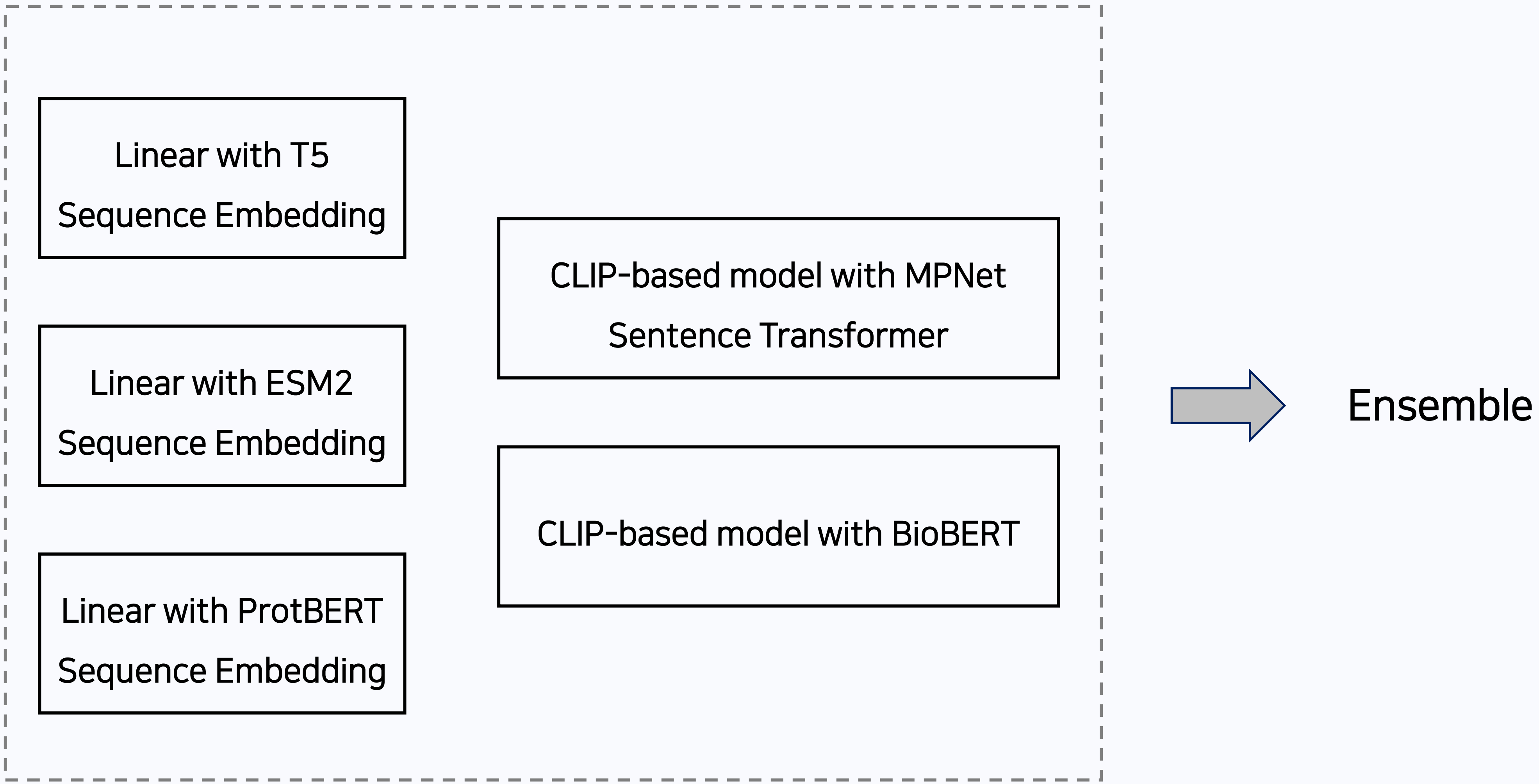
CAFA 5 Protein Function Prediction

Solution Introduction

Contents

kaggle

Architecture Preview



Contents

kaggle

Data Engineering – Linear Regression Based Models

kaggle

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

Data Engineering – CLIP Based Models

kaggle

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

kaggle

CV Strategy – Splitting Dataset

