Fusion Methods for Multimodal Biological Embeddings: A Project Report

Student Name

1 Introduction

In this project, we implement and test three multimodal fusion methods for integrating DNA, RNA, and protein embeddings extracted from pretrained foundation models. These modalities represent different levels of biological information, and their fusion aims to improve downstream molecular property prediction.

We follow the methodology proposed in the BioLangFusion framework, which includes two major steps:

1. **Codon-level alignment** of DNA, RNA, and protein embeddings.
2. **Fusion of aligned embeddings** using:
   - Concatenation Fusion
   - Multiple-Instance Learning (MIL) Fusion with Gated Attention
   - Cross-Modal Multihead Attention Fusion

Synthetic embedding tensors were used for testing the implementations.

2 Modality Alignment

The pretrained language models used in practice tokenize sequences at different biological granularities. For example:

- DNA embeddings: 6-mers
- RNA embeddings: 1-mer (single nucleotides)
- Protein embeddings: amino acids, derived from nucleotide triplets

To fuse them effectively, all embeddings must be aligned to the same biological unit. We follow the paper's design and choose the **codon** (3 nucleotides) as the shared resolution.

Let an RNA sequence have length T. Protein length is:

$$T' = \frac{T}{3}.$$

Alignment is performed as follows:

$$\tilde{E}_{DNA} = \mathrm{TConv}_{k=2,s=2}(E_{DNA}), \qquad \tilde{E}_{RNA} = \mathrm{AvgPool}_{k=3,s=3}(E_{RNA}),$$

yielding:

$$\tilde{E}_{DNA}, \tilde{E}_{RNA}, E_{Prot} \in \mathbb{R}^{T' \times d_m}.$$

All fusion methods described below assume aligned embeddings.

3 Concat Fusion

Concat Fusion is the simplest fusion strategy. At each aligned codon position, the modality embeddings are concatenated:

$$Z_{\text{concat}}(t) = \mathrm{MLP}(\tilde{E}_{DNA}[t]) \parallel \tilde{E}_{RNA}[t] \parallel E_{Prot}[t].$$

Because DNA embeddings often have much higher dimensionality, a projection MLP is applied beforehand to avoid dominance of one modality.

The final fused representation is:

$$Z_{\text{concat}} \in \mathbb{R}^{T' \times (d'_{DNA} + d_{RNA} + d_{Prot})}.$$

4 Multiple-Instance Learning (MIL) Fusion

Instead of treating modalities equally, MIL Fusion learns modality-level attention weights.

Each modality is first projected to a shared latent dimension:

$$H_m = W_m \tilde{E}_m.$$

Then, a mean-pooled summary vector is computed:

$$\bar{h}_m = \frac{1}{T'} \sum_{t=1}^{T'} H_m[t].$$

Attention weights use a gated network:

$$\alpha_m = \frac{\exp\left(W^\top \left[\tanh(V_m \bar{h}_m + b_m) \odot \sigma(U_m \bar{h}_m + c_m)\right]\right)}{\sum_{i \in \{\text{DNA,RNA,Prot}\}} \exp(\cdot)}.$$

The fused output is:

$$Z_{\text{MIL}} = \sum_{m \in \{\text{DNA, RNA, Prot}\}} \alpha_m H_m.$$

An entropy regularization term is optionally added:

$$H_{\text{attn}}(\alpha) = -\sum_m \alpha_m \log \alpha_m.$$

5 Cross-Modal Multihead Attention Fusion

While previous methods treat modalities independently at each position, this method allows modalities to attend to one another.

All embeddings are projected to shared dimension d:

$$H_m = W_m \tilde{E}_m.$$

Then a global context is formed:

$$C = [H_{DNA}; H_{RNA}; H_{Prot}] \in \mathbb{R}^{3T' \times d}.$$

Each modality performs multihead attention over this context:

$$Z_m = g(\text{MultiHead}(H_m W_Q^m,\ CW_K^m,\ CW_V^m)).$$

Finally:

$$Z_{\text{fused}} = \text{LayerNorm}\left(\frac{Z_{DNA} + Z_{RNA} + Z_{Prot}}{3} + [Z_{DNA} \| Z_{RNA} \| Z_{Prot}] W_o\right).$$

Cross-modal attention enables discovery of subtle dependencies across modalities.

6 Testing Methodology

To validate implementation correctness, we created synthetic embeddings and applied a 256-dimensional projection head:

```
dna_z = proj_dna(dna_emb)
rna_z = proj_rna(rna_emb)
prot_z = proj_prot(protein_emb)
```

Testing for each fusion module was performed by instantiating the correct constructor:

```
fusion = FusionConcat(256, 256, 256, 256)
fusion = FusionMIL(256, 256, 256, 256, 128)
fusion = FusionCrossAttention(256, 256, 256, 256, 4)
```

Output shapes matched the expected

$$(200, 256)$$

for all methods.

7 Conclusion

We successfully implemented the three multimodal fusion methods described in the BioLang-Fusion framework: Concat Fusion, MIL Fusion, and Cross-Modal Attention. Our tests confirm correct input–output behavior and consistency with the theoretical formulations.

Each method offers different trade-offs:

- Concat: simple, high-dimensional, no weighting.
- MIL: modality weighting via attention.
- Cross-Attention: rich cross-modality interactions.

These modules are now ready to be integrated into downstream prediction tasks.