1 Summary

1. Imitation learning with access to the expert's actions and rewards is too restrictive.
2. They propose a 2-stage method that does away with these 2 assumptions by relying on unaligned footage.
3. They first learn to map unaligned videos from multiple sources to a common representation using a self-supervised objective constructed over both time (temporal distance classification) and modality (cross-model temporal classification).
4. They then embed a single Youtube video in this representation to construct a reward function that encourages an agent to imitation human gameplay.

2 Self-supervised learning of representation

1. They argue that learning from Youtube videos is difficult because of the lack of frame-by-frame alignment and the presence of domain-specific variations in color, etc.
2. They argue that by learning a common representation across different demos, their representation will generalizes to the agent's observations.
3. Because they do not have ground truth labels, they adopt self-supervision to learn this representation.
4. They propose 2 self-supervision objs: temporal distance classification (TDC) and cross-modal temporal distance classification (CMC).

2.1 Temporal distance classification

1. The task is to predict the temporal distance δt between 2 frames of a single video sequence.
2. They argue that this task requires an understanding of how visual features move and transform over time, thus encouraging an embedding that learns meaningful abstractions of environment dynamics conditioned on agent interactions.
3. They have 2 functions: an embedding fun ϕ and a classifier τ_{tdc}, trained jointly to minimize CE loss.

2.2 Cross-modal temporal distance classification (CMC)

1. The Youtube videos came with sound.
2. Since the sound usually indicate salient events, a network that learns to correlate audio and visual observations should learn an abstract that emphasizes important game events.
3. The task is to predict the temporal distance between a given video frame and an audio snippet.
4. They introduce an embedding fun for the audio snipper ψ, which maps from a frequency-decomposed audio snippet to an N-dim embedding vector, $\psi(a)$.
5. The final loss is a weighted sum:

$$L = L_{tdc} + \lambda L_{cmc}$$

2.3 Model selection through cycle-consistency

1. A challenge in their approach is how to measure the quality of the embedding ϕ.
2. They propose to use cycle consistency for this purpose.
3. Let $V = \{v_1, v_2, \dots v_n\}$ and $W = \{w_1, w_2, \dots, w_n\}$. Define d to be the Euclidean distance in the associated embedding space:

$$d_\phi\left(v_i, w_j\right) = \left\|\phi\left(v_i\right) - \phi\left(w_j\right)\right\|_2$$

4. To evaluate the quality of the embedding ϕ, they first select $v_i \in V$ and determines its nearest neighbor in W:

$$w_j = \operatorname{argmin}_{w \in W} d_\phi\left(v_i, w\right)$$

5. They then find the nearest neighbor of w_j in V:

$$v_k = \operatorname{argmin}_{v \in V} d_\phi\left(v, w_j\right)$$

6. They say that v_i is cycle-consistent iff $|i-k| \leq 1$.

7. They define the one-to-one alignment capacity P_ϕ of the embedding space ϕ as the percentage of V that is cycle-consistent.

3 One shot imitation from Youtube footage

1. Given a Youtube video, they generate a sequence of "check point" every 16 frames along the embedded trajectories. They then construct the reward function:

$$r_{\text{imitation}} = \begin{cases} 0.5 & \text{if } \bar{\phi}\left(v_{\text{agent}}\right) \cdot \bar{\phi}\left(v_{\text{checkpoint}}\right) > \alpha \\ 0.0 & \text{otherwise} \end{cases}$$

   where $\bar{\phi}(v)$ is the zero-centered and l2-normalized version of ϕ.

2. They also require that the checkpoint be visitted in soft-order, i.e. if the last collected checkpoint is at $v^{(n)}$, then $v_{\text{checkpoint}} \in \left\{v^{(n+1)}, \ldots, v^{(n+1+\Delta t)}\right\}$.