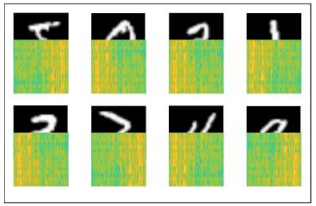
CLARGA: Multimodal Graph Representation Learning over Arbitrary Sets of Modalities

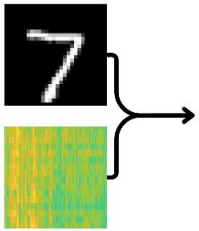
Research Goals

I want to classify numbers using images of handwritten digits and spoken pronunciations...



Training Dataset

CLARGA

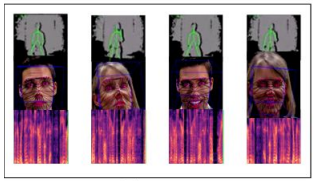


Multimodal Input

CLARGA

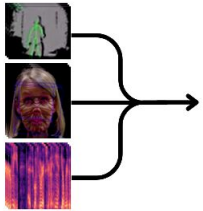
Accurate Classification

I want to diagnose depression from clinical interview recordings using pose, face, and speech features...



Training Dataset

CLARGA



Multimodal Input

CLARGA

ND

Accurate Classification

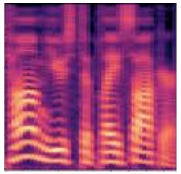
Achieve high accuracy and speed for any given task w/o changing CLARGA's underlying architecture. All we should need is a training dataset.

An Overview of CLARGA

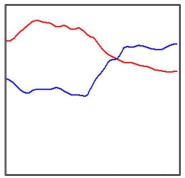
Modality Encoders

→ Vision Transformer → H_1 (Image)

"Apple Tree" → Language Transformer → H_2 (Text)



→ Fully Connected Layers → H_3 (Audio)



→ 1D Convolutions + LSTM → H_4 (Sensor)

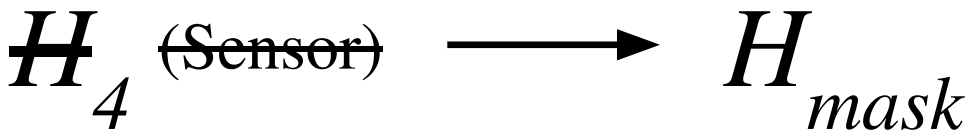
Handling Missing Modalities

H_1 (Image)

H_2 (Text)

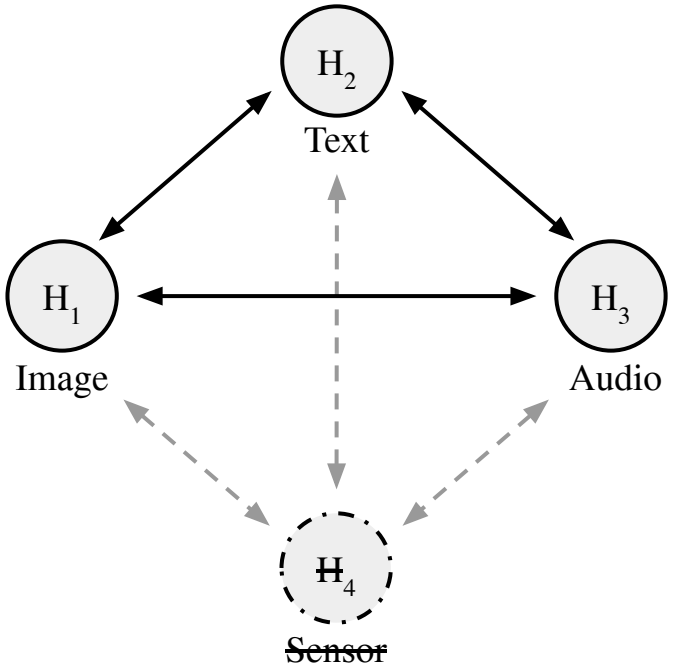
H_3 (Audio)

