

CAFA 5 Protein Function Prediction

Solution Introduction



Contents

1

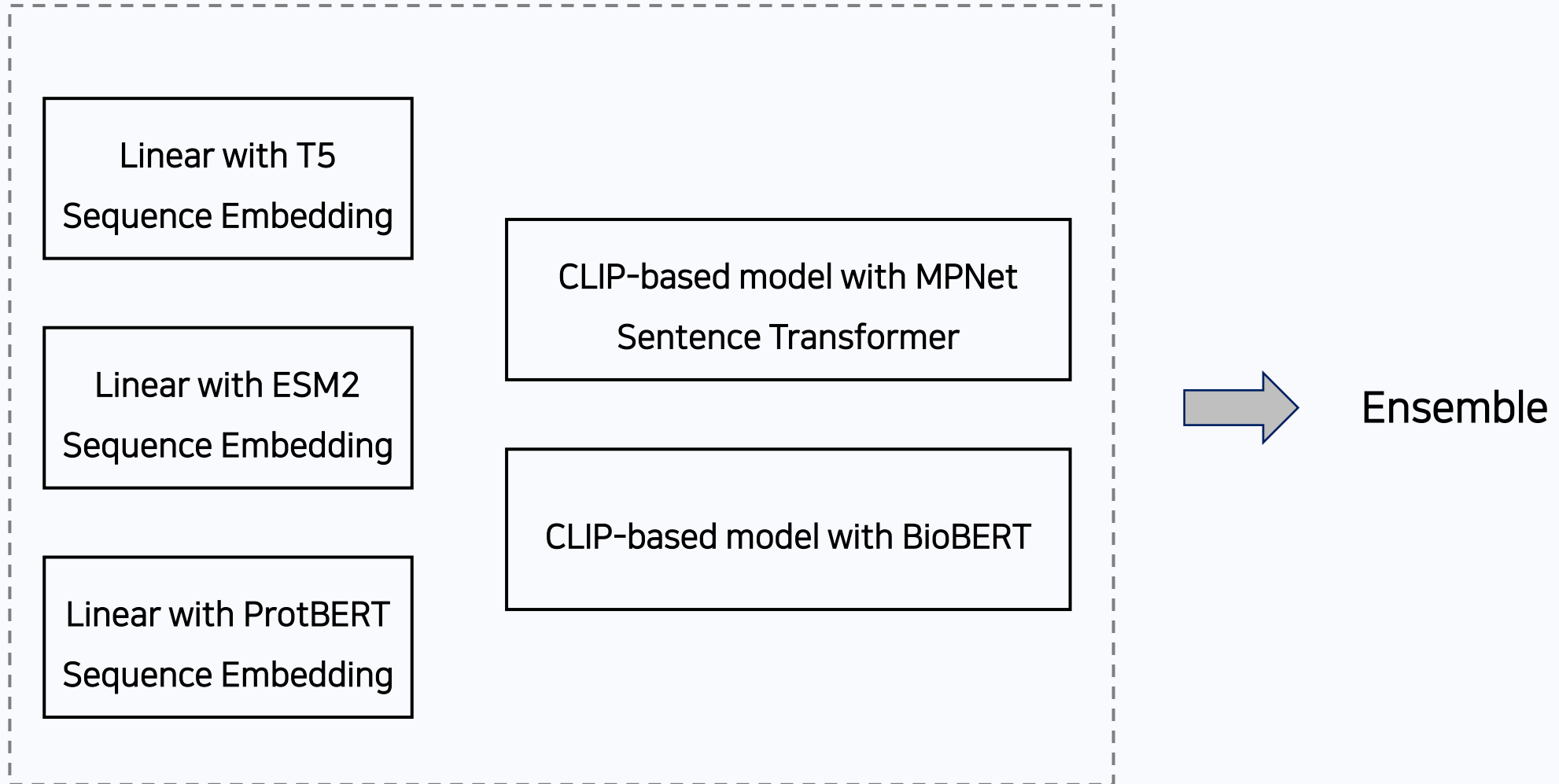
Data Engineering

2

CV Strategy

3

Architecture



Contents

1

Data Engineering

2

CV Strategy

3

Architecture

Sequence-side Features

1. T5 / ESM2 / ProtBERT feature vector
2. length of sequence
3. protein structure feature
4. mean & std. of amino acid property feature
5. ratio of each amino acids in sequence
6. ratio of each amino acids' group in sequence
7. taxonomic identifier

