2021 AIAC Competition Track1 Multimodal Video Similarity

一起去吃吉野家
2021.10.16

Introduction

Introduction

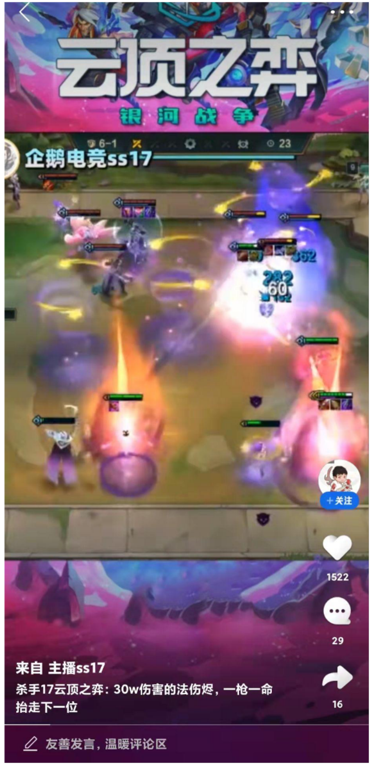
Task: Multimodal Video Similarity

Metric: Spearman rank correlation of (video embedding cosine similarity, annotated label)

Dataset :

	Pointwise	Pairwise	Test
Data size	100w	6.8w	4.3w
Describe	video	video pair, annotated similarity	video , no annotated info

Introduction



来自 主播ss17

杀手17云顶之弈：30w伤害的法伤烬，一枪一命抬走下一位

友善发言，温暖评论区

Video title : 杀手17云顶之弈：30w伤害的法伤烬，一枪一命抬走下一位

Frame feature : n_frame * 1536 (EfficientNetB3)

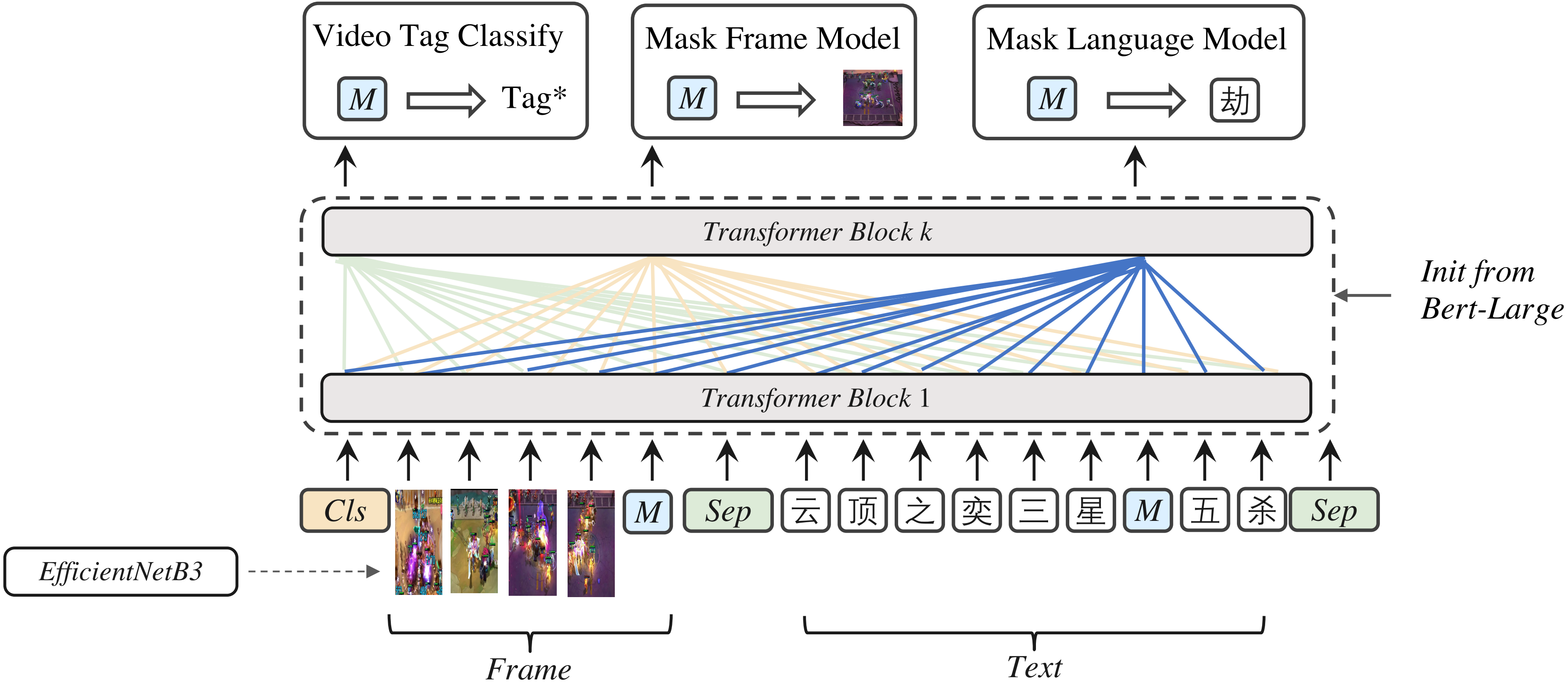
Asr text : 新闻是有盾，就是稳稳的哈，我的灵感。正在涌现。一张师傅张两张。反抢七座。拿下。你这攻速很快吗