Group Stratified Sampling

Group key : Sequence

Stratifying key : GO Term

Create 25 fold buckets

Local Fold

Global Fold

5 Folds

5 Folds

5 Folds

5 Folds

5 Folds

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

kaggle

Architecture – Linear Regression Based Models

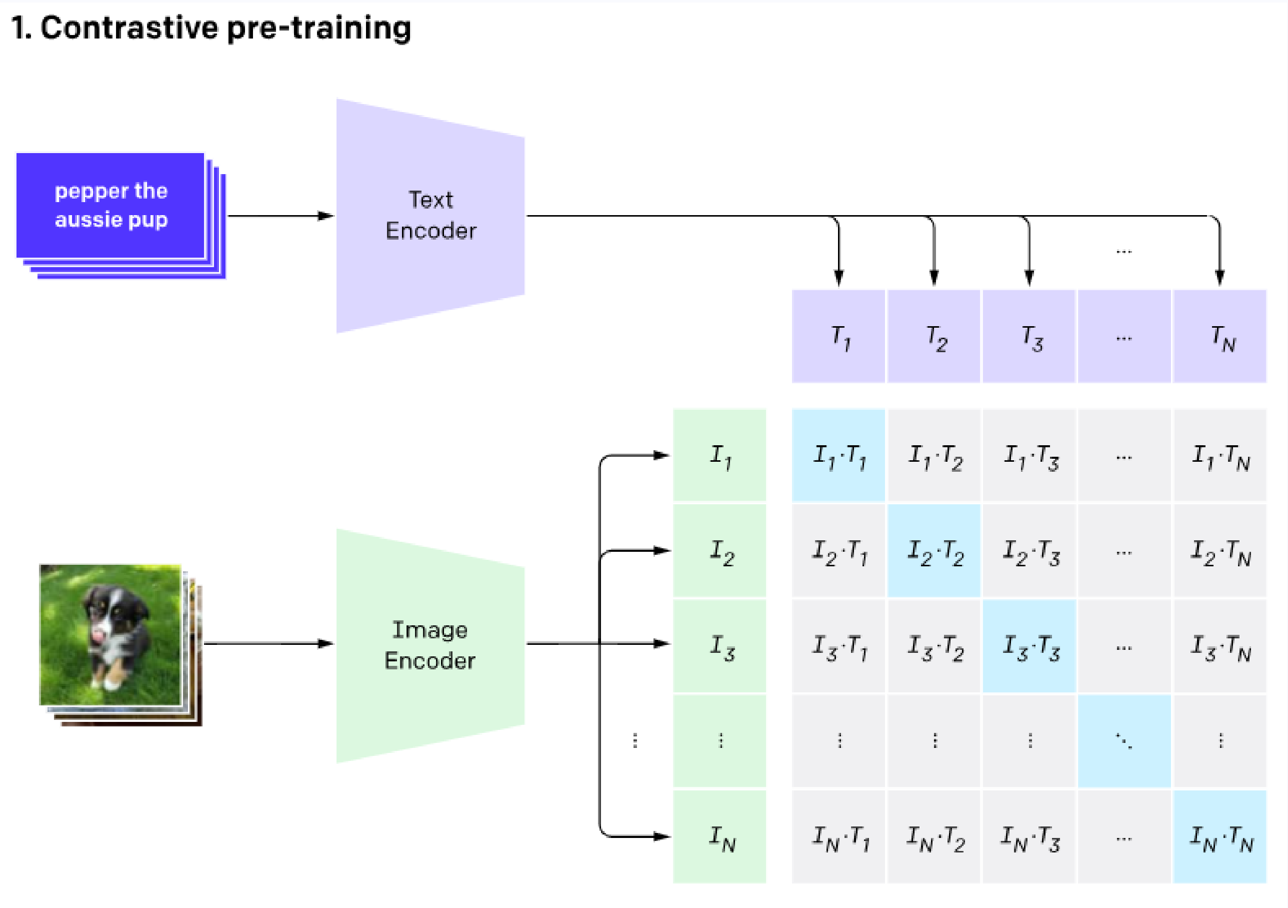
Sequence-side features → **Linear Regression** (For all GO terms) → Output on all GO terms (Treat as probability)