Sequence-side Features

1. T5 feature vector
2. length of sequence
3. protein structure feature
4. mean & std. of amino acid property feature
5. ratio of each amino acids in sequence
6. ratio of each amino acids' group in sequence
7. taxonomic identifier

GO Term-side Features

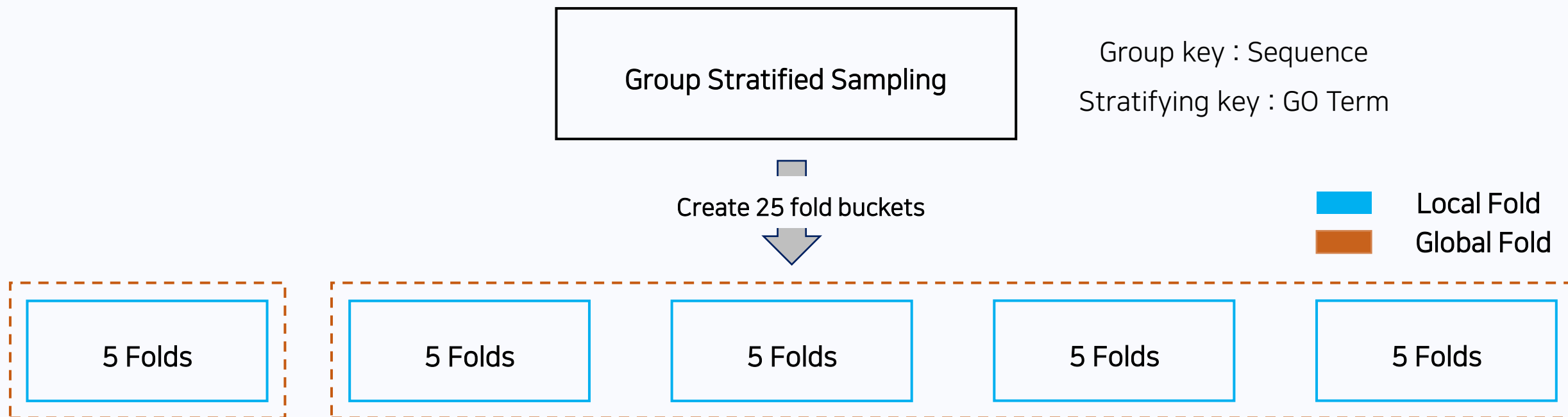
1. MPNet(ST) / BioBERT feature vector
2. GO Term type
3. word2vec embedding
4. GNN embedding

Contents

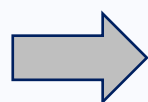
1 Data Engineering

2 CV Strategy

3 Architecture



Linear-based model uses Local Fold & CLIP-based model use Global Fold



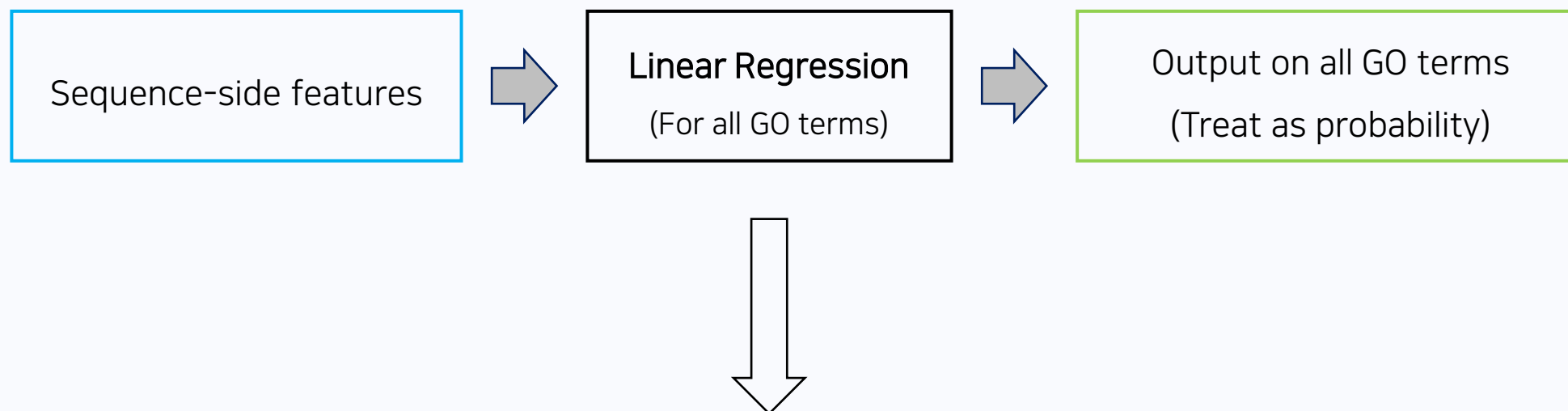
Twin CV allow to train linear-based model with data representing same distribution as total data

