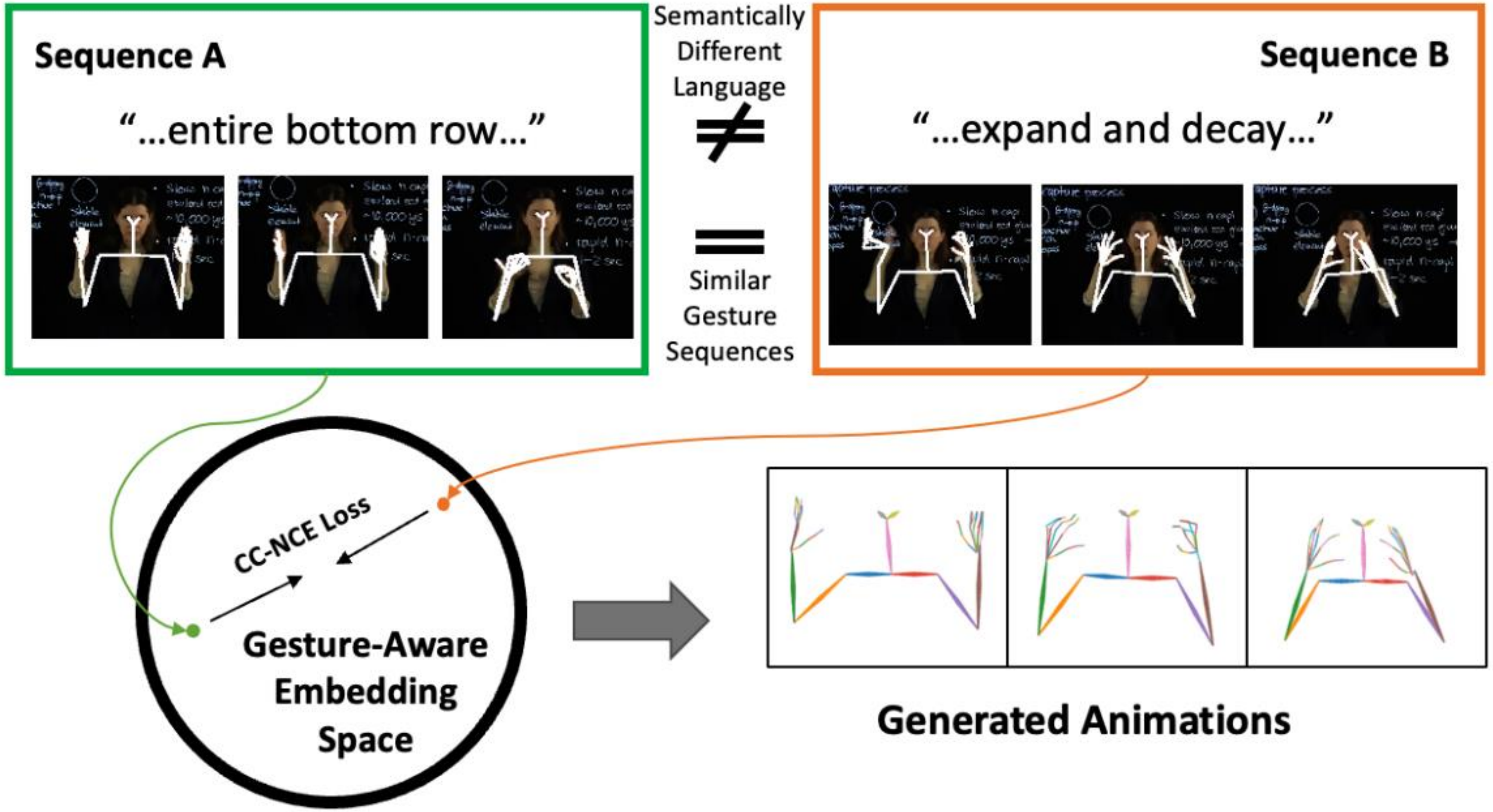
Crossmodal clustered contrastive learning: Grounding of spoken language to gesture

