NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Movie Box Office Prediction With Self-Supervised and Visually Grounded Pretraining

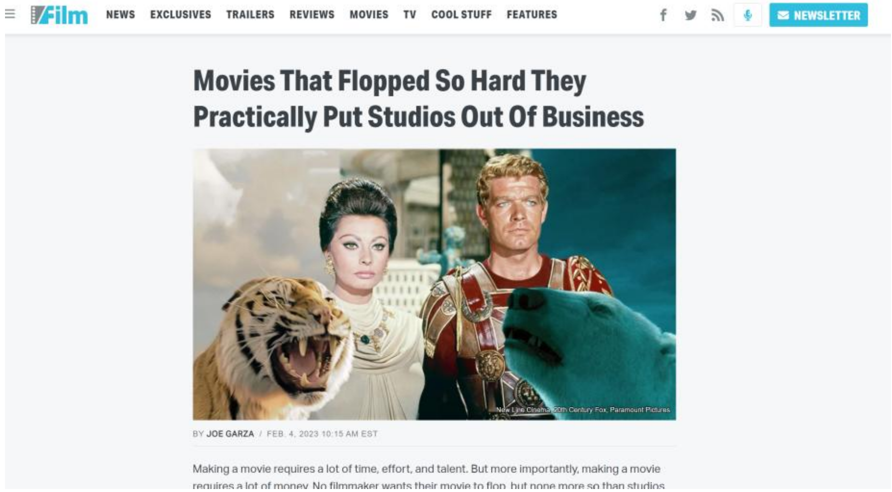
Nanyang Technological University

Authors:
Qin Chao (Speaker)
Eunsoo Kim
Boyang Li

13 July 2023

Introduction to Movie Box Office Prediction

- Movie investment carries significant risks
  - "Cleopatra (1963)", nearly ruined 20th Century Fox
  - "The Golden Compass (2007)", caused New Line Cinema absorbed into Warner Bros.
  - "Cutthroat Island(1995) “ made Carolco Pictures ceased to exist.



Movies That Flopped So Hard They Practically Put Studios Out Of Business

New Line Cinema, 20th Century Fox, Paramount Pictures

BY JOE GARZA / FEB. 4, 2023 10:15 AM EST

Making a movie requires a lot of time, effort, and talent. But more importantly, making a movie requires a lot of money. No filmmaker wants their movie to flop, but none more so than studios

Justice League Makes a HUGE Movie Flop an Essential Part of DC Lore

The Jonah Hex film might have been a huge box office flop, but that isn't continuity.

BY KATE O'DONOGHUE PUBLISHED SEP 2, 2022

DC Cancelled Gal Gadot's Major Wonder Woman Future Plans

By Richard Nebens Posted: June 30, 2023

Gal Gadot's future as Wonder Woman came to a screeching halt thanks to Warner Bros.' new plans for the greater DC Universe.

LATEST NEWS

The Last of Us Season 2 Cast: Every Actor Expected to Appear

Indiana Jones 5 Cast, Characters & Actors

First Look at The CW's Horrifying Two-Face Revealed by DC Show (Photos)