Contents

1 Data Engineering

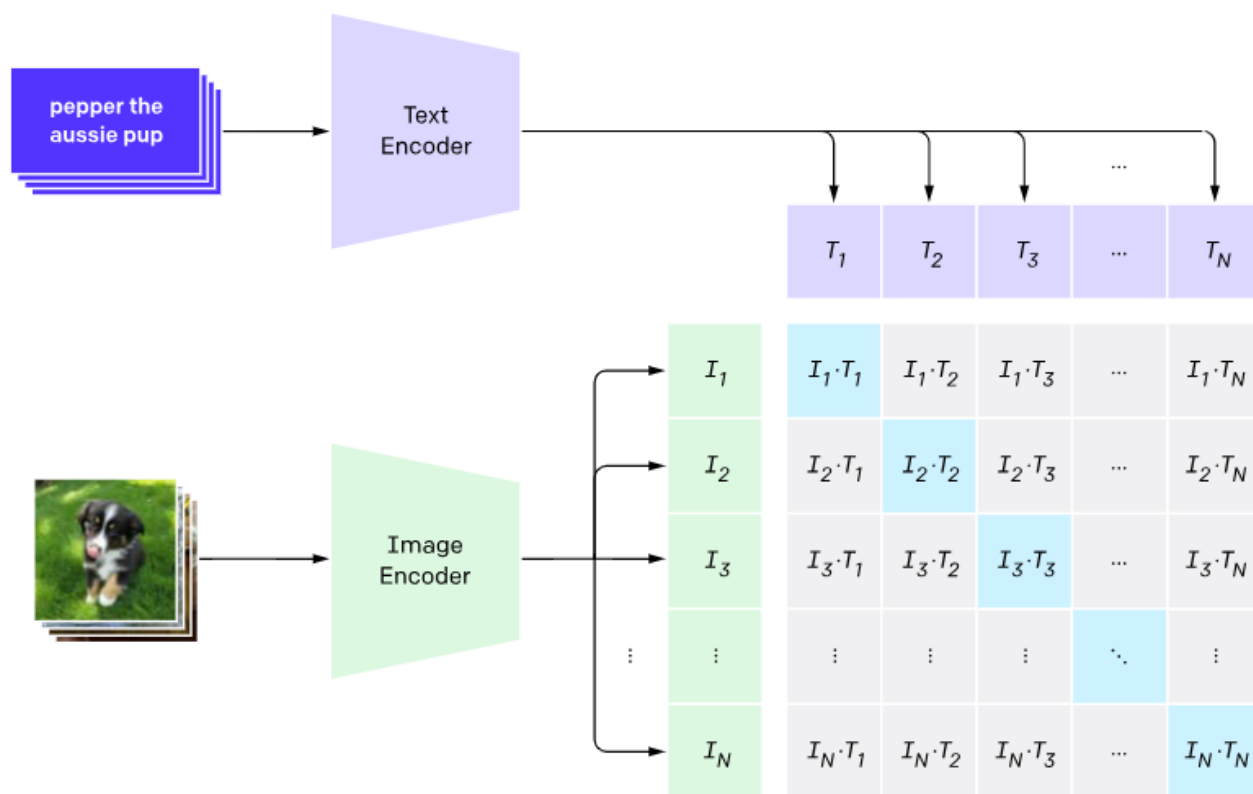
2 CV Strategy

3 Architecture



This simple model allow **high speed training**
for all GO Terms

1. Contrastive pre-training