Ruihai Hou

2021/12/15

Crossmodal grounding



Self-supervised clustering

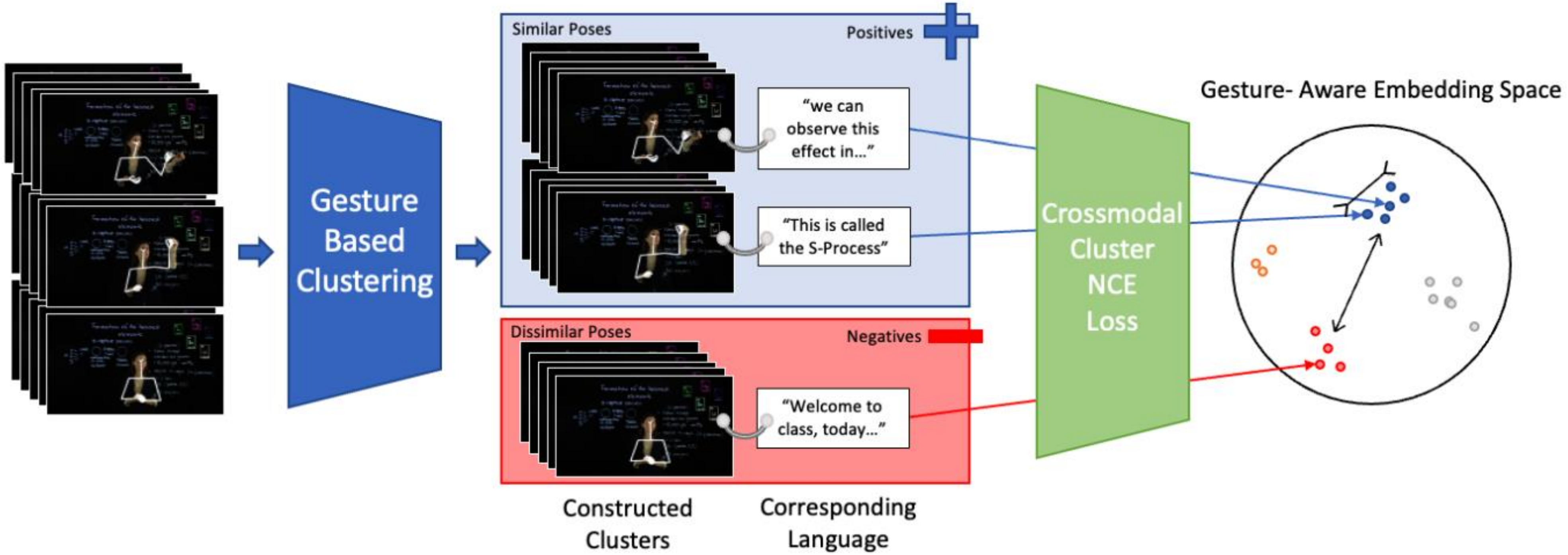


Fig. 3. Our proposed approach of self-supervised clustering in the output space of gestures, then utilizing the constructed clusters to sample negative and positives for the Crossmodal Cluster NCE loss to learn a gesture-aware language embedding space.

Threshhold

- iterate through the data and find the mean $\hat{\mu}$ and standard deviation $\hat{\sigma}$ of the pairwise dot-product similarity (referred to as Sim) of two arbitrary sequences of gestures. This metric is updated using a moving average continuously.

Batch Clustering

- Arbitrarily chosen anchor pose sequence y_a^b
- other pose sequences in the batch $y^b[\sim L]$ calculate similar score
- pose sequences whose similarity score greater than the threshold

  $(Sim(y_a^b, y^b[\sim L]) \geq \hat{\mu} + \hat{\sigma})$

  assign to a batch-wise cluster

Algorithm 1 Recursive Batch Clustering

- z^b: is the encoded audio and language representation
- y^b: corresponding ground truth pose
- $L = torch.zeroes(|B|)$: vector to check if clustered
- $Batch_D = dict()$: dictionary for batch-wise clusters
- $\hat{\mu}, \hat{\sigma}$: mean and std. dev for similarity scores
- Sim: Similarity Function
- C^b batch-wise cluster index

