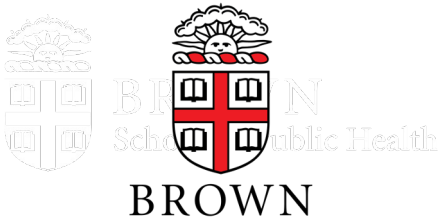
Non-Pairwise Multimodal Contrastive Loss

Michal Golovanevsky & Zhirui Li

11/6/2023

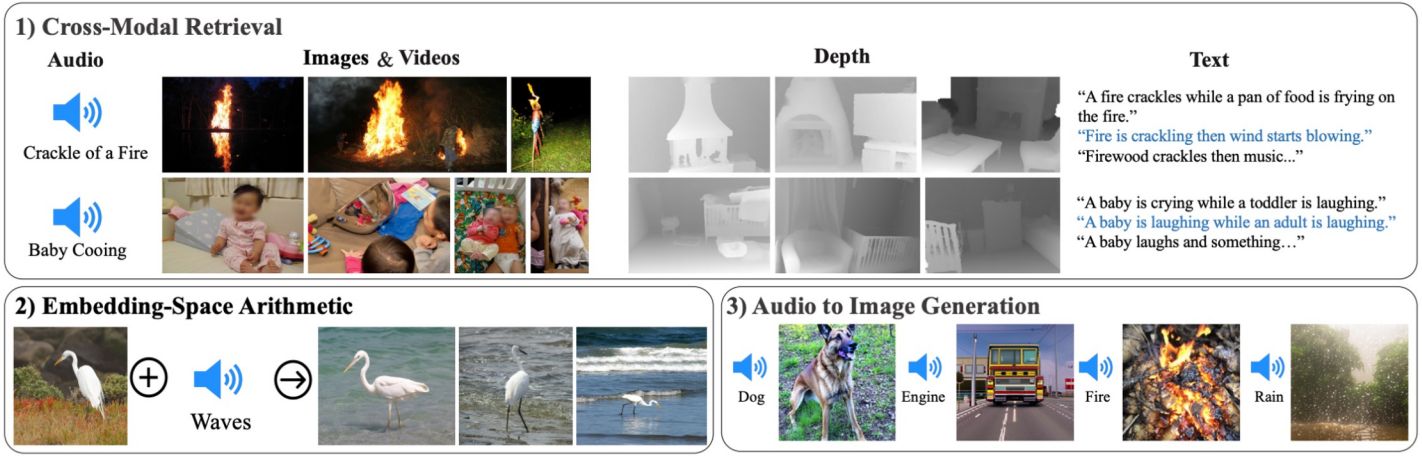


Inspiration/motivation

IMAGEBIND: One Embedding Space To Bind Them All

Rohit Girdhar* Alaaeldin El-Nouby* Zhuang Liu Mannat Singh
Kalyan Vasudev Alwala Armand Joulin Ishan Misra*
FAIR, Meta AI

https://facebookresearch.github.io/ImageBind



Model Overview (Recap)

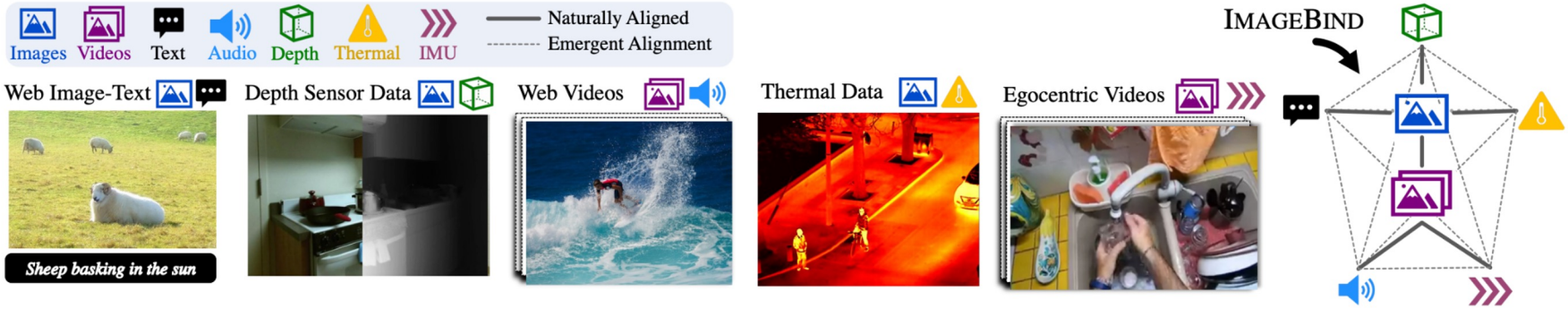


Figure 2. IMAGEBIND overview. Different modalities occur naturally aligned in different data sources, for instance images+text and video+audio in web data, depth or thermal information with images, IMU data in videos captured with egocentric cameras, *etc*. IMAGEBIND links all these modalities in a common embedding space, enabling new emergent alignments and capabilities.

