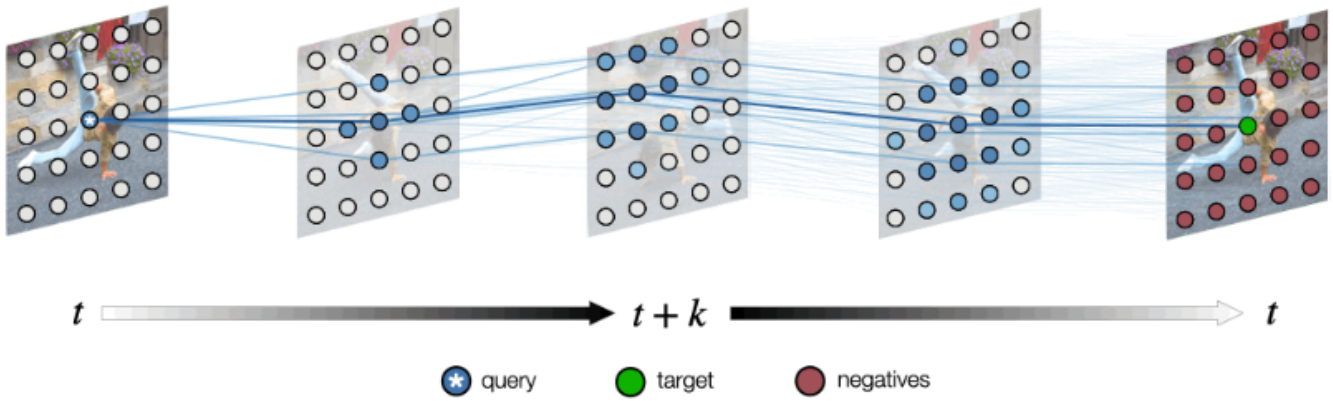
Space-Time Correspondence as a Contrastive Random Walk

Allan Jabri, Andrew Owens and Alexei Efros (2020)

Anil Keshwani @ PINLab Reading Group

March 31, 2021

Summary



- The paper proposes a self-supervised method for learning representations for visual correspondences across time: constructing palindromic video cycles.
- This can then be used for labelling e.g. of objects, semantic labels or pose keypoints.
- The authors represent video as graphs of frame patches where directed edges connect the same patches across time steps (i.e. across frames).
- They train by minimising the cross-entropy loss of nodes ending up at their starting position, which is true by construction in the palindromic video sequences.
- They use edge dropout as a regularizer and present a means of adapting the network at test-time via self-supervision.

Contrastive Random Walks on Video (1)

Video as a directed graph

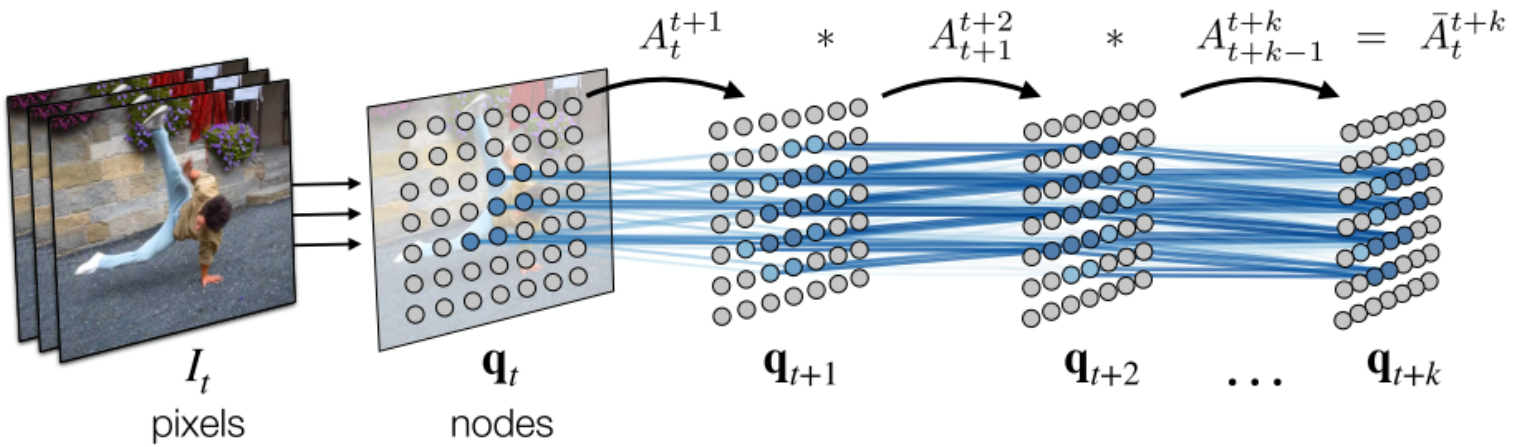
- Weighted edges connect nodes, image patches, in consecutive frames, $\mathbf{I}_t$ and $\mathbf{I}_{t+1}$
- A set of N nodes, q_t, is extracted from each frame, $\mathbf{I}_t$
- Encoder, ϕ, maps nodes to d-dimensional, l_2-normalised vectors
- Encoder yields an embedding matrix of nodes, $Q_t \in \mathbb{R}^{N \times d}$, from $\mathbf{I}_t$ (nodes are rows)
- Pairwise similarities are computed from the encoder representations: $d_\phi(q_1, q_2) = \langle \phi(q_1), \phi(q_2) \rangle$
- Non-negative "local" affinities obtained via softmax with temperature, τ, (hyperparameter) over edges departing from each node