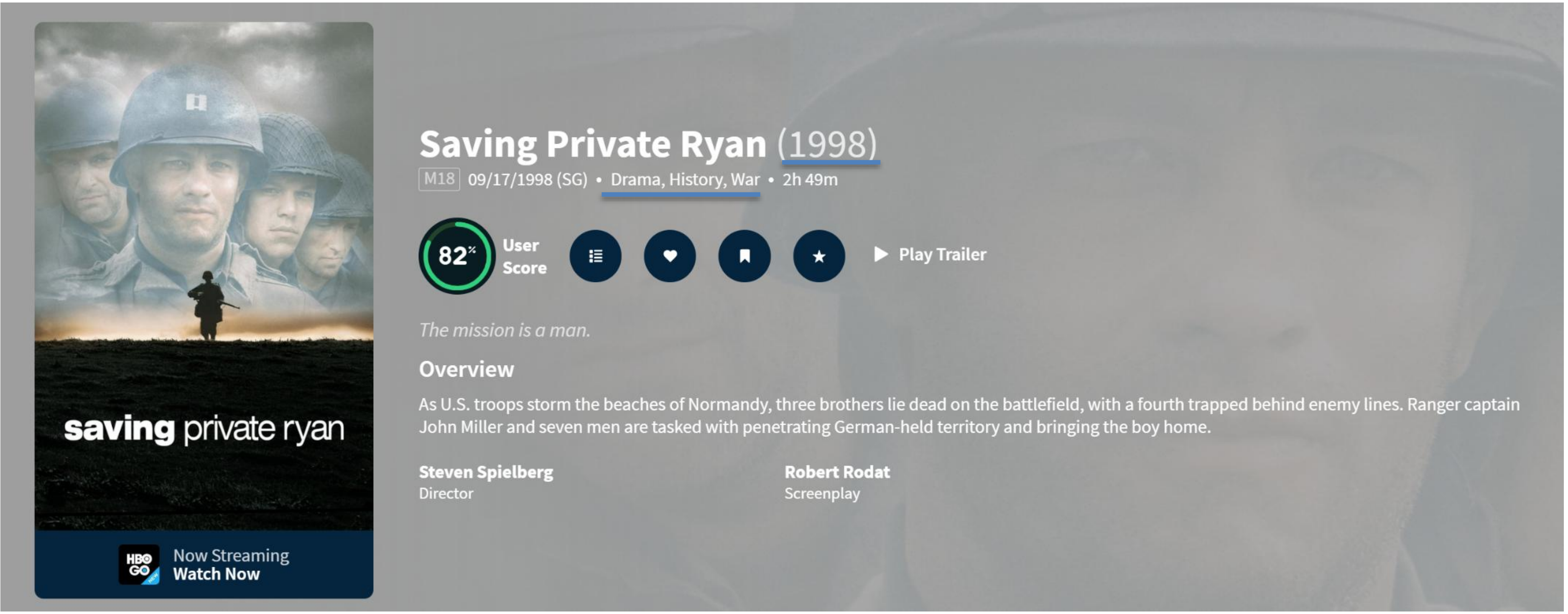
The Flash's Box Office Drop Is The 2nd Worst Behind Another Big Superhero Flop

The Flash movie continues to underperform and disappoint at the box office in its second weekend, echoing a previous superhero bomb.

BY NICHOLAS BEACH PUBLISHED 3 DAYS AGO

Introduction to Movie Box Office Prediction

- Features Collection: TMDB website



Saving Private Ryan (1998)

M18 09/17/1998 (SG) • Drama, History, War • 2h 49m

The mission is a man.

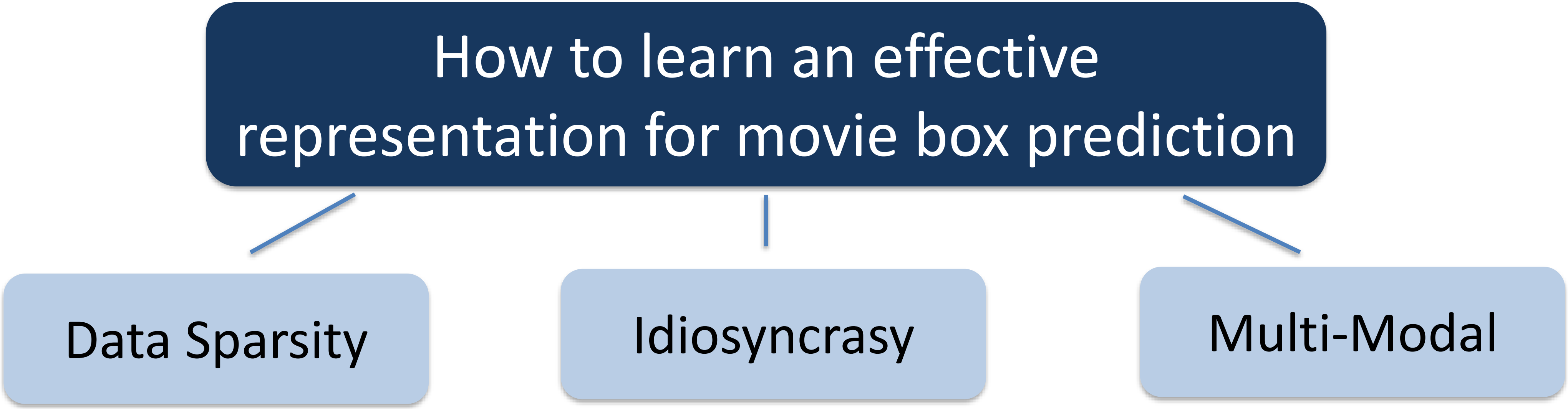
Overview

As U.S. troops storm the beaches of Normandy, three brothers lie dead on the battlefield, with a fourth trapped behind enemy lines. Ranger captain John Miller and seven men are tasked with penetrating German-held territory and bringing the boy home.