Follow-up Questions

- Can we create an embedding for all modalities without relying on pairwise connections?

Follow-up Questions

- Can we create an embedding for all modalities without relying on pairwise connections?

- What is the best way to bind modalities via contrastive learning?
  - Averaging the loss across many pairwise modalities?
  - Finding the “strongest” modality and binding through it?
  - Creating a contrastive objective that is not pairwise?

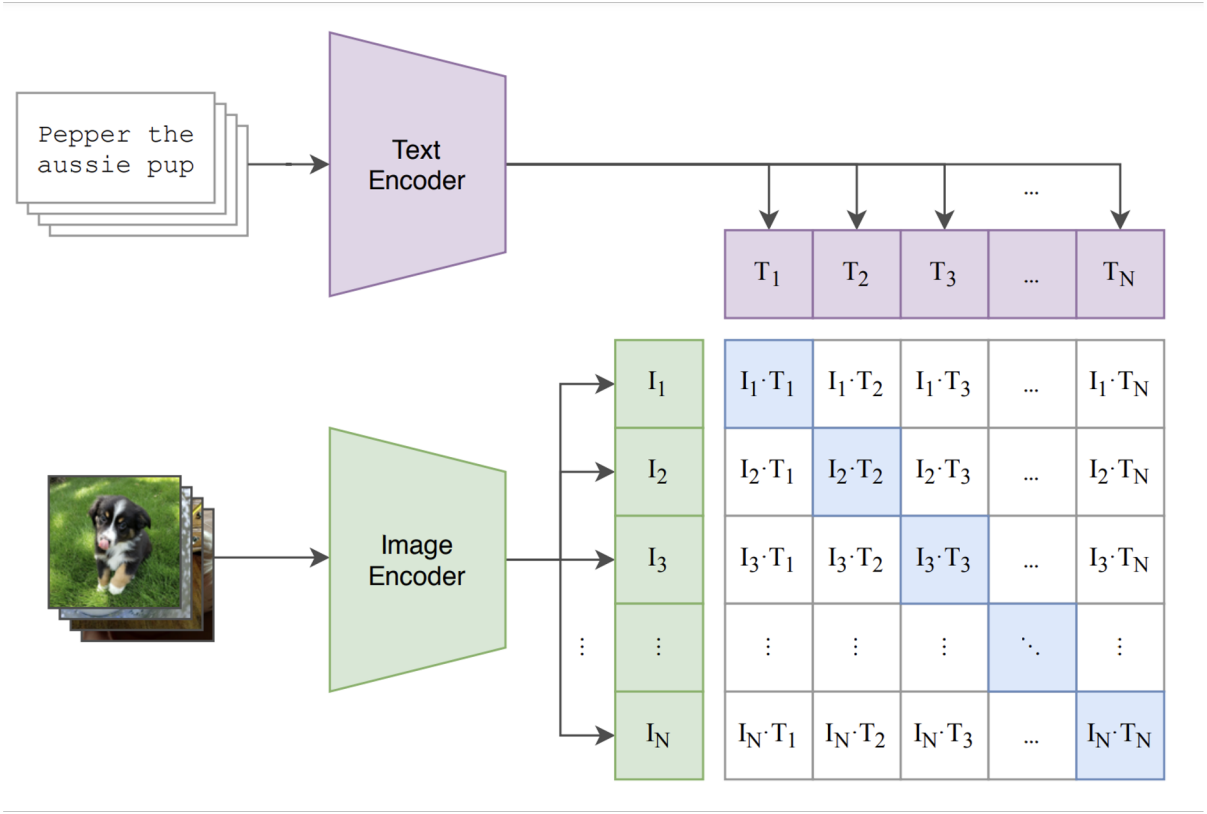
Follow-up Questions

- Can we create an embedding for all modalities without relying on pairwise connections?

- What is the best way to bind modalities via contrastive learning?
  - Averaging the loss across many pairwise modalities?
  - Finding the “strongest” modality and binding through it?
  - Creating a contrastive objective that is not pairwise?

- How much data is needed for contrastive learning to work?