Tag : [81774, 723622, 56044756, 55831581]

Category : 10802

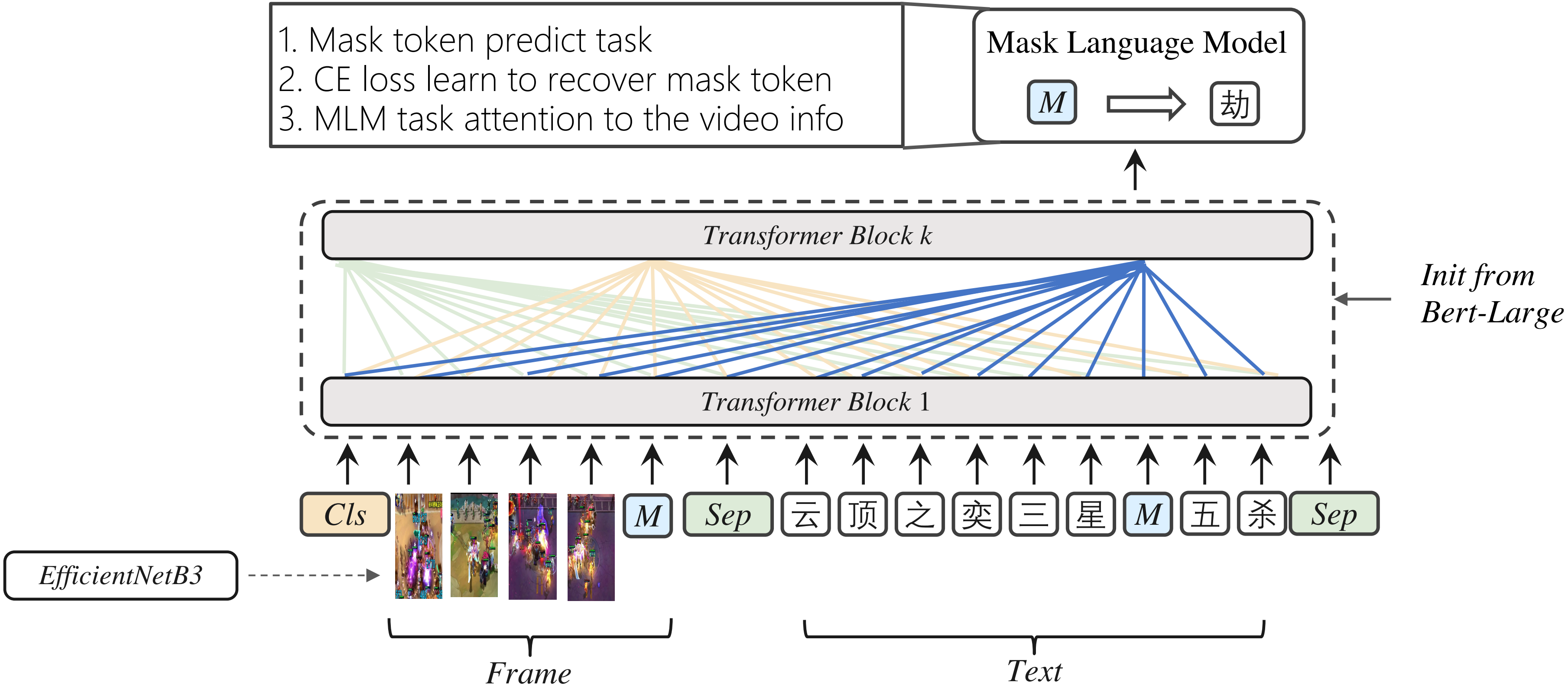
Method

Pretrain



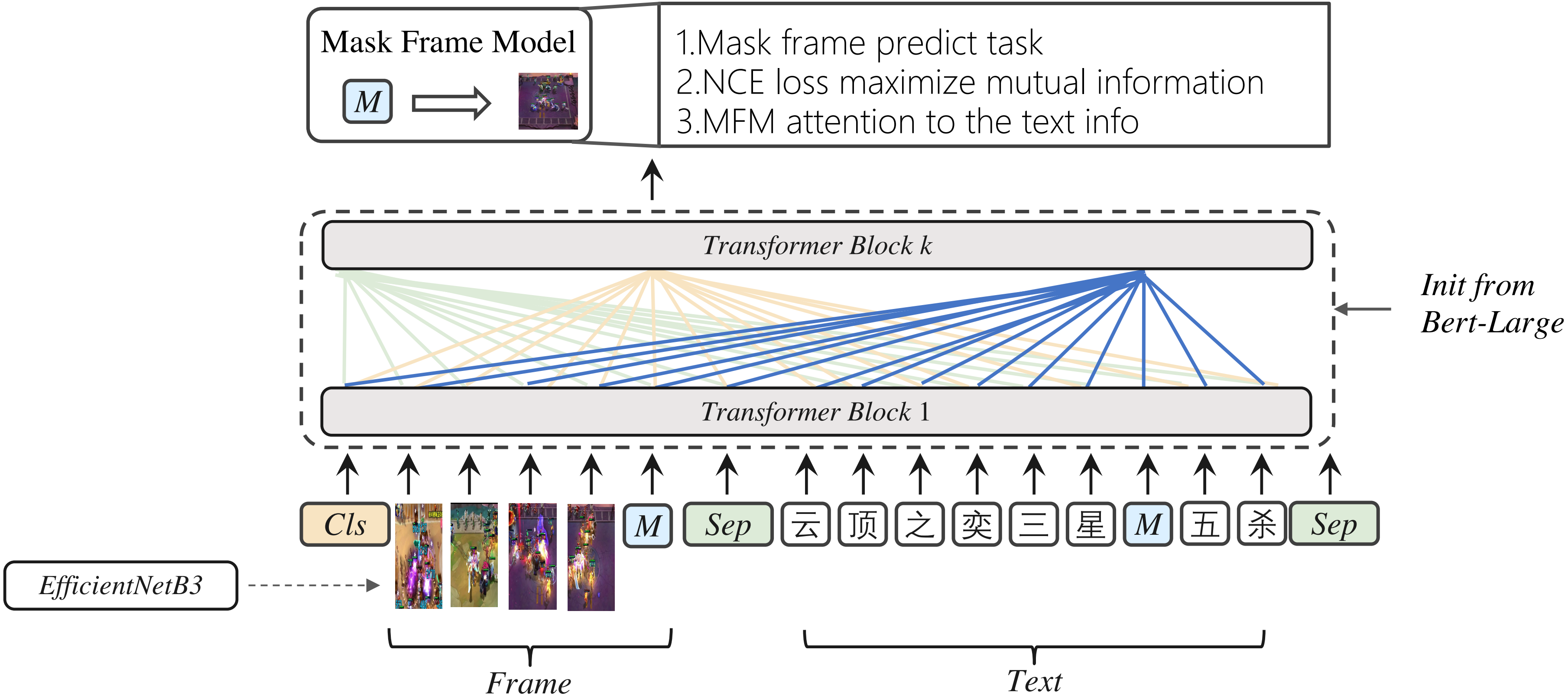
Video Tag Classify
M
Tag*
Mask Frame Model
M
Mask Language Model
M
劫
Transformer Block k
Transformer Block 1
Init from
Bert-Large
Cls
M
Sep
云
顶
之
奕
三
星
M
五
杀
Sep
EfficientNetB3
Frame
Text