Steven Spielberg
Director

Robert Rodat
Screenplay

Challenges



How to learn an effective representation for movie box prediction

Data Sparsity

Idiosyncrasy

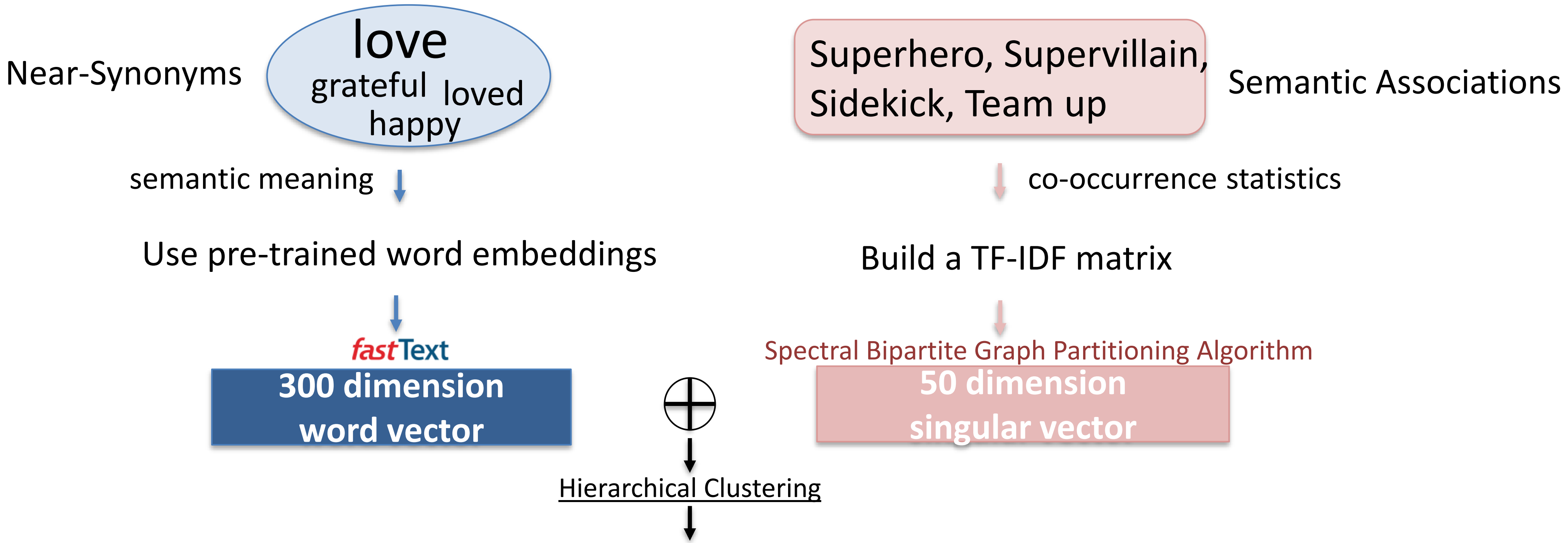
Multi-Modal

- Near-synonym keywords
- Missing keywords

- Robots in Sci-fi vs Robots on assembly line
- BERT embedding does not fit well

- Movie poster under-utilized

Keywords Clustering



Near-Synonyms

love
grateful loved
happy

semantic meaning

Use pre-trained word embeddings

*fast*Text

300 dimension word vector

Superhero, Supervillain, Sidekick, Team up

Semantic Associations

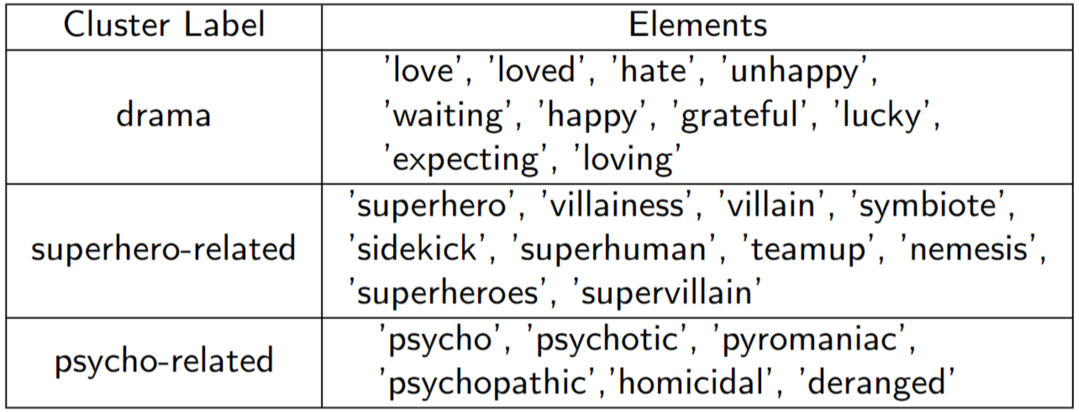
co-occurrence statistics

Build a TF-IDF matrix

Spectral Bipartite Graph Partitioning Algorithm