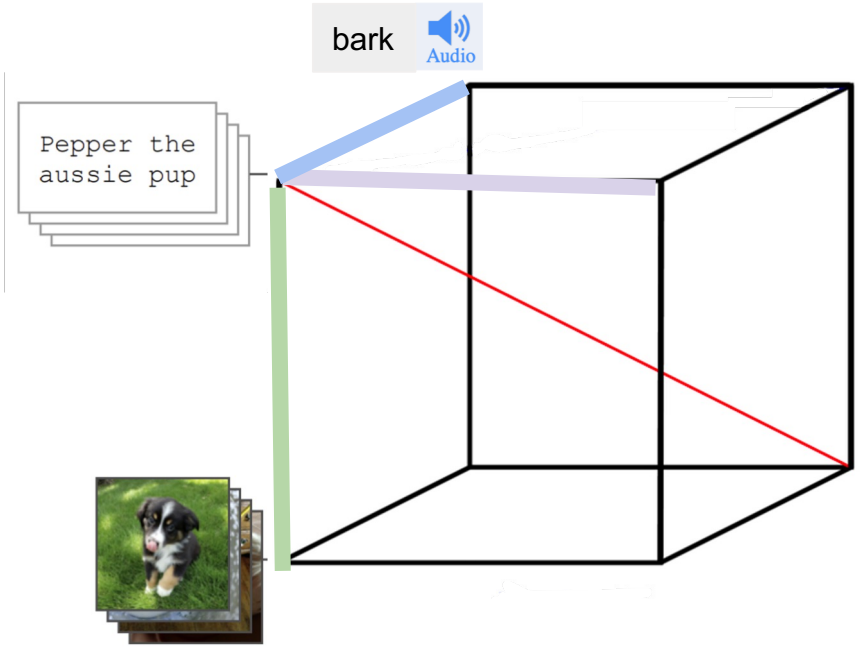
CLIP (Recap)



InfoNCE loss:

- maximises the similarity between correct pairs and minimises the similarity between incorrect pairs

Visual representation of multidimensional contrastive loss



New loss:

- maximises the similarity between correct ~~pairs~~ modality groups and minimises the similarity between incorrect ~~pairs~~ modality groups

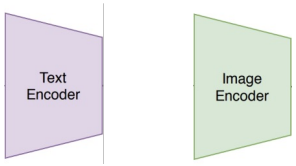
InfoNCE (from CLIP/ ImageBind)

$$L_{\text{InfoNCE}} = -\log\left(\frac{\exp(s_{\text{positive}}/\tau)}{\exp(s_{\text{positive}}/\tau) + \sum_{i=1}^{K}\exp(s_{\text{negative}_i}/\tau)}\right)$$

$$s(x_i, y_j) = \frac{f(x_i)^T g(y_j)}{\|f(x_i)\|\|g(y_j)\|}$$

InfoNCE (from CLIP/ ImageBind)

$$L_{\text{InfoNCE}} = -\log\left(\frac{\exp(s_{\text{positive}}/\tau)}{\exp(s_{\text{positive}}/\tau) + \sum_{i=1}^{K}\exp(s_{\text{negative}_i}/\tau)}\right)$$

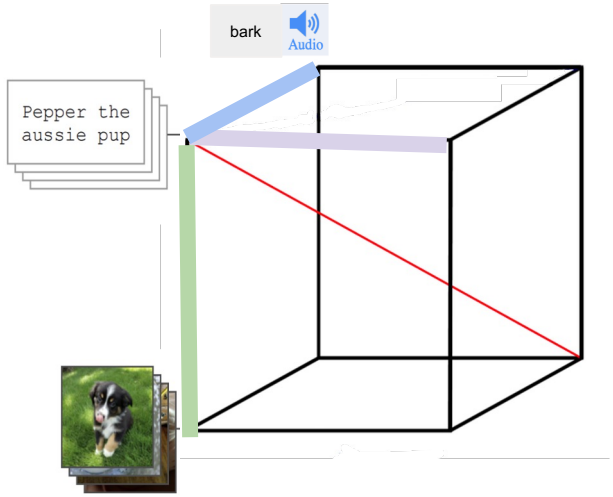


$$s(x_i, y_j) = \frac{f(x_i)^T g(y_j)}{\|f(x_i)\|\|g(y_j)\|}$$

New Loss Version 1: Using the Distance formula to Measure Similarity

Compute the central plane/line:

$$C_s = \frac{1}{N} \sum_{i=1}^{N} f_i(m_i)$$



Compute the distance from it to each modality:

$$D_{i,s} = \| f_i(m_i) - C_s \|^2$$

*Note distance captures “dissimilarity” unlike dot product

New Loss Version 1: Using the Distance formula to Measure Similarity

$$L_{\text{InfoNCE},i} = -\log \frac{\exp(-D_{i,s}/\tau)}{\sum_{k=1}^{K} \exp(-D_{k,s}/\tau)}$$