↓

This simple model allow high speed training
for all GO Terms

Architecture – CLIP Based Models

kaggle

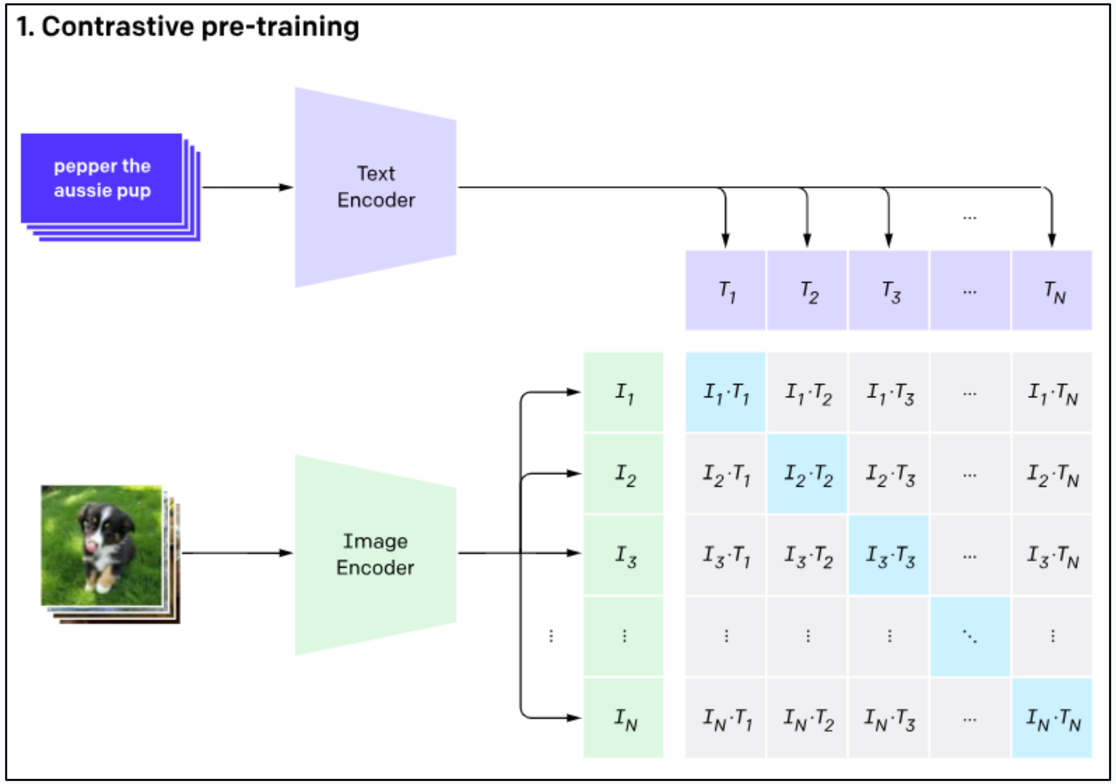


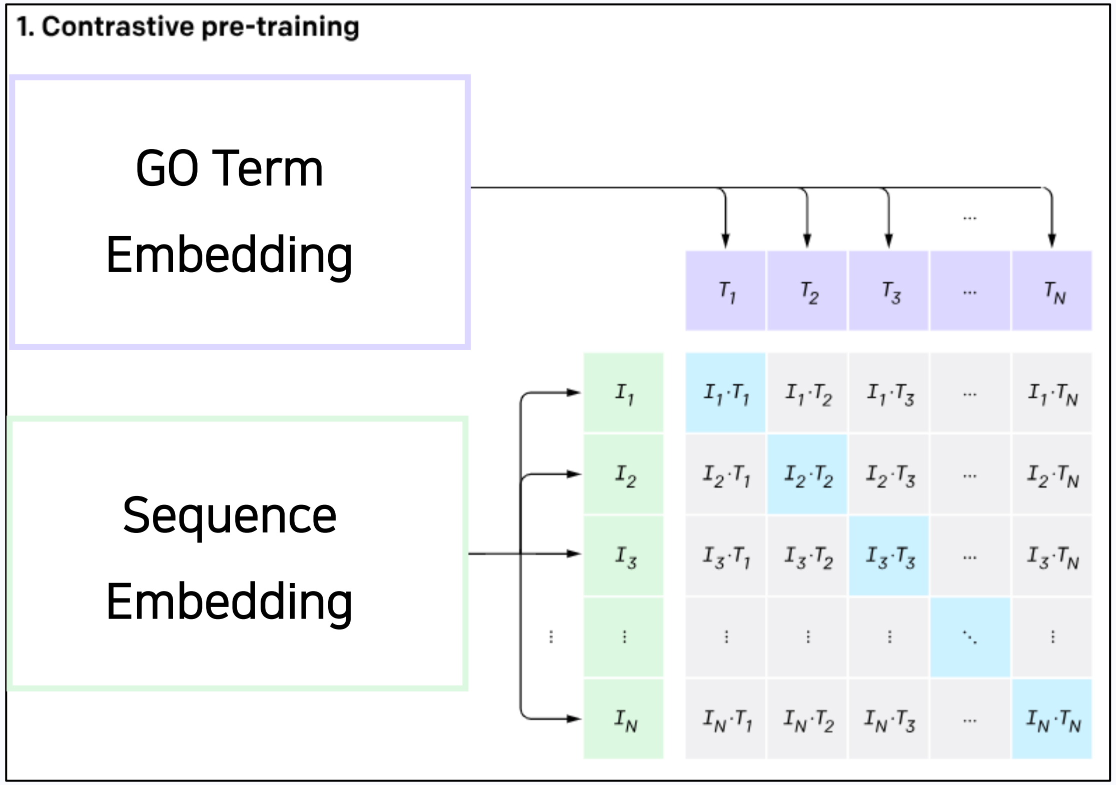
1 Using both query & content sides embedding vector