1

Using both query & content sides embedding vector

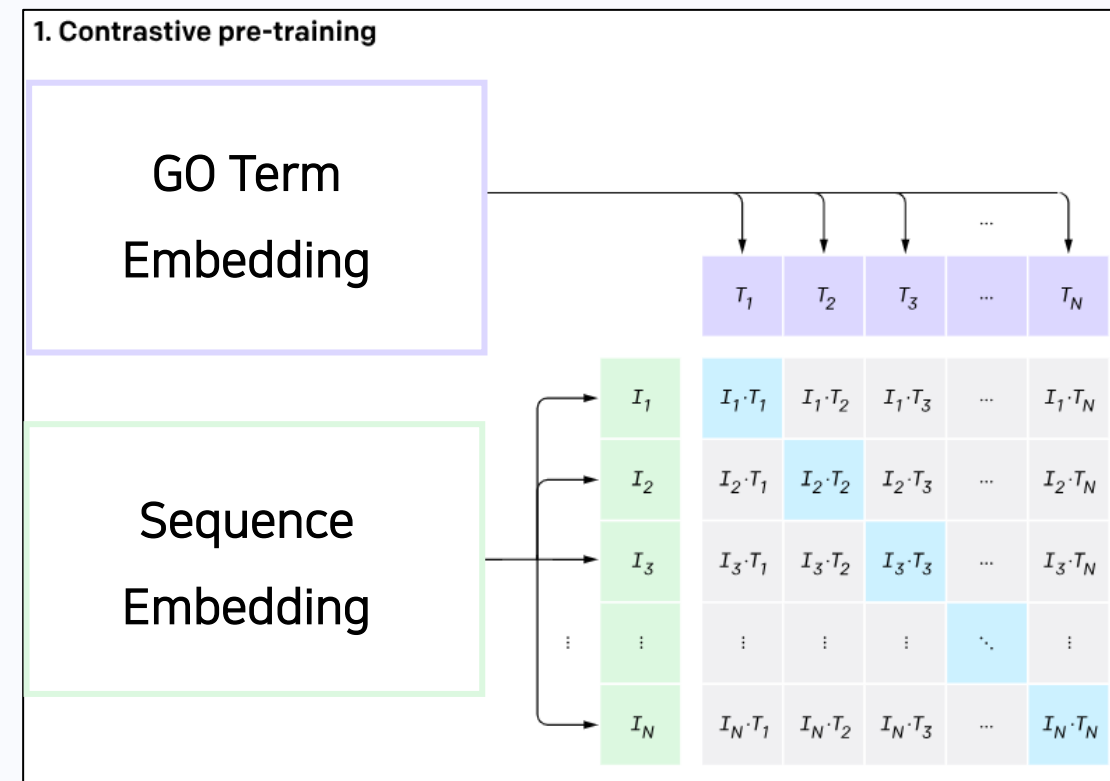
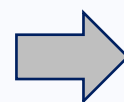
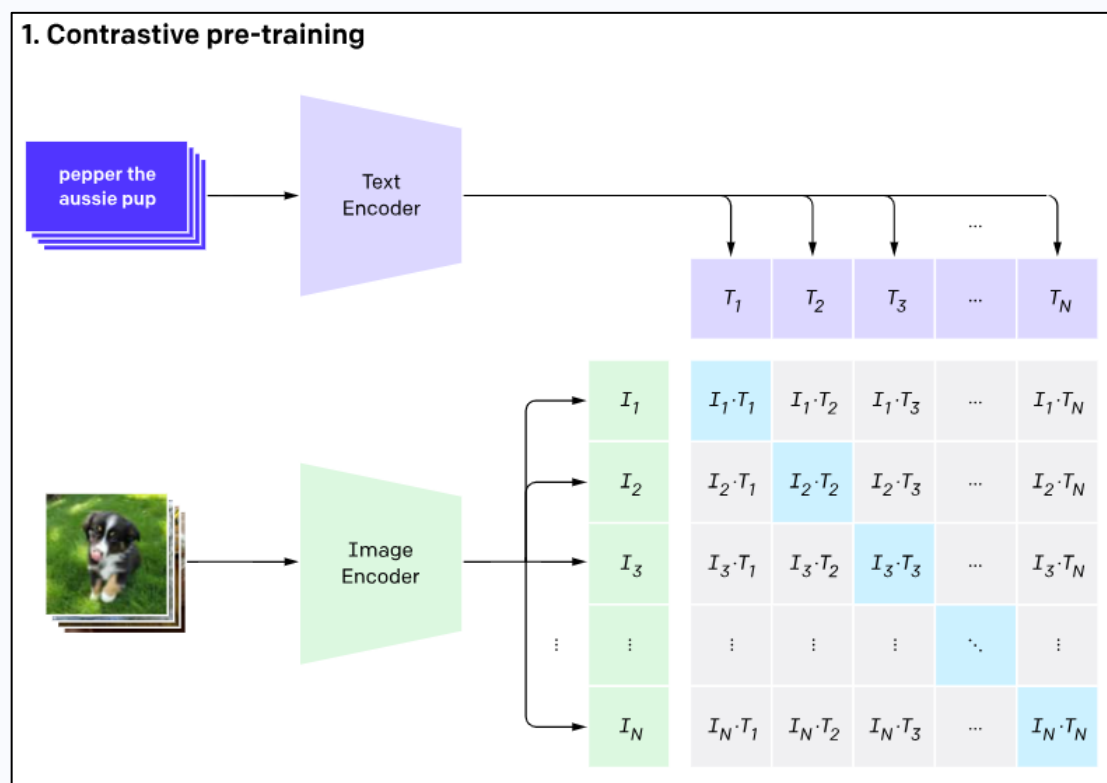
2