Distance formula instead of dot product

The negative examples

$$L_{\text{N-Way InfoNCE, Version 1}} = \frac{1}{N} \sum_{i=1}^{N} L_{\text{InfoNCE},i}$$

New Loss Version 1.5: Using the Distance formula to Measure Similarity

$$L_{\text{InfoNCE},i} = -\log \frac{\exp(-D_{i,s}/\tau)}{\sum_{k=1}^{K} \exp(-D_{k,s}/\tau)}$$

Distance formula instead of dot product

The negative examples

$$L_{\text{N-Way InfoNCE, Version 1.5}} = \sum_{i=1}^{N} \lambda_i \cdot L_{\text{InfoNCE},i}$$

Learned weight for each modality, instead of simple average

New Loss Version 2: Dot product with average (OvO)

Similarity function:

- One modality at a time
- Dotted with the average of all other modalities

$$s(x_j) = \frac{f_j(x_j)^T \left(\frac{1}{N-1} \sum_{i \neq j}^{N} f_i(x_i) \right)}{\|f_j(x_j)\| \left\| \frac{1}{N-1} \sum_{i \neq j}^{N} f_i(x_i) \right\|}$$

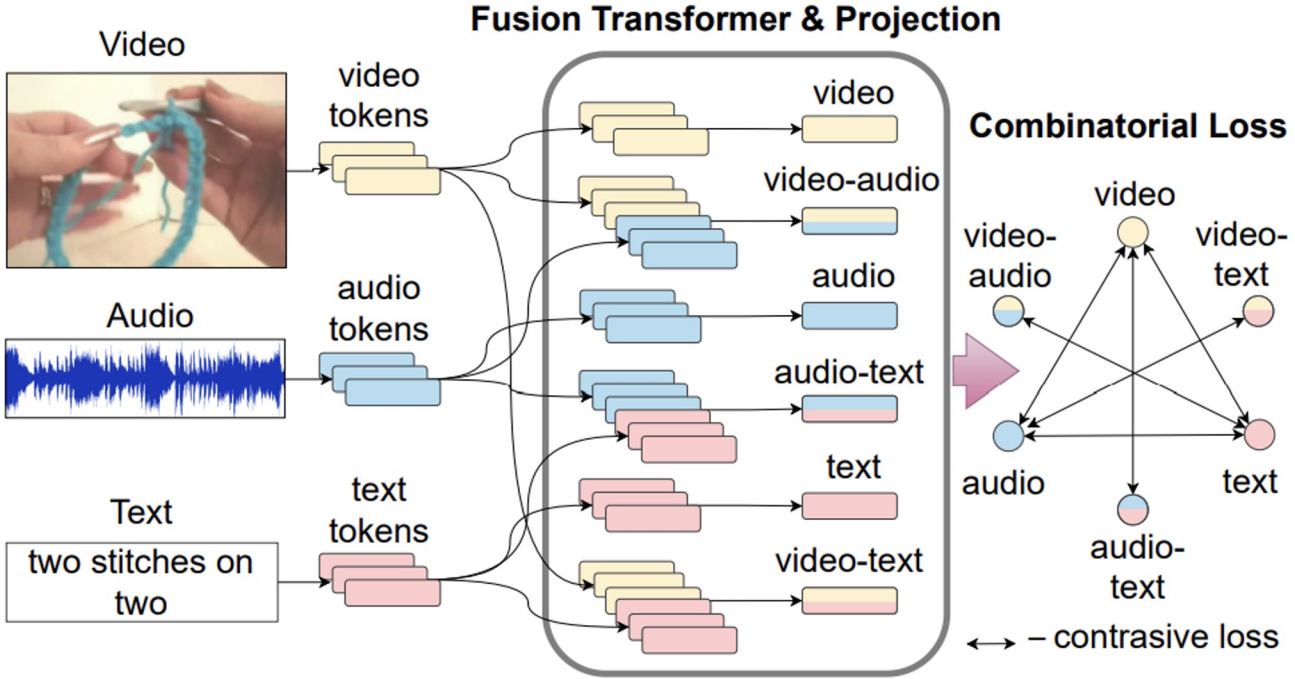
Everything else is the same as traditional InfoNCE:

$$L_{\text{InfoNCE}}(x_j) = -\log \left(\frac{\exp(s(x_j)/\tau)}{\exp(s(x_j)/\tau) + \sum_{i \neq j}^{N} \exp(s(x_i)/\tau)} \right)$$

$$L_{\text{Total}} = \frac{1}{N} \sum_{j=1}^{N} L_{\text{InfoNCE}}(x_j)$$

Other Works and Applications

"Everything At Once"



