LEARNING TEMPORAL EMBEDDINGS FOR COMPLEX VIDEO ANALYSIS

BY RAMANATHAN, TANG, MORI, AND LI

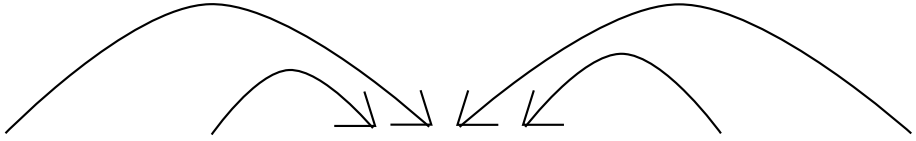
Chad Voegele

PROBLEM

What can we learn about videos

without supervision?

MOTIVATION



... quick fox jumps over dog ...

WORD2VEC FOR VIDEOS?

words $\approx$ frames

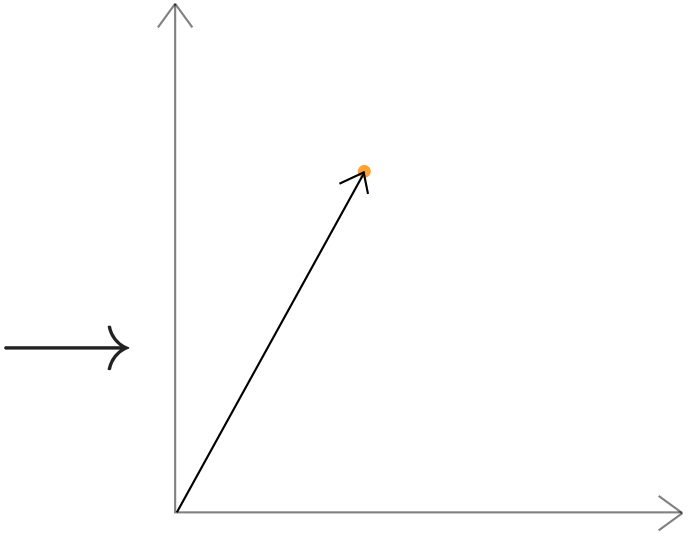
sentences $\approx$ video segments

WORD2VEC FOR VIDEOS?

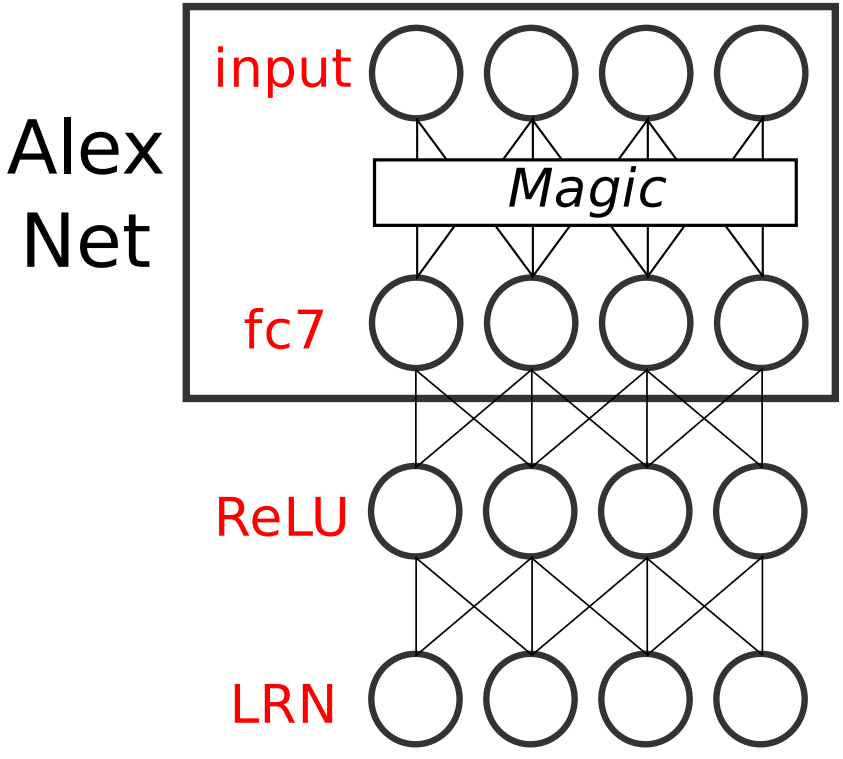
ISSUES

1. Frames are not discrete.
2. Visual similarity between neighboring frames.
3. Representation of context.

FRAME EMBEDDING



FRAME EMBEDDING



EMBEDDING OBJECTIVE

$$\begin{aligned}\text{similarity}(a, b) &= \frac{a \cdot b}{\|a\| \|b\|} \\ &= a \cdot b\end{aligned}$$

EMBEDDING OBJECTIVE

$$f_{v_j} \cdot h_{v_j} \gg f_{-} \cdot h_{v_j}$$

EMBEDDING OBJECTIVE

$$\min_{embedding} \sum_{v \in \mathcal{V}} \sum_{v_j \in v} \sum_{v_- \neq v_j} \max\left(0, 1 - \left(f_{v_j} - f_-\right) \cdot h_{v_j}\right)$$

