Aligning Step-by-Step Instructional Diagrams to Video Demonstrations

概要

- タスクは、aligning video segments to instructional diagrams
  - 家具の組み立て動画と、家具の組み立て説明画像(図で説明している)
  - 動画を与えた時に、対応する画像を選択肢から選択
  - diagram(説明書画像)を与えた時に対応する動画を選択
- IAW(IKEA Assembly in the Wild)というデータセット作成
- 対照学習(contrastive learning)やった
  - CLIPと全く同じことを(video, image)でやっただけ
- lossは独自のものを導入
  - 画像が時系列順になっていたりなどいろいろ特有な制約があるから

Video-Instruction Alignment

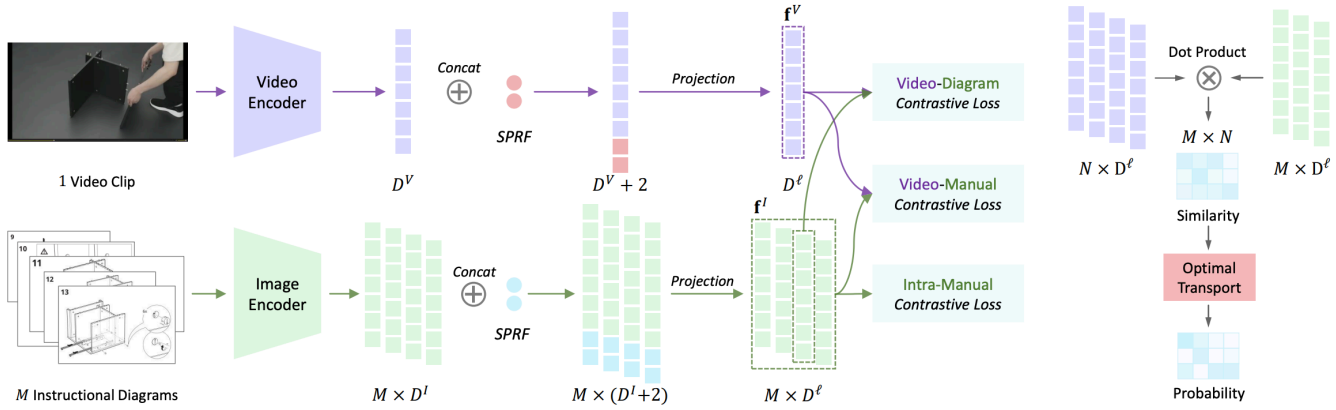
- やりたいことの定式化
- N個のvideo clip $\{V_i\}_{i=1}^N$, M個のinstructional diagrams $\{I_j\}_{j=1}^M$が与えられる
- 画像と動画をそれぞれencoderにいれて同じ次元のembedding spaceにprojectionした時のベクトルをf^V, f^Iとする
- 2つのベクトルの類似度を計算する関数をf_{sim}とする
- このとき、動画が与えられて対応する画像を選択するのは以下のように記述できる

$$j^* = \underset{j=1,\cdots,M}{\arg\max}\, f_{sim}\left(f^V, f_j^I\right)$$

- 逆に、画像が与えられて動画を選択するのは以下

$$i^* = \underset{i=1,\cdots,N}{\arg\max}\, f_{sim}\left(f_i^V, f^I\right)$$

Architecture



(a) Training stage. (b) Inference stage.

- video, imageをそれぞれvideo encoder, image encoderに入れる
- SPRFをconcatenateしてprojectionで次元を揃えて対照学習

Sinusoidal Progress Rate Feature

- positional encoding的なこと
- videoに対しては以下

$$r^V = \frac{(t_{start} + t_{end})}{2t_{duration}}$$

- j番目のimageに対しては以下

$$r^I = \frac{j}{M}$$

loss

- 確率は以下

$$p_{ij}^{V2I} = \frac{\exp(f_{\text{sim}}(\mathbf{f}_i^V, \mathbf{f}_j^I)/\tau)}{\sum_{b=1}^{B} \exp(f_{\text{sim}}(\mathbf{f}_i^V, \mathbf{f}_b^I)/\tau)}$$

$$p_{ji}^{I2V} = \frac{\exp(f_{\text{sim}}(\mathbf{f}_i^V, \mathbf{f}_j^I)/\tau)}{\sum_{b=1}^{B} \exp(f_{\text{sim}}(\mathbf{f}_b^V, \mathbf{f}_j^I)/\tau)}$$

(論文より引用)

- このとき普通の対照学習では損失以下

$$\mathcal{L}_{\text{infoNCE}} = -\frac{1}{2B}\left(\sum_{i=1}^{B}\log p_{ii}^{V2I} + \sum_{j=1}^{B}\log p_{jj}^{I2V}\right)$$

(論文より引用)

- この論文では以下の3つのlossを使う
  - Video-Diagram Contrastive Loss
  - Video-Manual Contrastive Loss
  - Intra-Manual Contrastive Loss

Video-Diagram Contrastive Loss

$$\mathcal{L}^{\text{VI}} = \frac{1}{2}\left(D_{JS}(\boldsymbol{p}^{V2I} \| \boldsymbol{q}^{V2I}) + D_{JS}(\boldsymbol{p}^{I2V} \| \boldsymbol{q}^{I2V})\right)$$

(論文より引用)

- D_{JS}はJensen-Shannon divergence

Video-Manual Contrastive Loss

$$\mathcal{L}^{VM} = \sum_{i=1}^{B} \frac{M_i}{\sum_{b=1}^{B} M_b} CE(\boldsymbol{p}_i^{V2I}, \boldsymbol{p}_i^{gt})$$

(論文より引用)

Intra-Manual Contrastive Loss

$$p_{jk}^{I2I} = \frac{\exp(f_{\text{sim}}(\mathbf{f}_j^I, \mathbf{f}_k^I)/\tau)}{\sum_{m=1}^{M} \exp(f_{\text{sim}}(\mathbf{f}_j^I, \mathbf{f}_m^I)/\tau)}$$

(論文より引用)

$$\mathcal{L}^{M} = \sum_{j=1}^{B} \frac{M_j}{\sum_{b=1}^{B} M_b} D_{JS}\left(\boldsymbol{p}_j^{I2I} \| \mathcal{N}(j, \theta)\right)$$

(論文より引用)

英単語

- cardinality : 基数、濃度
- sinusoidal : 正弦波