Other Works and Applications

"Everything At Once"

- Combinatorial Loss

$$\text{NCE}(x, y) = -\log\left(\frac{\exp(x^\top y/\tau)}{\sum_{i=1}^{B}\exp(x_i{}^\top y_i/\tau)}\right)$$

$$L = \lambda_{t\_v} L_{t\_v} + \lambda_{v\_a} L_{v\_a} + \lambda_{t\_a} L_{t\_a} + \\ + \lambda_{t\_va} L_{t\_va} + \lambda_{v\_ta} L_{v\_ta} + \lambda_{a\_tv} L_{a\_tv},$$

Other Works and Applications

Adaptive Contrastive Learning

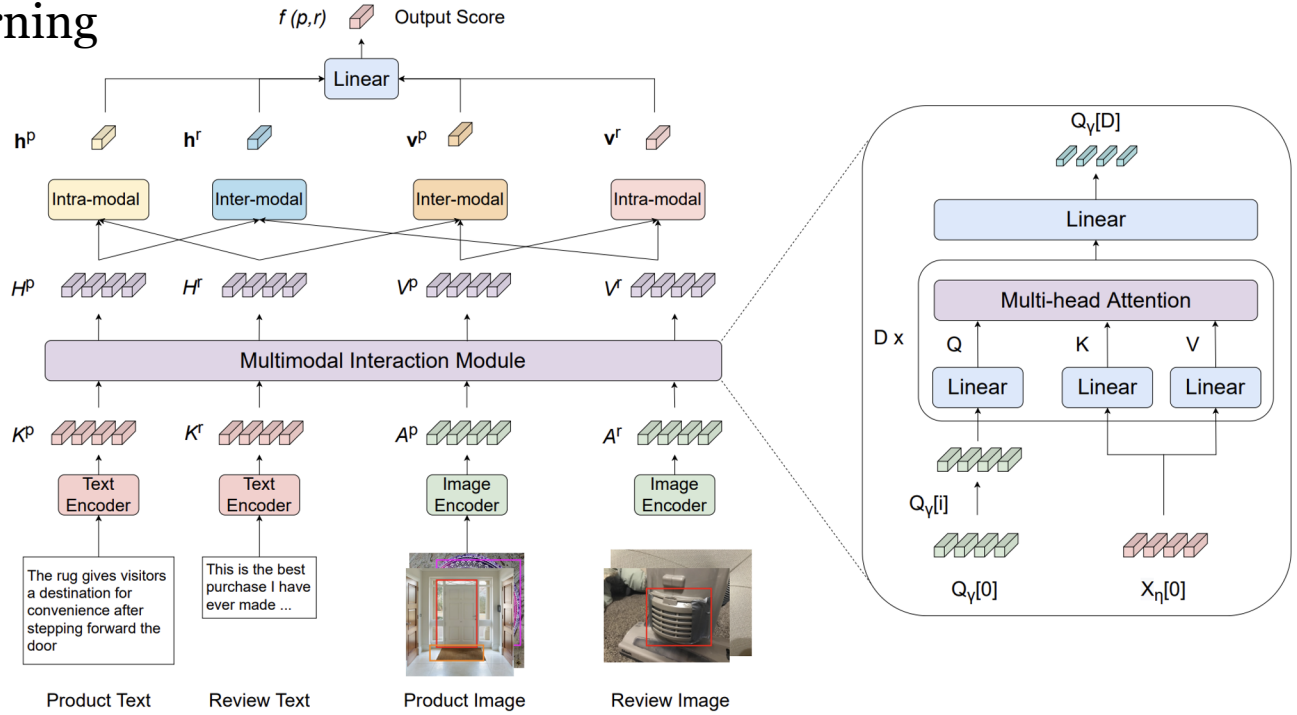


Figure 1: Diagram of our Multimodal Review Helpfulness Prediction model.