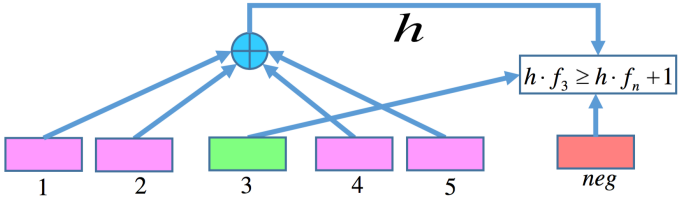
EMBEDDING OBJECTIVE

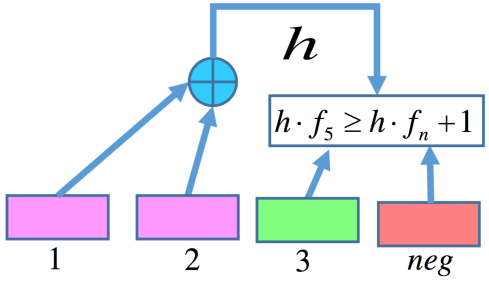
WANT

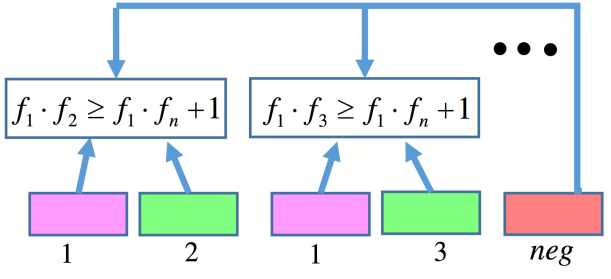
$$1 - \left(f_{v_j} - f_-\right) \cdot h_{v_j} < 0$$

$$\Leftrightarrow \quad f_{v_j} \cdot h_{v_j} > 1 + f_- \cdot h_{v_j}$$

FRAME CONTEXT







$$h_{v_j} = \frac{1}{2T}\sum_{t=1}^{T} f_{v_{j-t}} + f_{v_{j+t}} \qquad h_{v_j} = \frac{1}{T}\sum_{t=1}^{T} f_{v_{j-t}} \qquad h_{v_j} \in \{f_{v_k} \mid k \neq j\}$$

MULTI-RESOLUTION & NEGATIVES

EVENT RETRIEVAL

TASK

$$v \rightarrow \{v_j \in \mathcal{V} \,|\, \text{event}(v) = \text{event}(v_j)\}$$