$a = rand(|B|)$
$C^b = 0$
while L not all True **do**
  $C^b = C^b + 1$
  $L[a] = True$
  $y_a^b = y^b[a]$
  for $idx, score$ `in enumerate`$(Sim(y_a^b, y^b[\sim L])$ `)` **do**
    if $score \geq \hat{\mu} + \hat{\sigma}$ **then**
      $Batch_D[C^b]$ append $(y^b[idx], z^b[idx])$
      $L[idx] = True$
    end if
  end for
  $dissimseq, idx = TopK(sim, 1, largest = False)$
  $a = idx$
end while
return $Batch_D$

Global Clustering

- sample sequence y^b_{samp} from the batch cluster
- sample sequences y^g_{samp} from each of global clusters
- check whether y^b_{samp} belongs in an existing cluster in global clusters using $Sim(y^b_{samp}, y^g_{samp}) \geq \hat{\mu} + \hat{\sigma}$
- exceed the threshold, merge the batch cluster to the global cluster, else create a new global cluster

Algorithm 2 Global Clustering

– $Batch_D$: dictionary for batch-wise clusters
– $Global_D$: dictionary for global clusters
– C^g: global cluster index
– $\hat{\mu}, \hat{\sigma}$: mean and std. dev for similarity scores
– Sim: Similarity Function
y^g_{samp} = sample a pose sequence per cluster from $Global_D$
for $i, values$ in $Batch_D$ **do**
  $y^b_i, z^b_i = values$ (contains aligned poses & embeddings)
  y^b_{samp} = sample a single sequence from y^b_{clus}
  for $idx, score$ `in enumerate`$(Sim(y^b_{samp}, y^g_{samp}))$ **do**
    if $score \geq \hat{\mu} + \hat{\sigma}$ **then**
      $Global_D[idx]$ append (y^b_{clus}, z^b_{clus})
    else
      $C_g = C_g + 1$
      $Global_D[C_g + 1] = (y^b_{clus}, z^b_{clus})$
    end if
  end for
end for
return $Global_D$

Crossmodal Cluster NCE

$$y_i^+, z_i^+ = argmax\ (Sim(y_{cg}^g, y_i^b)), \forall\ [y_{cg}^g, z_{cg}^g] \in Global_D)$$

$$z_i^- = \left[Global_D \backslash z_i^+\right]$$

$$L_{cc-nce} = -\mathbb{E}_z \left[\log \frac{\exp(F(z)^T F(z_c^+))}{\exp(F(z)^T F(z_c^+)) + \exp(F(z)^T F(z_c^-))}\right]$$

Experimental Results

Fig. 4. Generated keypoints superimposed on ground truth images for easy comparison. The usage of contrastive learning produces gestures closer to the ground truth (L_{MoCo}, $L_{patchwise}$, $Ours$)

Experimental Results

Model	FID ↓					
Speaker:	maher	bee	lec_cosmic	oliver	colbert	Mean
Ours	48.52 ± 5.39	100.03 ± 20.74	44.43 ± 9.71	54.06 ± 9.38	5.85 ± 0.84	50.58 ± 7.15
Without L_{cc-nce} [1]	**21.38 ± 3.89**	**65.67 ± 11.35**	**23.14 ± 11.03**	**46.48 ± 1.12**	6.77 ± 0.05	**32.69 ± 3.90**
L_{cc-nce} replaced by L_{MoCo} [19]	32.15 ± 20.83	74.892 ± 24.17	27.38 ± 16.71	48.78 ± 2.13	6.57 ± 0.16	39.66 ± 12.38
L_{cc-nce} replaced by $L_{patchwise}$ [33]	26.45 ± 3.74	70.23 ± 10.52	38.95 ± 4.02	49.47 ± 9.47	**5.48 ± 0.85**	33.30 ± 3.74

Table 2. Ablation of various contrastive loss mechanisms for 5 speakers in PATS in the task of generation of gestures in terms of coverage (FID). *Ours* utilizes the proposed L_{cc-nce} loss, whereas *Without* L_{cc-nce} utilizes no contrastive learning at all, as proposed in [1]. L_{cc-nce} is replaced by two other contrastive learning mechanisms L_{MoCo} [19] and $L_{patchwise}$ [33] for comparison.

- CC-NCE produces better L1 scores than other baselines

Conclusion

- Crossmodal Cluster NCE loss can guide the latent space to learn the similarities and dissimilarities in the constructed clusters in the gesture domain