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

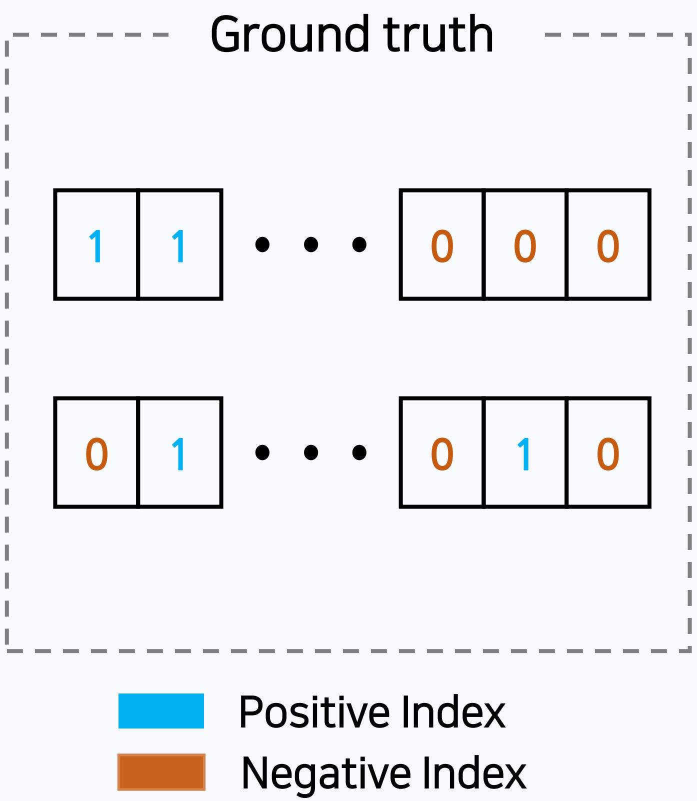




But, what do we do to migrate original model when it is not a multiclass classification task?