Pretrain



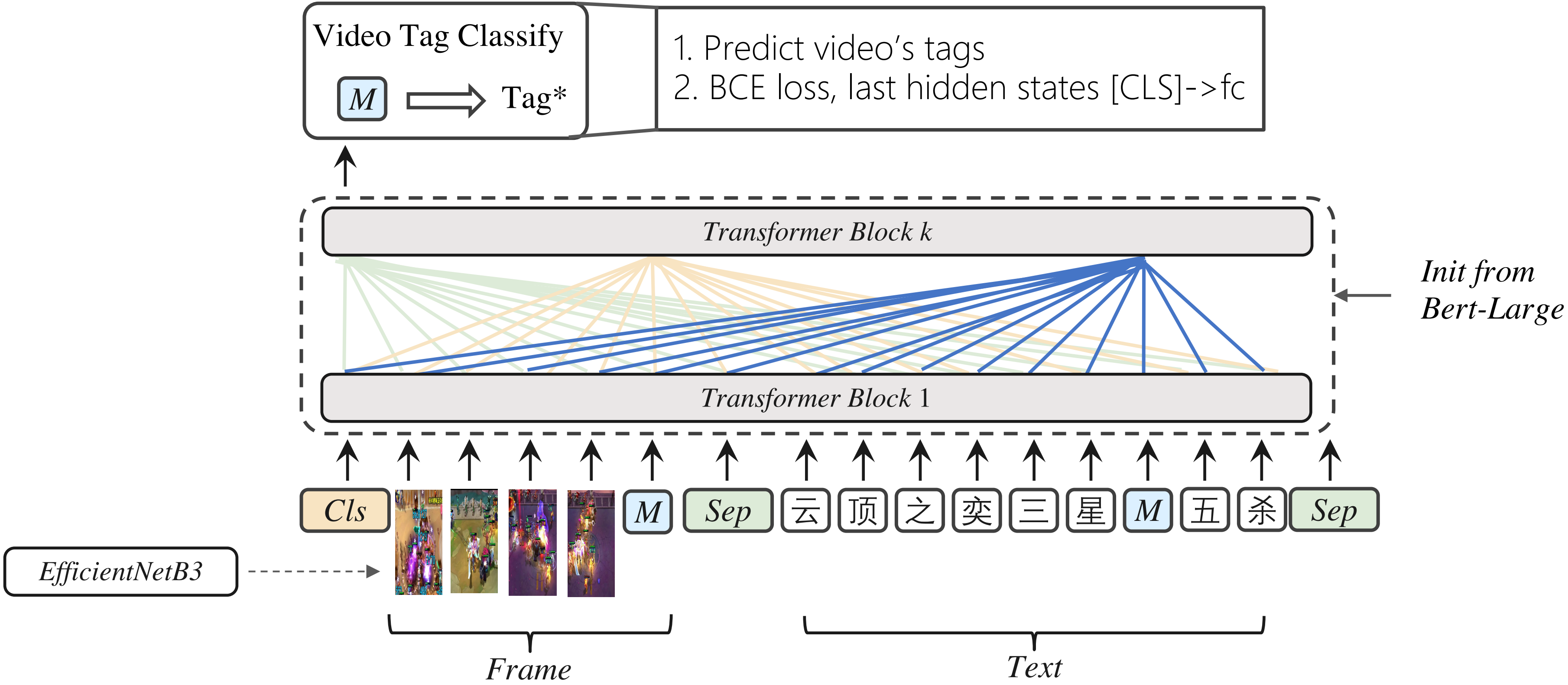
1. Mask token predict task
2. CE loss learn to recover mask token
3. MLM task attention to the video info