50 dimension singular vector

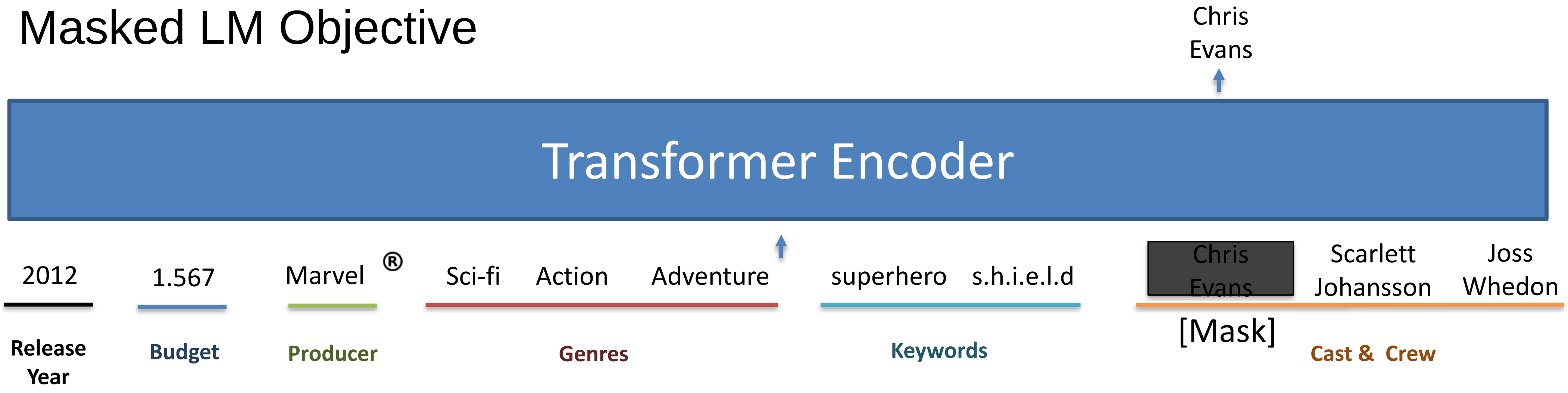
Hierarchical Clustering



Cluster Label	Elements
drama	'love', 'loved', 'hate', 'unhappy', 'waiting', 'happy', 'grateful', 'lucky', 'expecting', 'loving'
superhero-related	'superhero', 'villainess', 'villain', 'symbiote', 'sidekick', 'superhuman', 'teamup', 'nemesis', 'superheroes', 'supervillain'
psycho-related	'psycho', 'psychotic', 'pyromaniac', 'psychopathic','homicidal', 'deranged'

Self-Supervised Learning Pretraining

- Masked LM Objective



- Numerical Embedding (compute the distance to an anchor vector)

$$\mathrm{NE}_i(x) = \exp\left(-\frac{\|x - q_i\|_2}{\sigma^2}\right)$$, where $\{q_i\}_{i=0}^{D-1}$ are D evenly spaced number over [-10, 10]