Other Works and Applications

Adaptive Contrastive Learning

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{B} \text{sim}(\mathbf{t}_i^1, \mathbf{t}_i^2) + \sum_{j=1,k=1,j\neq k}^{B} \text{sim}(\mathbf{t}_j^1, \mathbf{t}_k^2)$$

$$\mathbf{t}^1, \mathbf{t}^2 \in \{\mathbf{h}^p, \mathbf{h}^r, \mathbf{v}^p, \mathbf{v}^r\}$$

Other Works and Applications

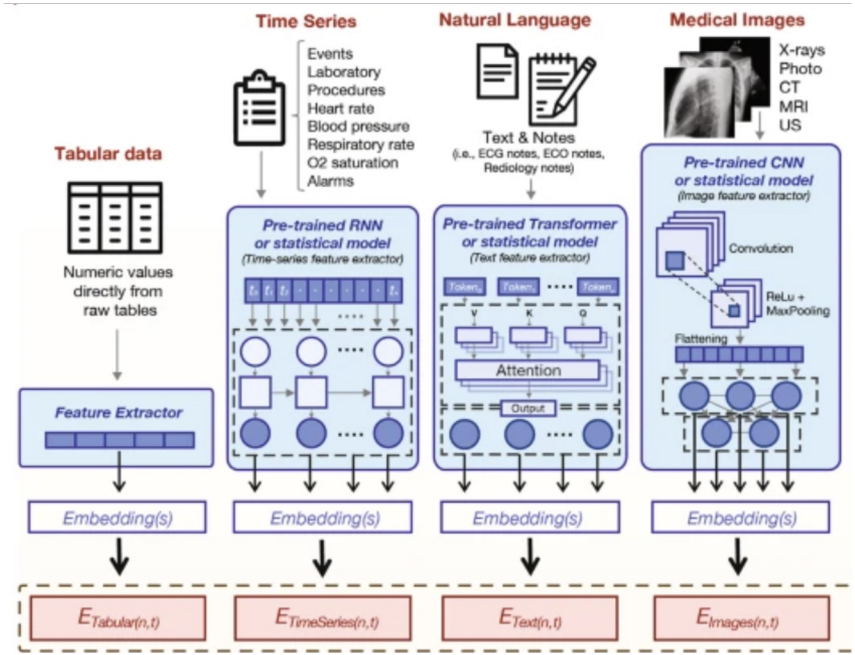
Adaptive Contrastive Learning

$$\mathcal{L}_{\text{AdaptiveCE}} = -\sum_{i=1}^{B} \epsilon_i^p \cdot \text{sim}(\mathbf{t}_i^1, \mathbf{t}_i^2) + \sum_{j=1,k=1,j\neq k}^{B} \epsilon_{j,k}^n \cdot \text{sim}(\mathbf{t}_j^1, \mathbf{t}_k^2)$$

$$\epsilon_{j,k}^n = [\text{sim}(\mathbf{t}_j^1, \mathbf{t}_k^2) - o^n]_+ \qquad \epsilon_i^p = [o^p - \text{sim}(\mathbf{t}_i^1, \mathbf{t}_i^2)]_+$$

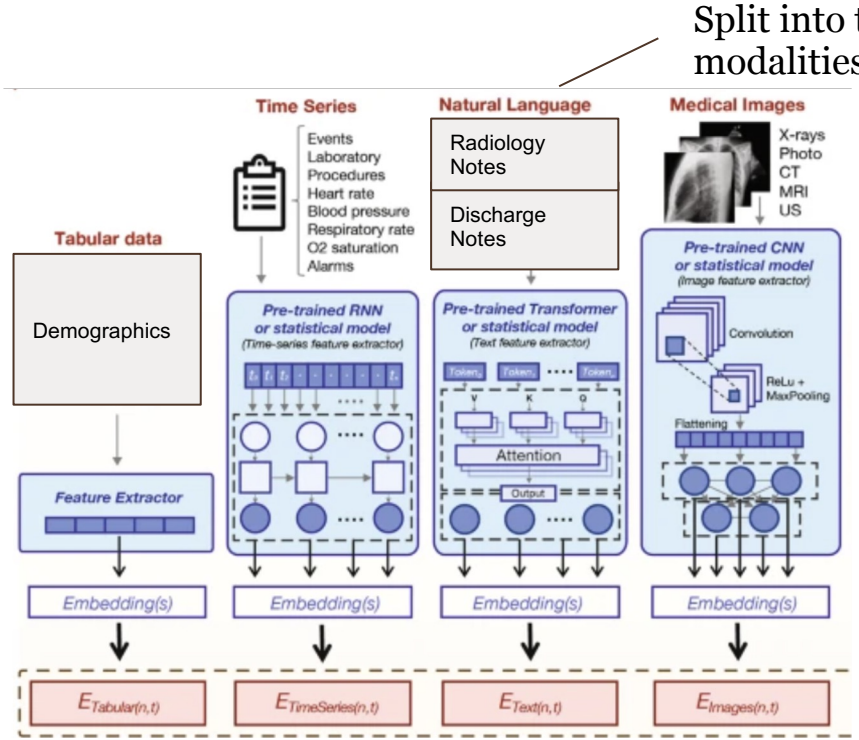
Other Works and Applications

MIMIC-CXR Dataset:



