Static Interpreter Structure

Inputs:
- base model weights and gradients
- dataset feature name embeddings
- cosine similarity matrix between all feature embeddings

Output:
- cosine similarities between each feature embedding and target embedding