Affinity matrices are emergent given the representations.

$$A_t^{t+1} = softmax(Q_t \cdot Q_{t+1}^T)_{ij} = \frac{exp(d_\phi(q_t^i, q_{t+1}^j)/\tau)}{\sum_{l=1}^{N} exp(d_\phi(q_t^i, q_{t+1}^l)/\tau)}; \text{ softmax row-wise}$$

Contrastive Random Walks on Video (2)

- Non-negative "local" affinities obtained via softmax with temperature, τ, (hyperparameter) over edges departing from each node
- Affinity matrix for "global" graph is composition of local stochastic matrices via a Markov chain



Long-range correspondence is multiple steps (Markov assumption enables product)

$$\bar{A}_t^{t+k} = \prod_{i=0}^{k-1} A_{t+i}^{t+i+1} = P(X_{t+k}|X_t);\ X \text{ node position r.v.}$$

Guiding the Walk with Palindromes

$$L_{CE}(\bar{A}_t^{t+k}, Y_t^{t+k}) = -\sum_{i=1}^{N} \log P(X_{t+k} = Y_t^{t+k}(i) | X_t = i)$$

- Walk can be viewed as *chain of contrastive learning problems*: maximise similarity of query and target in adjacent frames and minimise other similarities
- Longer-range correspondence: labels of t and $t+k$ provide *implicit supervision* of intermediate frames
  - Simple cases: paths do not overlap, e.g. smooth or high sampling frequency video
  - Harder cases: transition probability split across latent correspondences (paths) e.g. deformation or one-to-many matches

Self-supervision: Palindromic video sequences enable use of cycle-consistency objective. No need to infer intermediate latent views.

$$L_{cyc}^{k} = L_{CE}(\bar{A}_t^{t+k} \cdot \bar{A}_{t+k}^{t}), I) = -\sum_{i=1}^{N} \log P(X_{t+2k} = i | X_t = i)$$

Edge Dropout and Self-supervised Adaptation at Test-time

Edge dropout: Correspondence between image segments in consecutive frames (many patches/nodes to many others with similar affinity) led to use of edge dropout with rate δ to regularize the graph.

$$B_{ij} = \frac{\tilde{A}_{ij}}{\sum_l \tilde{A}_{il}} \text{ with } \tilde{A} = dropout(A, \delta)$$

$$L^k_{c\tilde{y}c} = L_{CE}(\bar{B}^{t+k}_t \cdot \bar{B}^t_{t+k}), I)$$

Test-time training: fine-tune the model parameters every 5 timesteps, applying Adam for 100 updates with input frames $\{I_{t-m}, \ldots, I_t, \ldots, I_{t+m}\}$, prior to propagating labels to I_t; i.e. $m = 10$.

Improves object propagation especially recall of the region similarity metric, which measures how often more than 50% of the object is segmented.

Pithy Implementation Details

1. **Pixels to Nodes**: 64×64 patches from 7×7 grid over 256×256 image $\rightarrow$ 49 nodes per frame.
2. **Spatial jittering**: Prevent matching on borders (robustness; learn non-trivial representations)
3. **Encoder**: `l2_norm(Linear(ResNet-18(node)))` $\rightarrow v \in \mathbb{R}^{128}$
4. **Curriculum Learning ("Shorter paths")**: Optimise sub-cycles to enforce curriculum and encourage palindromic paths (i.e. node visited at t is visited at $2k - t$
5. **Training**: Train ϕ using unlabelled Kinetics400 with Adam for $2 \cdot 10^6$ updates with learning rate 1×10^{-4}, $\tau = 0.07$ and $\delta = 0.1$ for length-10 (frame) videos resizing frames to 256×256 before encoding.

Experiments

Evaluation on video label propagation tasks: objects, keypoints and semantic "parts".