~~H_4 (Sensor)~~

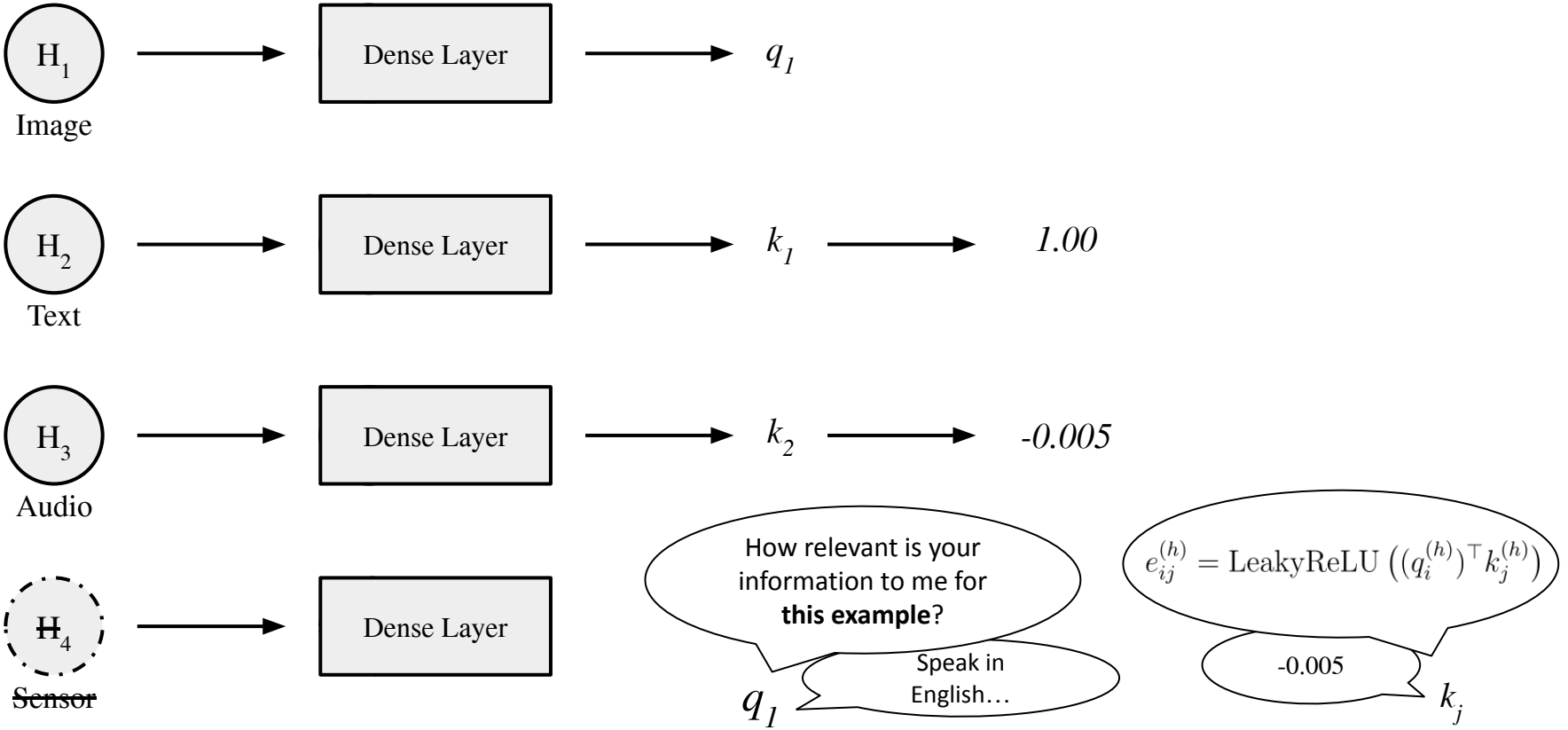


- Replace missing modalities w/ a mask
- Optimize mask throughout training to minimize loss of performance