Architecture – CLIP Based Models

Introduction on Dynamic Negative Sampling technique for binary classification

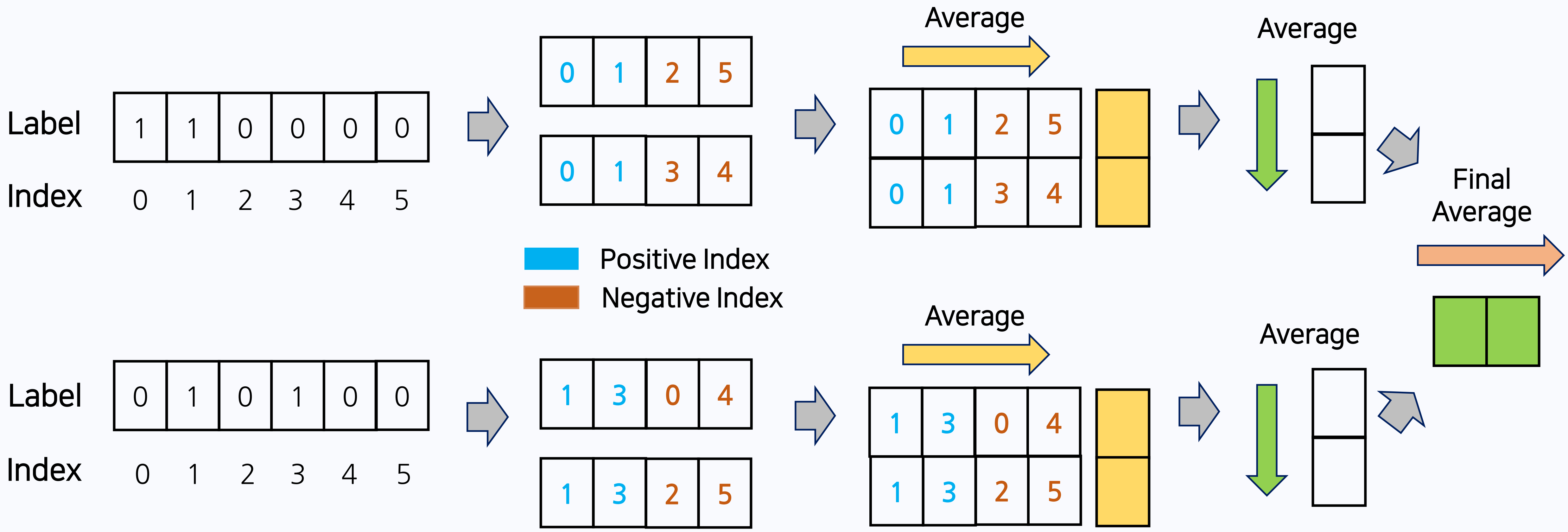


Calculating loss with Dynamic Negative Sampling Process

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



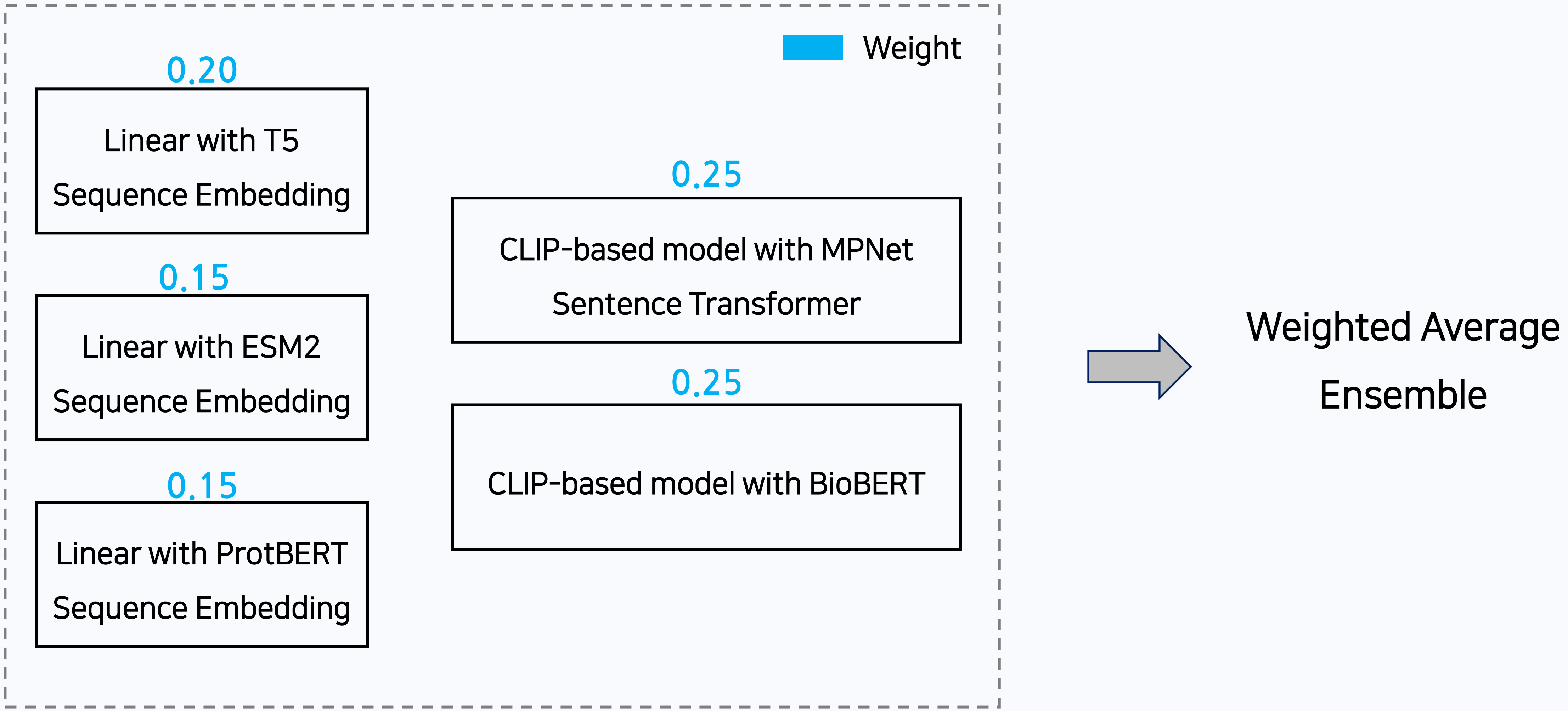