Calculating output with matrix multiplication



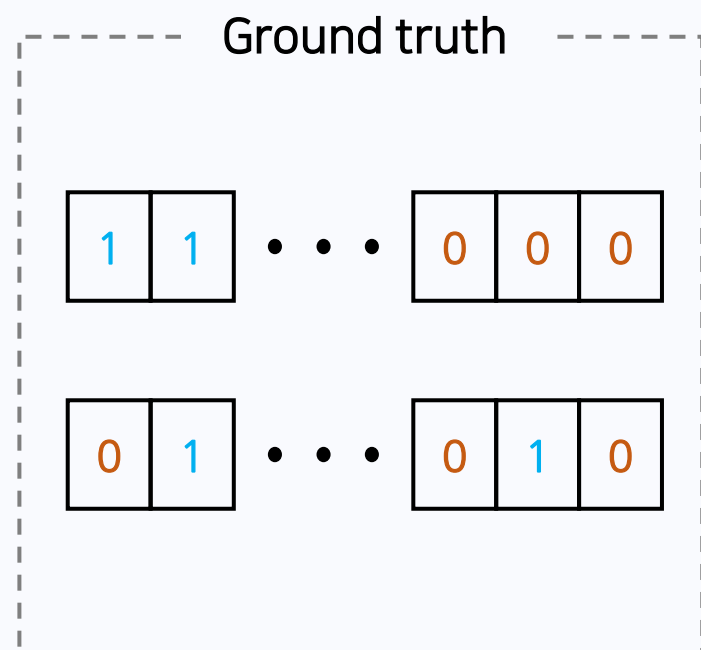
Matrix multiplication allow to calculate probability on all contents from a input query with fast speed



Architecture – CLIP Based Models



But, what do we do to migrate original model when it is not a multiclass classification task?

Introduction on **Dynamic Negative Sampling** technique for binary classification