Visual Grounding

- ‘robot’ have different meaning in film

Sparrow, Amazon’s latest warehouse robot, leverages computer vision and artificial intelligence to recognise and handle millions of items

VS

PACIFIC RIM

- Use off-the-shelf object detection model (e.g., VinVL, 1500+ object labels)
    - Extract local feature maps for all the detections (discard tiny objects and titles)

Visual grounding

- Contrastive Learning Objective

$\mathcal{Z}_i = \{z_m\}_{m=1}^{M}$: visual features for ***M*** objects on the poster $\xleftrightarrow{\text{similar}}$ $\mathcal{X}_i = \{x_k\}_{k=1}^{K}$: contextualized embeddings of the ***K*** keywords

- Many-to-Many

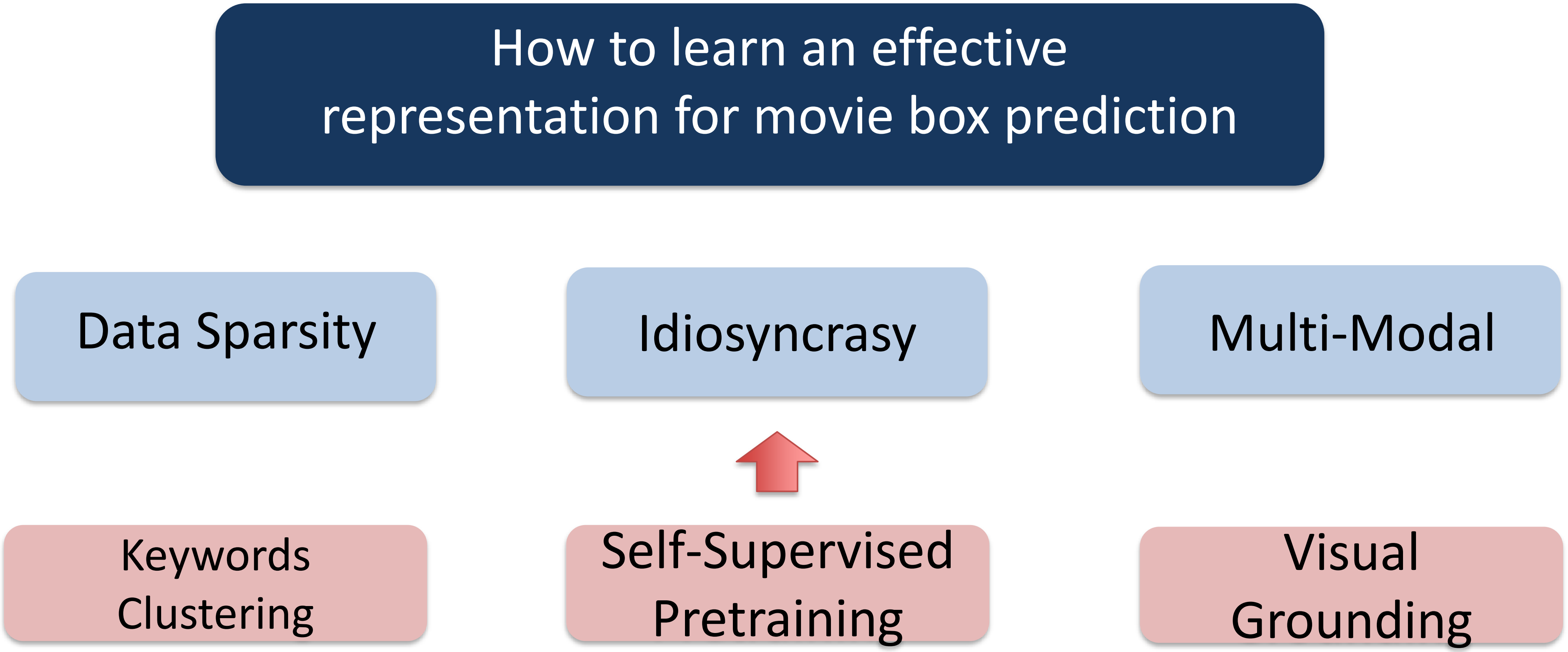
Positive pairs: $(\boldsymbol{x}, \boldsymbol{z}) \in \mathcal{X}_i \times \mathcal{Z}_i$

$$\mathcal{L}_{\mathrm{VG}} = -\frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{\mathrm{sim}(i,i)}{\mathrm{sim}(i,i) + \sum_{(i',j')} \mathrm{sim}(i',j')}\right)$$

$$\downarrow$$

$$\mathrm{sim}(i,i) = \sum_{(x,z)\in\mathcal{X}_i\times\mathcal{Z}_i} \exp\left(\frac{x^{\top} z}{\|x\|_2\|z\|_2}\right)$$