Architecture – CLIP Based Models

Dynamic Negative Sampling shows the highest performance among other techniques

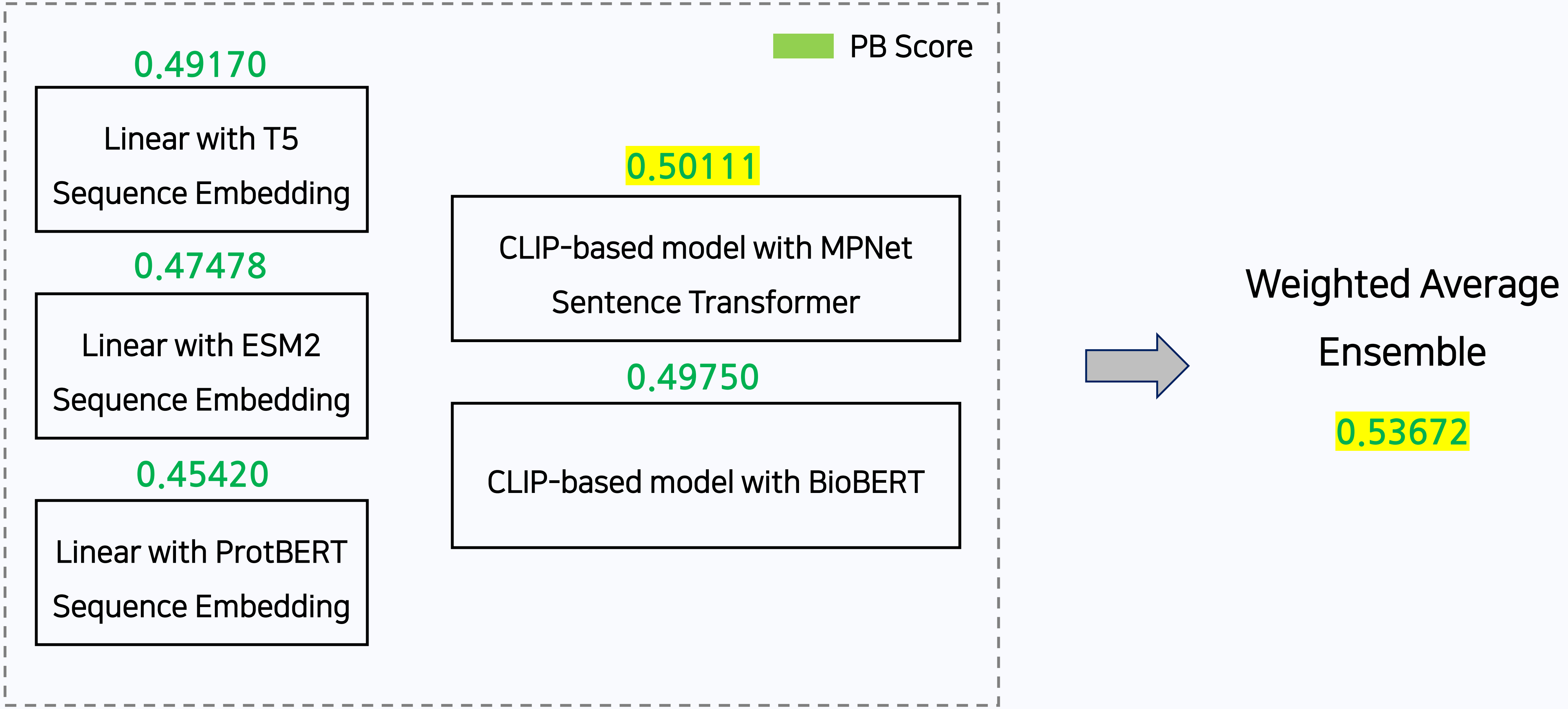
Techniques	CV	LB
No operation	0.464073	0.41247
Applying Weight Multiplier (n_negatives / n_positives)	0.460397	0.43574
*Applying Dynamic Negative Sampling (n_combinations=1)	0.458883	0.42919
*Applying Dynamic Negative Sampling (n_combinations=4)	0.457073	0.43746
*Applying Dynamic Negative Sampling (n_combinations=8)	0.457457	0.44945

* negative_sampling_ratio = 1.0

Architecture - Ensemble



Architecture – Evaluation Score



Thank You

The End

Appendix – Public Model Evaluation Score