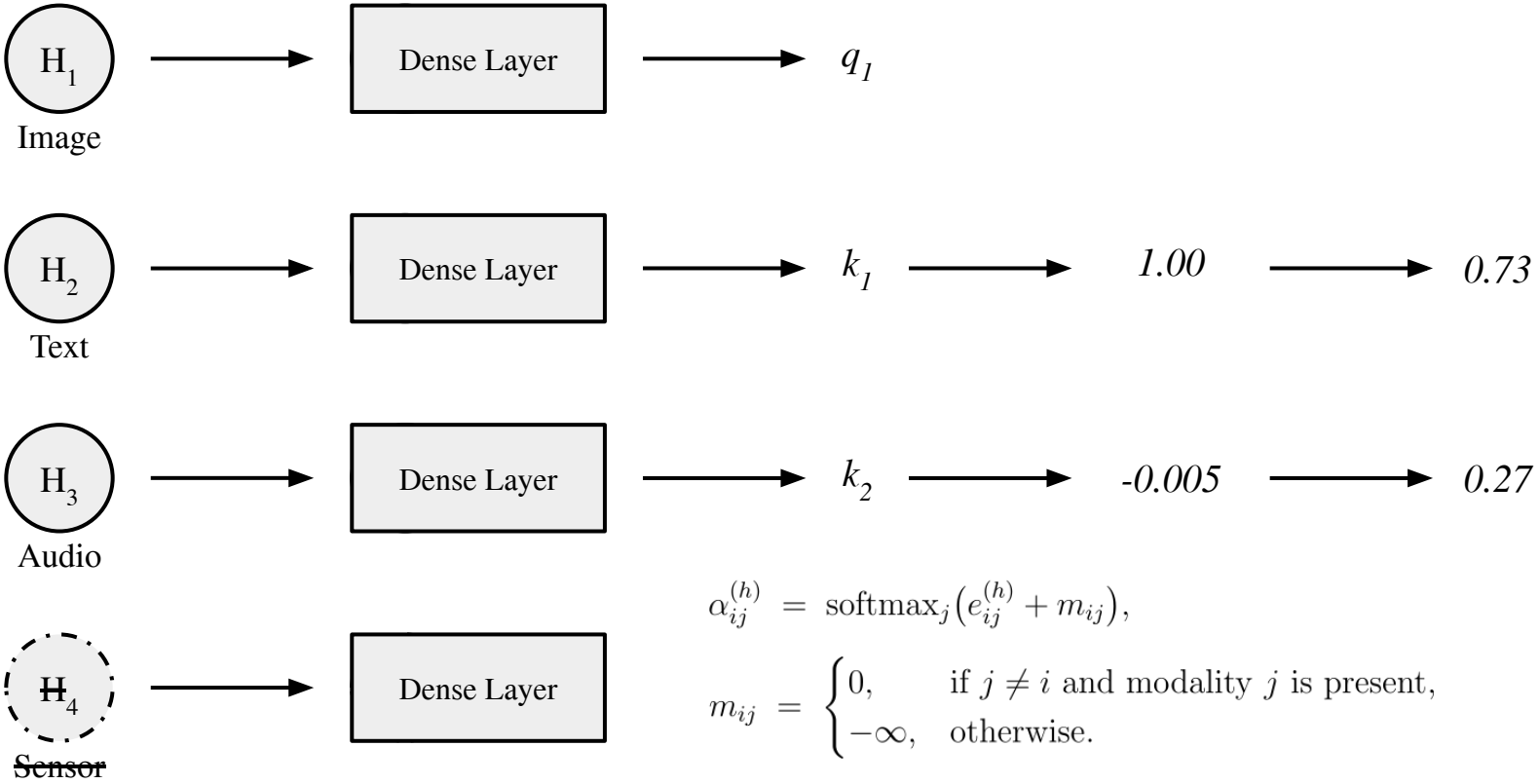
Constructing a Graph



Adaptive Graph Attention



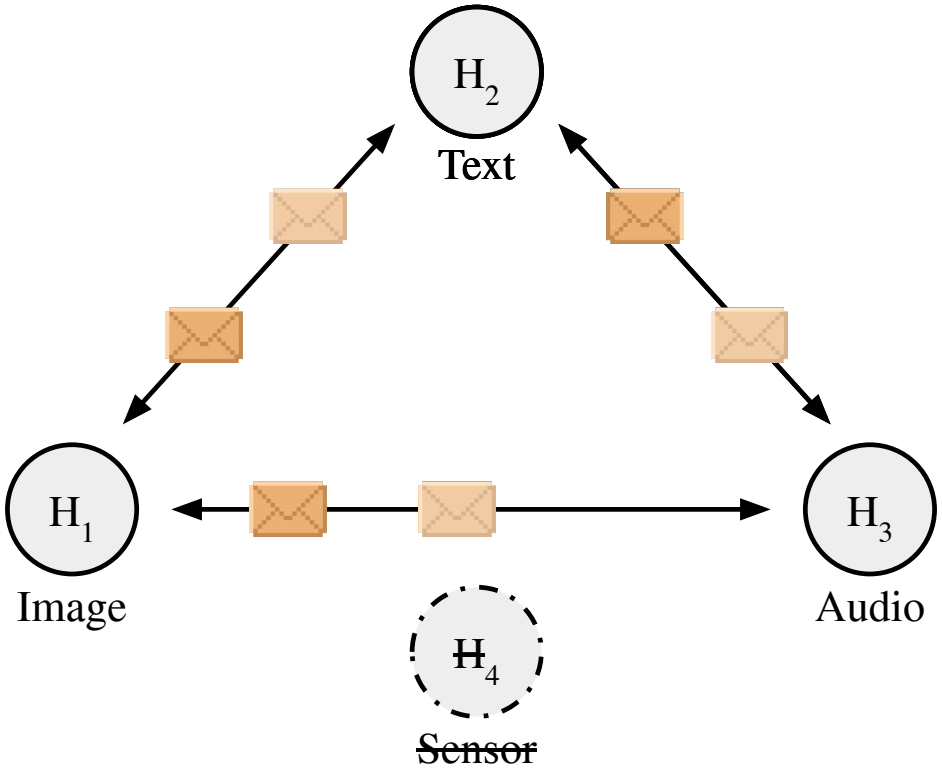
Adaptive Graph Attention



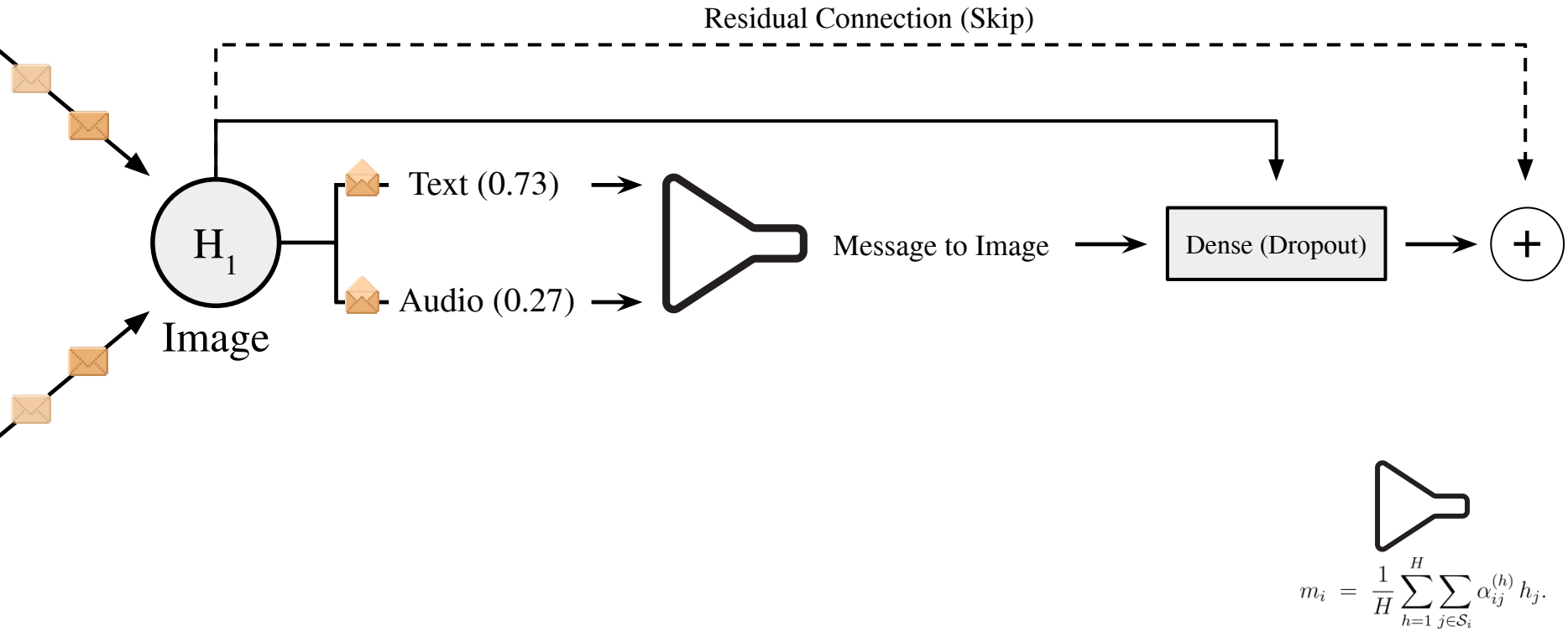
$$\alpha_{ij}^{(h)} = \mathrm{softmax}_j\left(e_{ij}^{(h)} + m_{ij}\right),$$

$$m_{ij} = \begin{cases} 0, & \text{if } j \neq i \text{ and modality } j \text{ is present,} \\ -\infty, & \text{otherwise.} \end{cases}$$

Residual Graph Attention Layers (GAT)