Pretrain



1.Mask frame predict task
2.NCE loss maximize mutual information
3.MFM attention to the text info

Pretrain



Video Tag Classify

M ⇨ Tag*

1. Predict video’s tags
2. BCE loss, last hidden states [CLS]->fc

Transformer Block k

Transformer Block 1

Init from Bert-Large

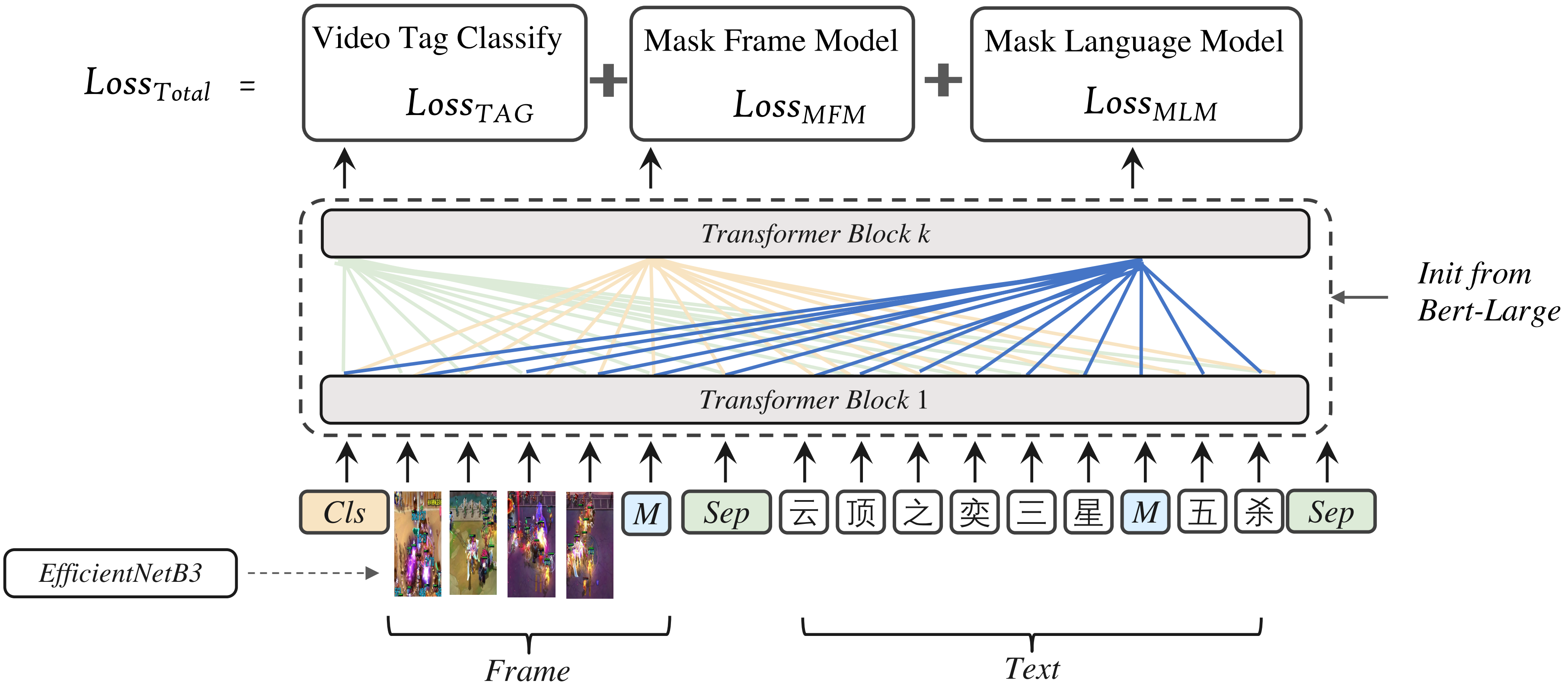
Cls *M* *Sep* 云 顶 之 奕 三 星 *M* 五 杀 *Sep*

EfficientNetB3

Frame

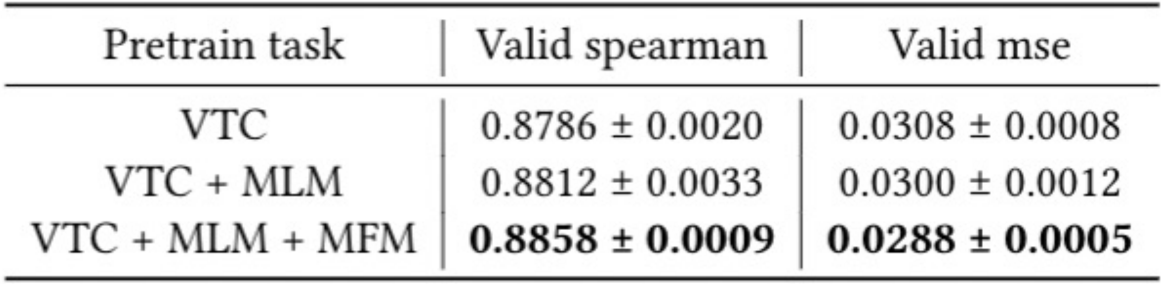
Text

Pretrain


$Loss_{Total}$ =