METHOD

For each $v_j \in \mathcal{V}$,

1. Uniformly sample 4 frames from v_j.
2. Compute and average the frame embeddings.

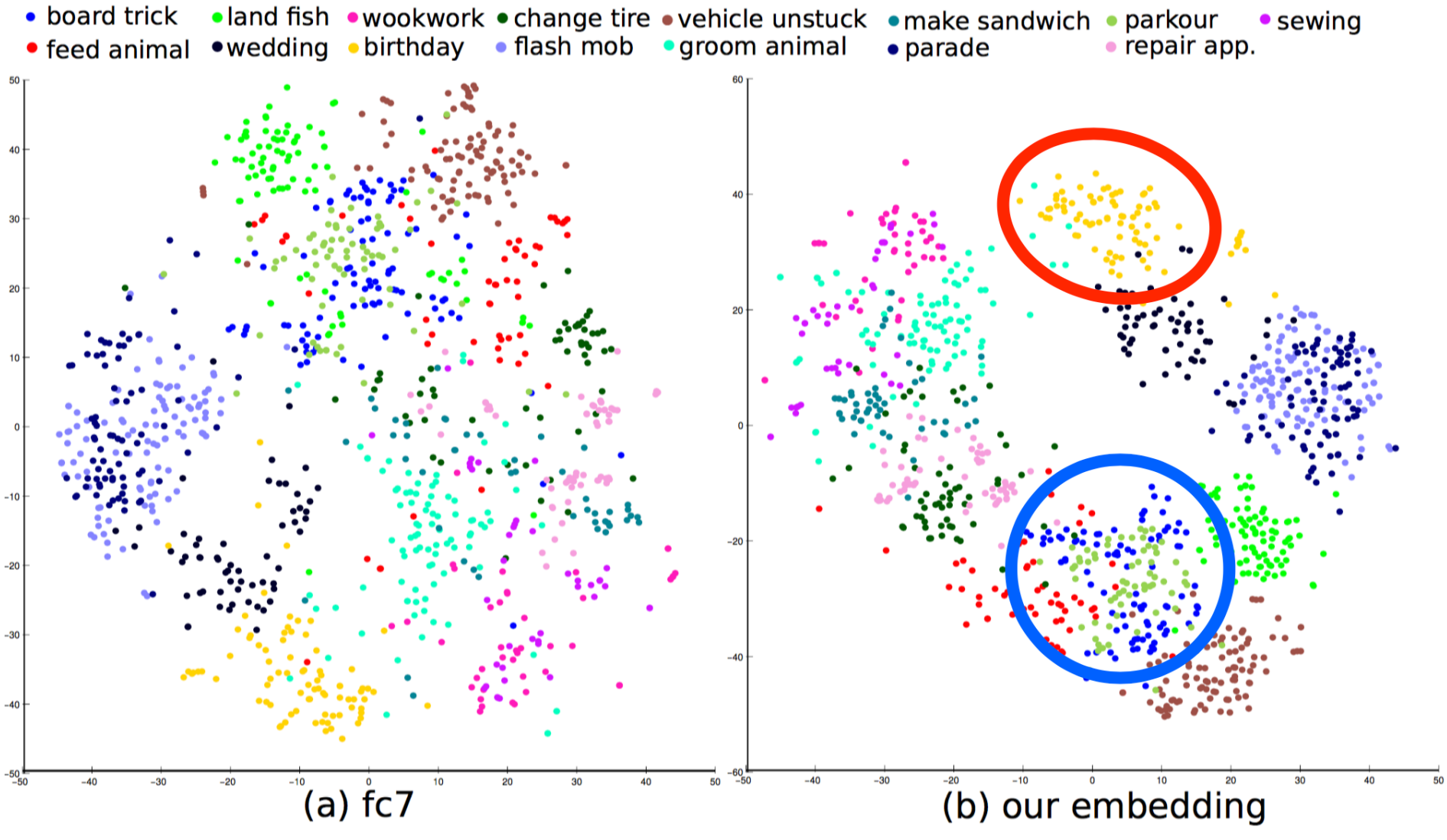
Then,

1. Sort $\left\{ \bar{f}_v \cdot \bar{f}_{v_k} \,\middle|\, v_k \neq v \right\}$

EVENT RETRIEVAL

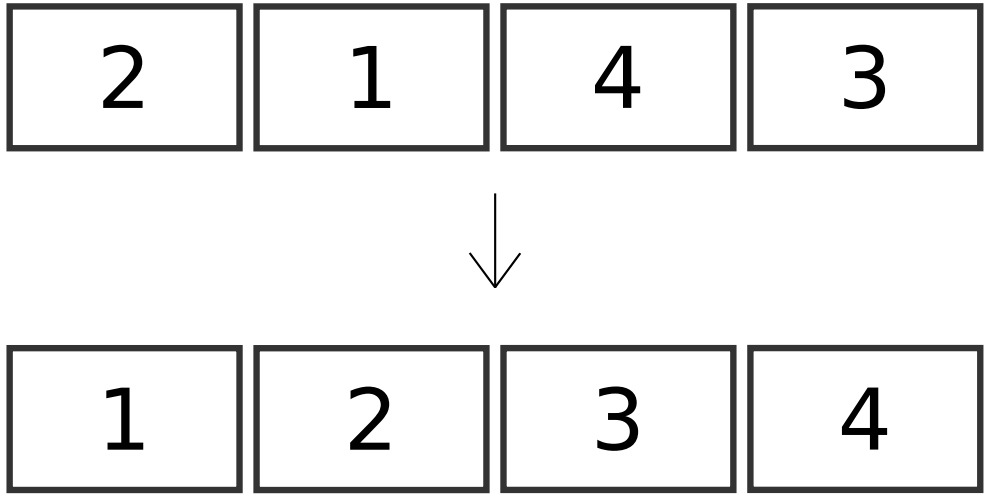
Method	mAP (%)
Chance	6.53
Two-stream pre-trained	20.09
fc6	20.08
fc7	21.24
Model (no future)	21.30
Model (no hard neg.)	24.22
Model (best)	**25.07**

EVENT RETRIVEAL



SAMPLE VIDEOS

TEMPORAL ORDER RECOVERY



TEMPORAL ORDER RECOVERY

METHOD

Given

$$\left\{ \boldsymbol{s}_{v_j} \,\middle|\, \boldsymbol{s}_{v_j} \in \boldsymbol{v}_j \right\}$$

Until done,

1. Average last two frame embeddings.
2. Find next frame as frame with highest similarity.

TEMPORAL ORDER RECOVERY

Method	Kendall Tau
Chance	50
Two-stream	42.05
fc6	42.43
fc7	41.67
Model (pairwise)	42.03
Model (no future)	40.91
Model (best)	**40.41**

TEMPORAL ORDERING FOR PHOTOS

DISCUSSION

- How are long-distance dependencies captured?
- Can we estimate the quality of embeddings independent of application?
- Hyper-parameter tuning: fps sampling, embedding dimension, negative selection, context representation

SOURCES

- Word2Vec: An Introduction
- Unsupervised Learning of Visual Representations using Videos by Nitish Srivastava
- Visualizing Data using t-SNE by van der Maaten
- Fox Over Dog Picture
- Groundhog Day, 1993, Columbia Pictures
- Efficient Estimation of Word Representations in Vector Space by Mikolov