- Predict pixel-wise labels in target nodes given ground truth for first frame.
  - source nodes $\mathbf{q}_s$ with labels $L_s \in \mathbb{R}^{N \times C}$
  - target nodes $\mathbf{q}_t$
  - K_t^s the transition matrix from sources to targets where only top-k transitions retained per target node
  - **Propagated labels**: $L_t = K_t^s \cdot L_s$
- Temporal context provided as queue of last m frames*
- Efficiency: Restrict source nodes considered to spatial neighbourhood of query node (*local attention*)
- **Baseline comparators**: ResNet-18 (output of the penultimate residual block) for node embeddings
  - Pre-trained visual features: ImageNet, MoCo, VINCE
  - Based on colorization (Task-specific): CorrFlow, Mast and UVC

*for discussion

Results

Outperforms all of the listed competing self-supervised methods when using context.

Object Segmentation

- Data: DAVIS 2017, a popular benchmark for video object segmentation
- Metrics: mean and recall for *boundary alignment* and *region similarity*

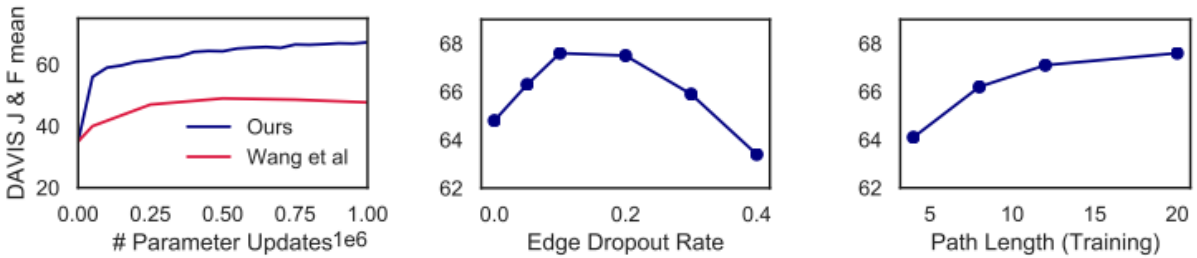
Pose Tracking

- Evaluate on JHMDB benchmark, which involves tracking 15 keypoints
- Model sees sufficiently hard negative samples from same image at training time hence learns features that discriminate beyond color
- They propagate keypoints independently (naively)

Video Part Segmentation

- Evaluate on Video Instance Parsing (VIP) benchmark (propagate part labels e.g. arm, leg, hair, shirt etc.)
- More temporal context ($m = 4$): outperforms supervised approach

Variations of the Model



(a) performance vs. training time (b) effect of edge dropout (c) effect of path length

- **Edge dropout** (simulate partial occlusion $\rightarrow$ robustness): Tried $\{0, 0.05, 0.1, 0.2, 0.3, 0.4\}$ and moderate δ yields significant improvement on DAVIS benchmark
- **Path length**: Trained with clips of length $2, 4, 6$ or 10 (path lengths $4, 8, 12$ or 20) and longer clips accelerated convergence and improved performance (DAVIS benchmark). In contrast to previous work. Authors attribute this to soft-attention mechanism which marginalises over ambiguity.
- **Improvement with training**: Downstream performance on DAVIS improves as more data is seen during self-supervised training

Closing Remarks and Setting the Work in Context (1)

Temporal correspondence: Work does contrastive data association via soft-attention, as a means for learning representations directly from pixels. By contrast to previous approaches like "tracking as repeated detection" or older ones like optical flow.
$\rightarrow$ **Future direction**: Incorporate Transformers to improve predictions across time.

Graph Neural Networks and Attention: Work uses cross-attention between nodes of adjacent frames to learn to propagate node identity through a graph. Pretext task is instance discrimination across space and time.

Graph Partitioning: Work models pixel groups ("partitions") in dynamic scenes *implicitly* so as to scale to real, large-scale video data.
$\rightarrow$ **Future direction**: Incorporate more explicit entity estimation.

Closing Remarks and Setting the Work in Context (2)

Graph Representation Learning approaches solve for distributed representations of nodes and vertices given connectivity in the graph. This work uses graph matching for representation learning, using cycle-consistency to supervise a chain of matches without inferring correspondence between intermediate pairs of graphs*.

Self-supervised Visual Representation Learning: Temporality defines natural pretext tasks in video (e.g. future prediction or motion estimation). Work implicitly determines which views to bring closer ("automatic view selection")

Self-supervised Correspondence and Cycle-consistency: $L^k_{c\tilde{y}c}$ loss is discriminative and permits association between regions that may have significant differences in their appearance (not true for colorization). Soft-attention mechanism allows for dense learning signal and marginalisation over ambiguity.

*for discussion

Thanks for your attention!