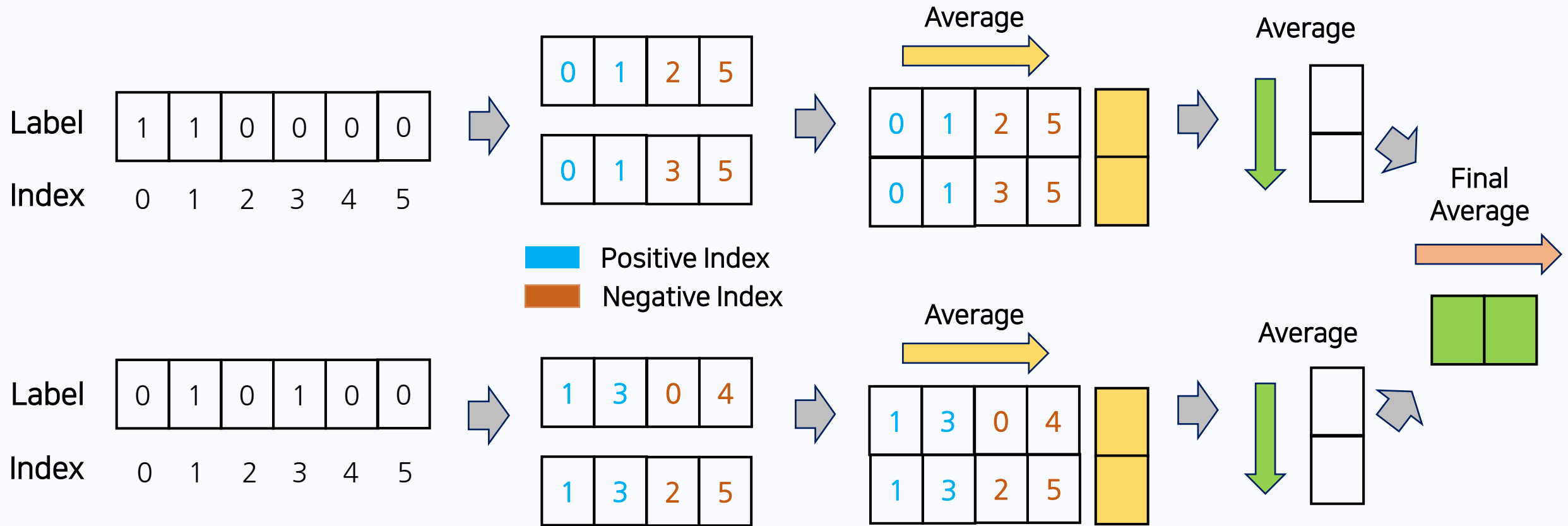
 Positive Index
 Negative Index

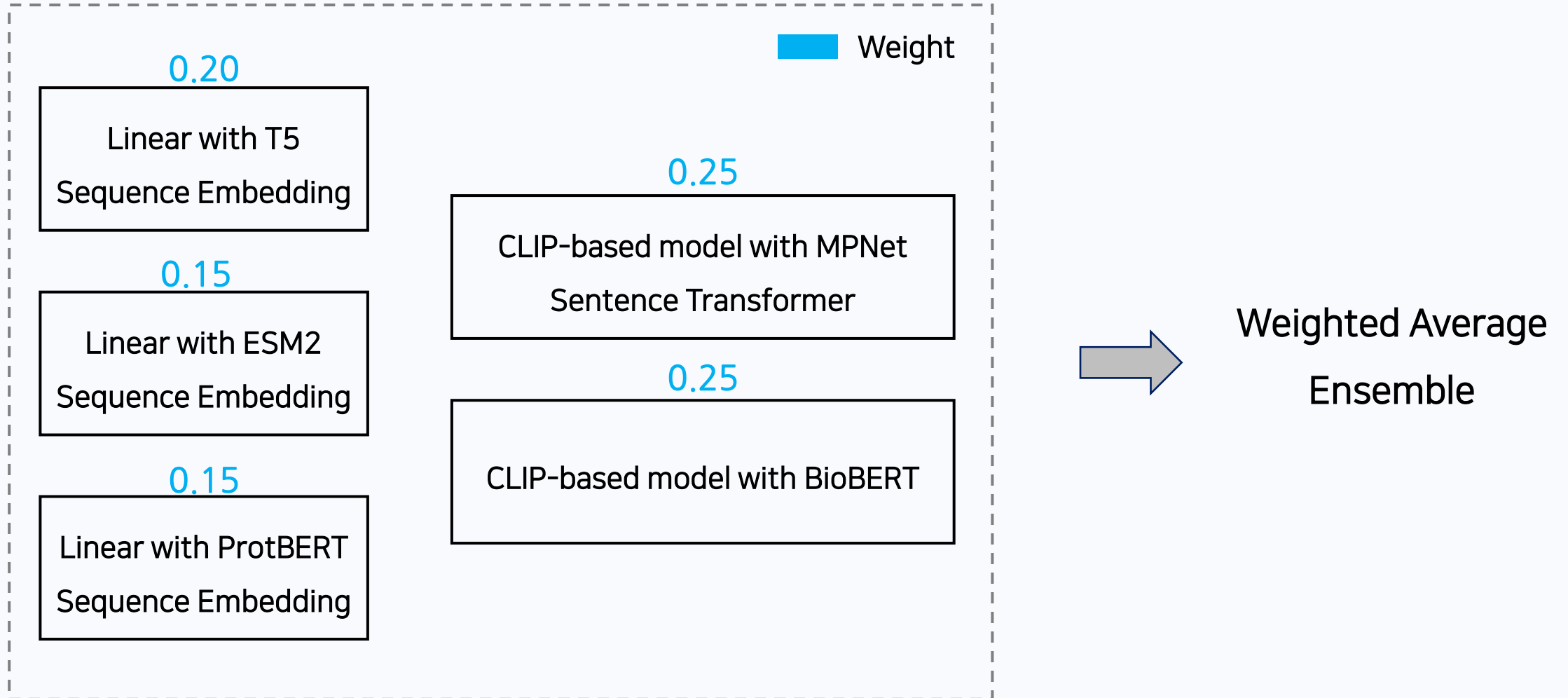
Calculating loss with Dynamic Negative Sampling Process