Video Tag Classify
$Loss_{TAG}$
+
Mask Frame Model
$Loss_{MFM}$
+
Mask Language Model
$Loss_{MLM}$
Transformer Block k
Transformer Block 1
Init from
Bert-Large
Cls
M
Sep
云
顶
之
奕
三
星
M
五
杀
Sep
EfficientNetB3
Frame
Text

Pretrain

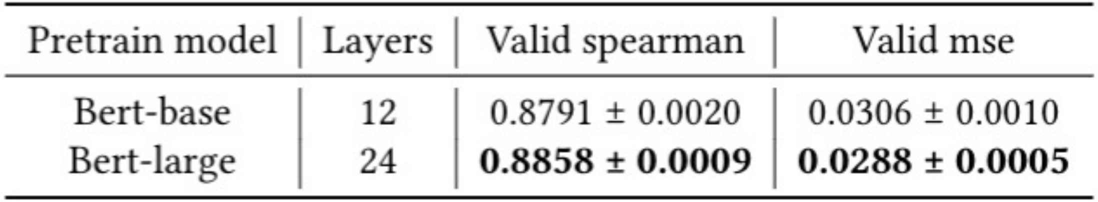
Different pretrain task



Pretrain task	Valid spearman	Valid mse
VTC	0.8786 ± 0.0020	0.0308 ± 0.0008
VTC + MLM	0.8812 ± 0.0033	0.0300 ± 0.0012
VTC + MLM + MFM	**0.8858 ± 0.0009**	**0.0288 ± 0.0005**