Other Works and Applications

MIMIC-CXR Dataset:



Outline of Experiments

- Baselines:
  - Concatenation model:
    - pre-trained embeddings + linear layers + concat
    - For applications like MIMIC with no good baselines

Outline of Experiments

- Baselines:
  - Concatenation model:
    - pre-trained embeddings + linear layers + concat
    - For applications like MIMIC with no good baselines
  - Original models presented in the papers
    - If needed, trained on a smaller subset of the data

Outline of Experiments

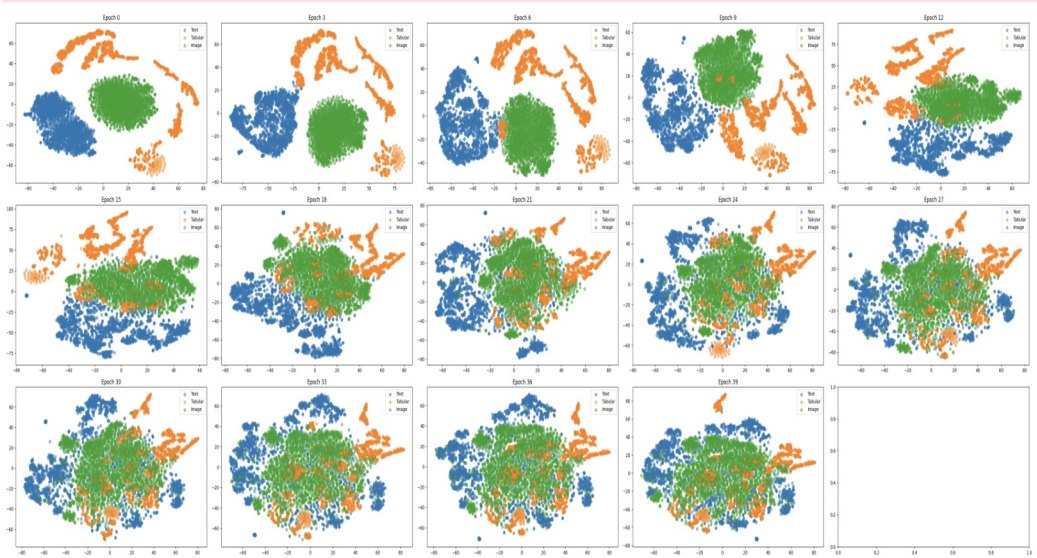
- Baselines:
  - Concatenation model:
    - pre-trained embeddings + linear layers + concat
    - For applications like MIMIC with no good baselines
  - Original models presented in the papers
    - If needed, trained on a smaller subset of the data
  - InfoNCE for all pairs of modality combinations

Outline of Experiments

- Baselines:
  - Concatenation model:
    - pre-trained embeddings + linear layers + concat
    - For applications like MIMIC with no good baselines
  - Original models presented in the papers
    - If needed, trained on a smaller subset of the data
  - InfoNCE for all pairs of modality combinations

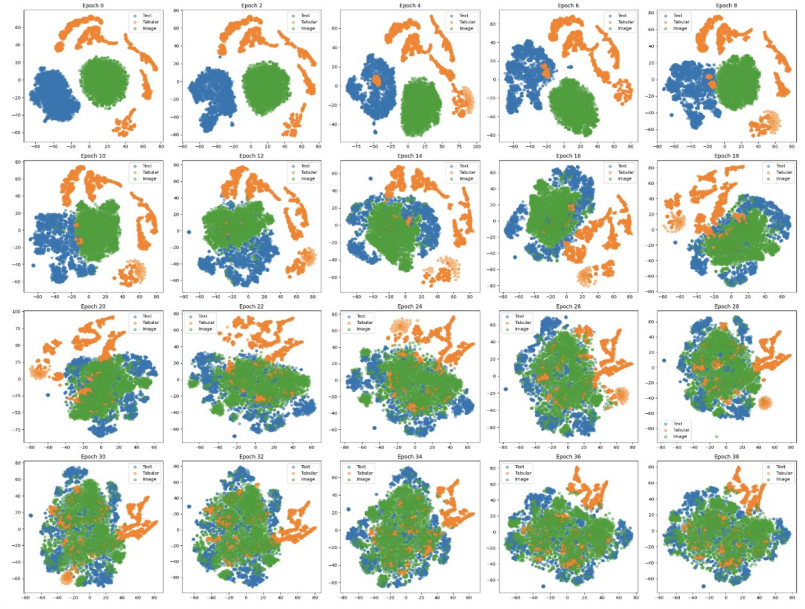
- Our new method:
  - Distance formula
    - Plus/minus learned parameter for weight instead of average
  - OvO formula
    - Plus/minus learned parameter for weight instead of average