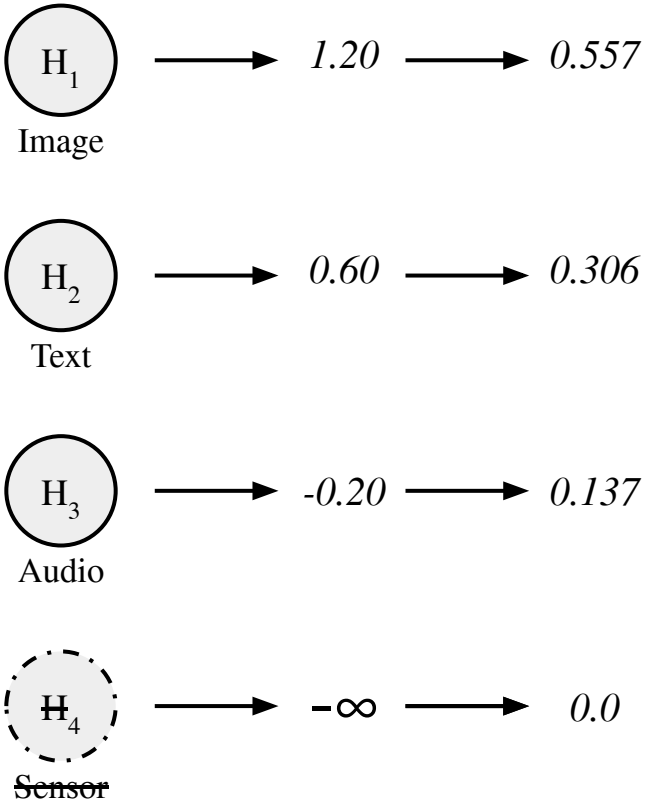


Residual Graph Attention Layers (GAT)



$$m_i = \frac{1}{H}\sum_{h=1}^{H}\sum_{j\in\mathcal{S}_i}\alpha_{ij}^{(h)}\, h_j.$$

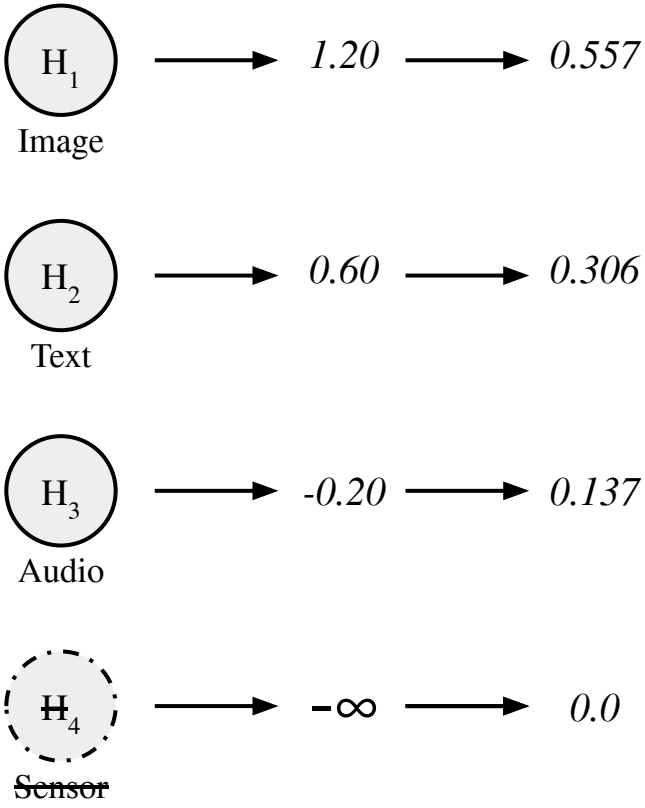
Modality Fusion



Overall, how relevant are each of you for **this example**?

q_F

Modality Fusion



$$\beta_i = \frac{\exp(s_i)}{\sum_{\ell \in \mathcal{S}} \exp(s_\ell)}$$

Dense Layer → Output

Hybrid Objective

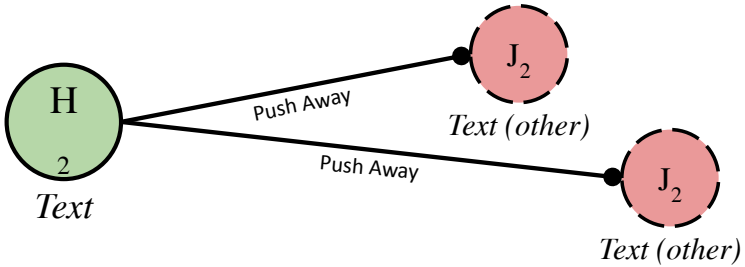
Two Goals:

- Minimize task loss
- Minimize contrastive alignment loss

Task Loss

- Error between model's prediction and the true label
- E.g., Cross-Entropy for classification

Contrastive Alignment Loss (InfoNCE)



Push away other examples' embeddings,
they should be more distinct