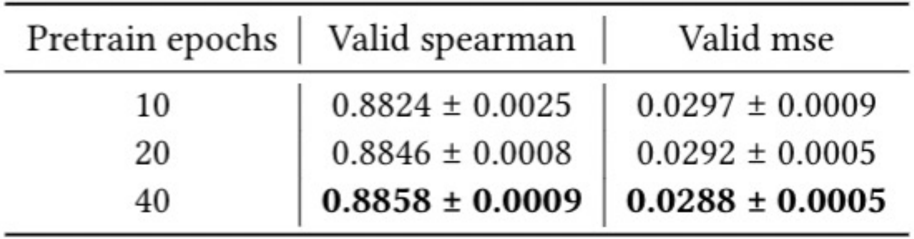
Pretrain

Larger model



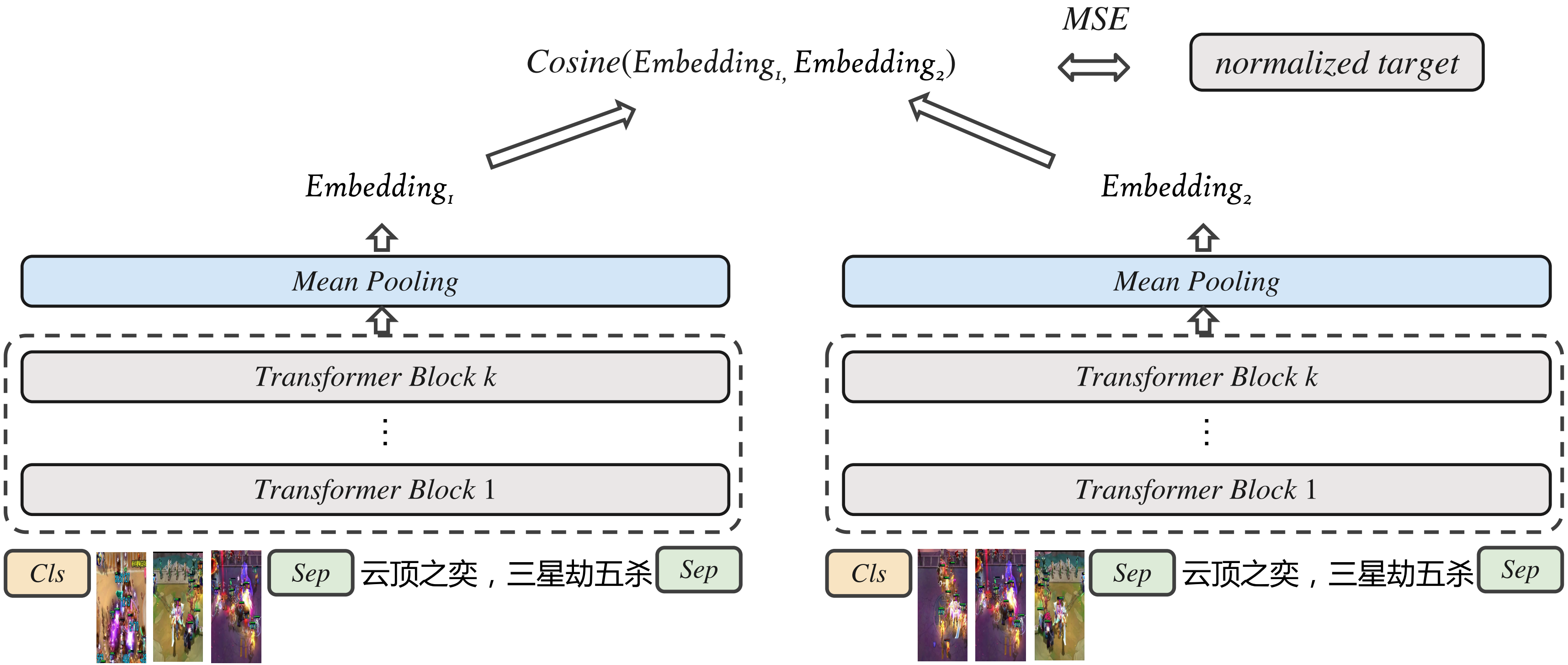
Pretrain model	Layers	Valid spearman	Valid mse
Bert-base	12	0.8791 ± 0.0020	0.0306 ± 0.0010
Bert-large	24	**0.8858 ± 0.0009**	**0.0288 ± 0.0005**