Preliminary Results on Amazon Reviews dataset

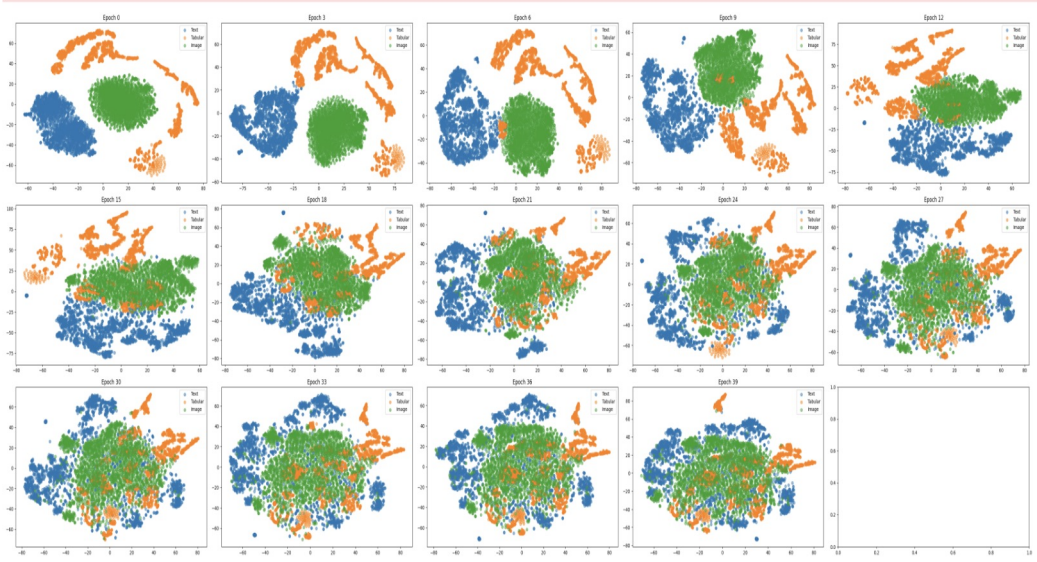


OvO

infoNCE



Preliminary Results on Amazon Reviews dataset

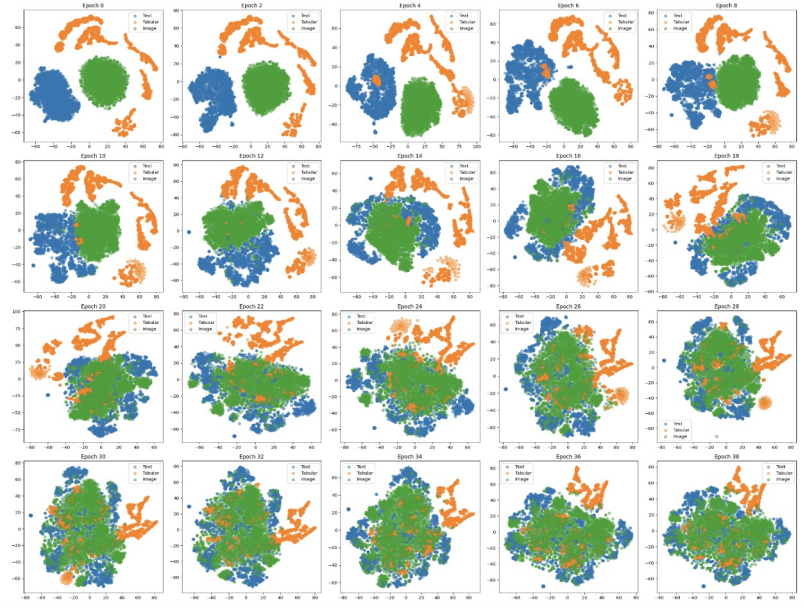


OvO

- Silhouette Score of 0.0002

infoNCE

- Silhouette Score of 0.0003



Thank you! Questions? Suggestions?