 videos, up to 700 actions
Approach 1: Unordered Frames

Pretty good baseline

Combination
Approach 2: Frames + Global Model

Tends to overfit, hard to train
3D Convolutions

Convolution across both space and time

[Diagram showing 3D space dimensions: Height, Width, Time]
3D Convolutions

Input \( X \in \mathbb{R}^{C_i \times T \times H \times W} \)

Kernel \( W \in \mathbb{R}^{C_o \times C_i \times t \times h \times w} \)

Bias \( b \in \mathbb{R}^{C_o} \)

Output \( Z \in \mathbb{R}^{C_o \times (T-t+2p_t+1) \times (H-h+2p_h+1) \times (W-w+2p_w+1)} \)

\[
Z_{c,d,a,b} = b_c + \sum_{l=0}^{C_i} \sum_{i=0}^{t} \sum_{j=0}^{h} \sum_{k=0}^{w} X_{k,d+i,a+j,b+k} \cdot W_{c,l,i,j,k}
\]

Very slow!
Factorized 3D Convolutions

“2+1D”

\[
O(HWTC^2)
\]

\[
O((HW+T)C^2)
\]
- Pre-train a network on ImageNet

- “Inflate” some 2D convolutions to 3D

$I^{2+1D}$

```
... → Conv 3x3 → ... → Conv 3x1x1
                      → Initialize weights with $1/t$
                      → Copy weights
... → Conv 1x3x3 → ...
```
Open Problem: What Tasks Should We Care About?

• Vision tasks are often proxies or initial steps in other applications
  - But they often don’t really capture the downstream task.

• Vision is good as a test bed.
  - New architectures
  - Pre-training