Pretrain longer



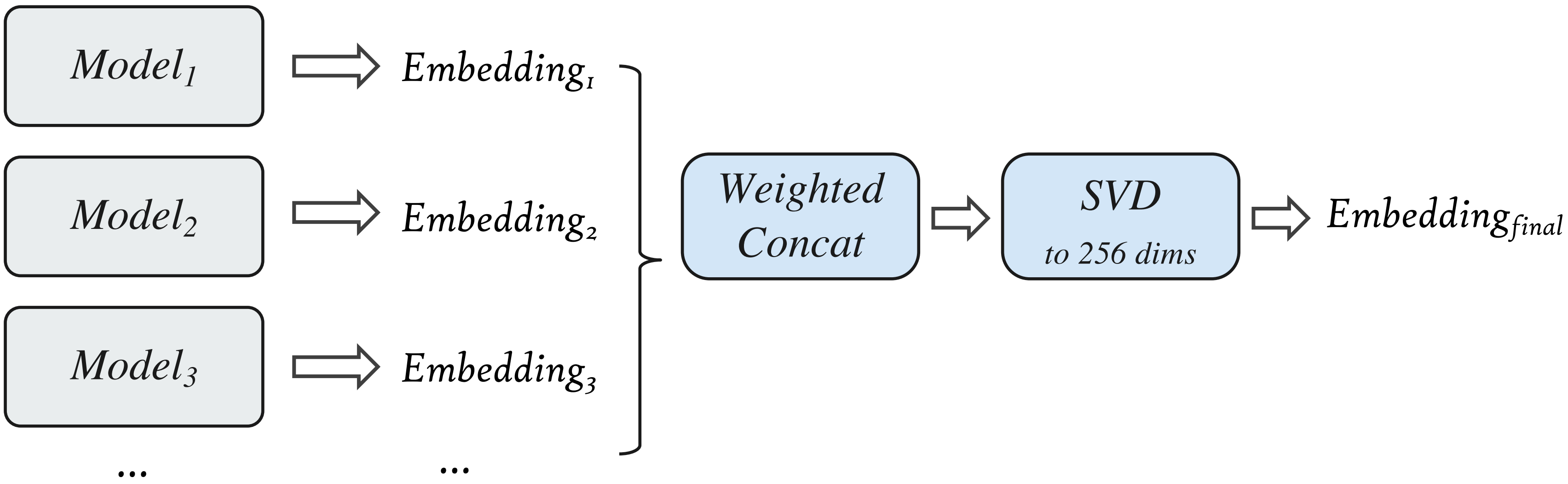
Pretrain epochs	Valid spearman	Valid mse
10	0.8824 ± 0.0025	0.0297 ± 0.0009
20	0.8846 ± 0.0008	0.0292 ± 0.0005
40	**0.8858 ± 0.0009**	**0.0288 ± 0.0005**

Finetune


MSE
Cosine(Embedding₁, Embedding₂)
normalized target
Embedding₁
Embedding₂
Mean Pooling
Mean Pooling
Transformer Block k
⋮
Transformer Block 1
Transformer Block k
⋮
Transformer Block 1
Cls
Sep
云顶之奕，三星劫五杀
Sep
Cls
Sep
云顶之奕，三星劫五杀
Sep

Ensemble

Weighted concat -> SVD



$Model_1$ → *$Embedding_1$*

$Model_2$ → *$Embedding_2$*

$Model_3$ → *$Embedding_3$*

… …

Weighted Concat → *SVD* *to 256 dims* → *$Embedding_{final}$*

Thanks