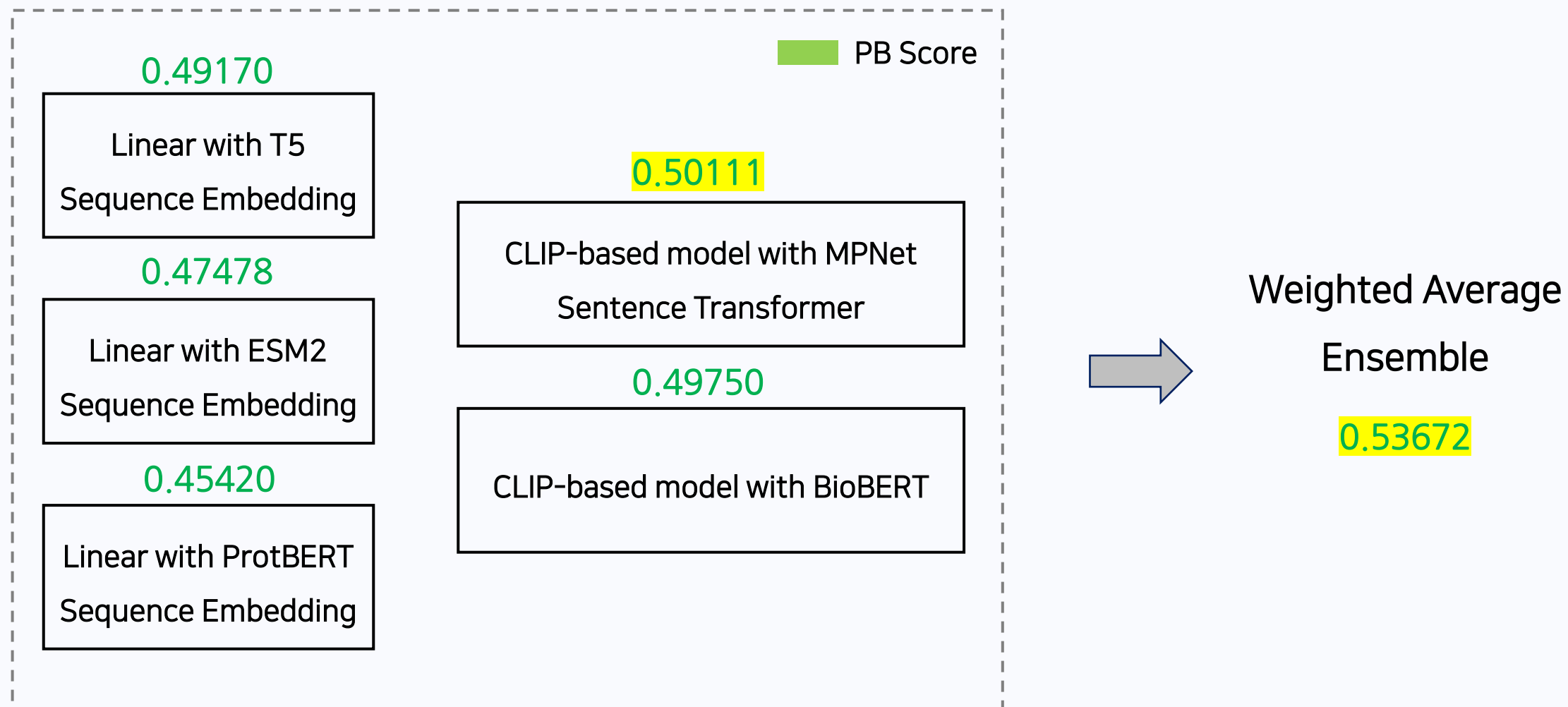
1. Shuffling negatives
2. Select number of negatives (*negative_sampling_ratio)
3. Select number of combinations (*n_combinations)
4. Average on elements' loss in each combinations
5. Average on all combinations' loss
6. Average on all batches' loss

* This is hyper-parameter

Architecture – CLIP Based Models







Thank You

The End



