# Homework 6

Using the assigned reading listed on the course page, answer the questions below with a short response. Note that we are looking for concise statements that show understanding, not quantity. The total discussion should roughly be a page.

## Multi-modal learning

1. Multi-modal representation learning refers to a task in which more than one sensory modality is used to learn neural representation.

   (a) To learn a meaningful representation in Harwath et al. (IJCV2019), what assumptions do they make about the relationship between audio and visual data? What does this reveal to us about the underlying representation of our world? Can you think of another task that might give rise to a better representation than aligning audio and visual data?

   (b) Now assume we use videos instead of images. What are the advantages of using videos over images to learn visual-audio joint representation?

   (c) The triplet loss used by Harwath et al. is composed of $\mathcal{L}_s$ and $\mathcal{L}_h$, explain the intuition behind these losses. Why could $\mathcal{L}_s$ be insufficient? What would happen if we removed all the negative samples and the loss is computed with only the positive samples?

## Contrastive learning

1. What is the relationship between the contrastive loss and the triplet loss? Discuss the similarities and differences between these two losses?

2. Answer the questions below after reading Wang et al. For all questions, it is fine to provide a high-level argument without all the mathematical details. However, it should be possible to extend your argument to rigorous proof. None of the questions depends on $M$, so treat it as any fixed positive integer.

   Let us consider the *ideal scenario* where the dataset can be partitioned into finitely many disjoint equal-size sets $\{A_j \colon |A_j| = L\}_{j=1}^{K}$, and the positive pairs are always from the same set. For instance, in image pre-training, each $A_j$ can be all augmentations of a single image. Following the notation in the paper, this means that $p_{\mathsf{pos}}$ is generated by first uniformly sampling an $A_j$, and then picking two i.i.d. samples from $A_j$.

   Recall that the contrastive loss is

$$\mathcal{L}_{\mathsf{contrastive}}(f;\tau,M) := \mathbb{E}_{\substack{(x,y)\sim p_{\mathsf{pos}} \\ \{x_i^-\}_{i=1}^{M} \overset{\text{i.i.d.}}{\sim} p_{\mathsf{data}}}} \left[ -\log \frac{e^{f(x)^\mathsf{T} f(y)/\tau}}{e^{f(x)^\mathsf{T} f(y)/\tau} + \sum_{i=1}^{M} e^{f(x)^\mathsf{T} f(x_i^-)/\tau}} \right]. \tag{1}$$

# Homework 6

Considering different output spaces, we may have three optimal encoders:

$$f_{\mathcal{S}}^* \in \underset{f_{\mathcal{S}}\colon \mathbb{R}^n \to \mathcal{S}^{m-1}}{\arg\min}\ \mathcal{L}_{\text{contrastive}}(f_{\mathcal{S}}; \tau, M) \qquad \text{(Fixed norm: unit sphere } \mathcal{S}^{m-1})$$

$$f_{\mathbb{R}}^* \in \underset{f_{\mathbb{R}}\colon \mathbb{R}^n \to \mathbb{R}^m}{\arg\min}\ \mathcal{L}_{\text{contrastive}}(f_{\mathbb{R}}; \tau, M) \qquad \text{(Unconstrained: } \mathbb{R}^m)$$

$$f_{\mathbb{B}}^* \in \underset{f_{\mathbb{B}}\colon \mathbb{R}^n \to \mathbb{B}^m}{\arg\min}\ \mathcal{L}_{\text{contrastive}}(f_{\mathbb{B}}; \tau, M), \qquad \text{(Bounded norm: unit ball } \mathbb{B}^m)$$

where $\mathbb{R}^n$ is the input space, and the unit ball is $\mathbb{B}^m := \{z \in \mathbb{R}^m \colon \|z\|_2 \le 1\}$.

Here we allow encoder to be *any function* that obeys the output space constraint.

(a) (Necessity of normalization). Is $f_{\mathbb{R}}^*$ well defined?
(Hint: Note that the positive pairs are from a disjoint partition. Consider an encoder that outputs a distinct vector for each partition. What happens if you scale outputs of $f^*$ by $w > 0$?)

(b) (*Optional*) (Alignment). We say an encoder $f$ is *sufficiently aligned* if it clusters features together for positive pairs. Specifically, for any $A_j$, any $x, y \in A_j$ (i.e., a positive pair), any $x^- \notin A_j$ (i.e., *not* from the same partition),

$$\|f(x) - f(y)\|_2 < \left\|f(x) - f(x^-)\right\|_2. \tag{2}$$

Show that, for any $f_{\mathcal{S}}$ or $f_{\mathbb{B}}$ that is not *sufficiently aligned*, it is not optimal w.r.t. $\mathcal{L}_{\text{contrastive}}(\ \cdot\ ; \tau, M)$ for some small $\tau > 0$.
(Hint: Consider the contrastive loss as a simple cross-entropy classification loss, where we want to linearly classify $f(x)$ from $\{f(x), f(x_1^-), f(x_2^-), \dots\}$ using the feature of the positive sample $f(y)$. If there is some "misclassification", what happens with a small enough $\tau$?)

(c) (*Optional*) (Uniformity and pariwise distances on the hypersphere). Consider any *sufficiently aligned* encoder $f_{\mathcal{S}}$ with a fixed temperature $\tau$. Suppose that with some magical power we obtain $f'_{\mathcal{S}}$ that has universally larger dot products (on distinct data pairs):

$$f'_{\mathcal{S}}(x)^{\mathsf{T}} f'_{\mathcal{S}}(y) > f_{\mathcal{S}}(x)^{\mathsf{T}} f_{\mathcal{S}}(y), \qquad \forall x \neq y. \tag{3}$$

i. What can you say about the new encoder's loss $\mathcal{L}_{\text{contrastive}}(f'_{\mathcal{S}}; \tau, M)$ versus the original $\mathcal{L}_{\text{contrastive}}(f_{\mathcal{S}}; \tau, M)$?

ii. What does this intuitively say about the distances between features on $\mathcal{S}^{m-1}$? (This need not be rigorous.)

iii. Based on your answers above, can we alternatively optimize for alignment plus expected pairwise dot-product / distances? If so, should it be large or small? If not, why?

**Submission**: Upload a PDF of your response through Canvas by **10/26 at 1pm**.