Challenges



How to learn an effective representation for movie box prediction

Data Sparsity

Idiosyncrasy

Multi-Modal

Keywords Clustering

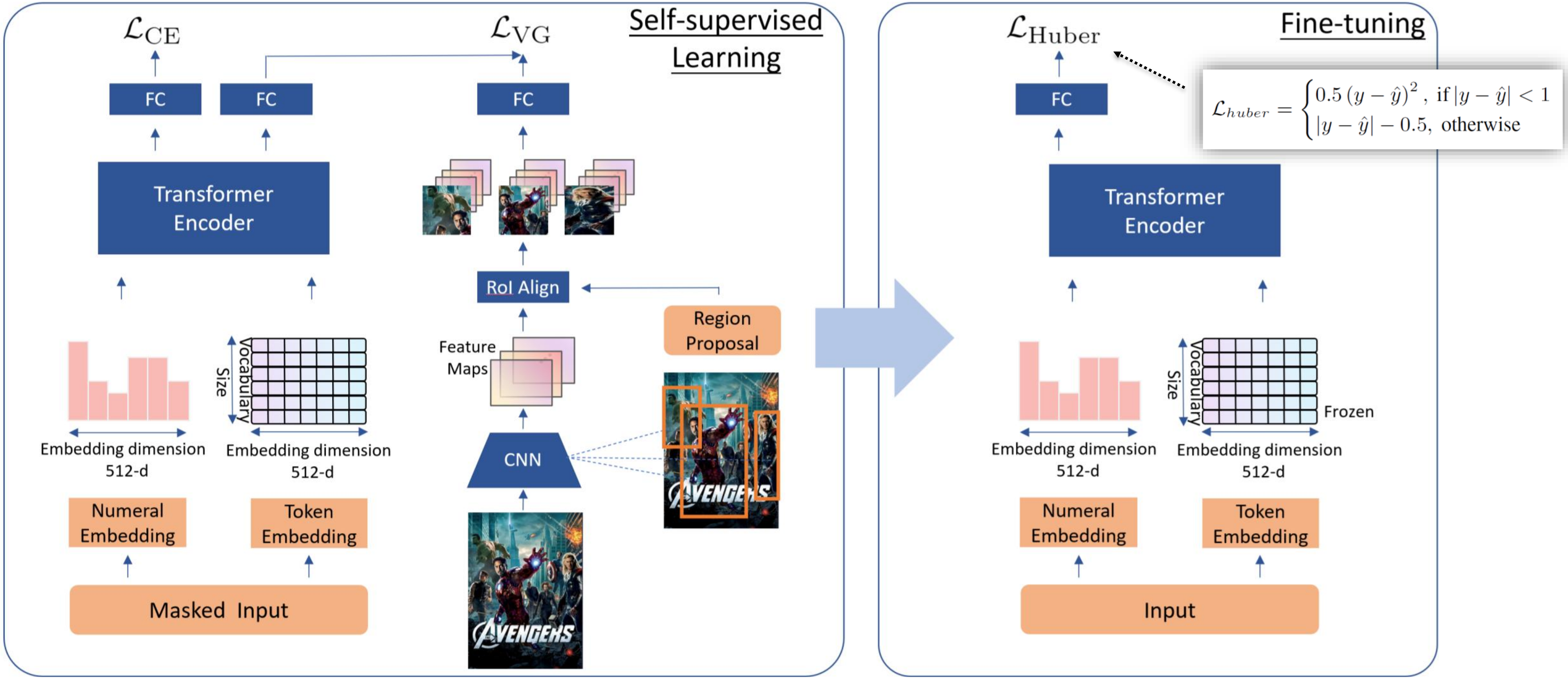
Self-Supervised Pretraining

Visual Grounding

Model Architecture

An example of input with textual and numerical features:

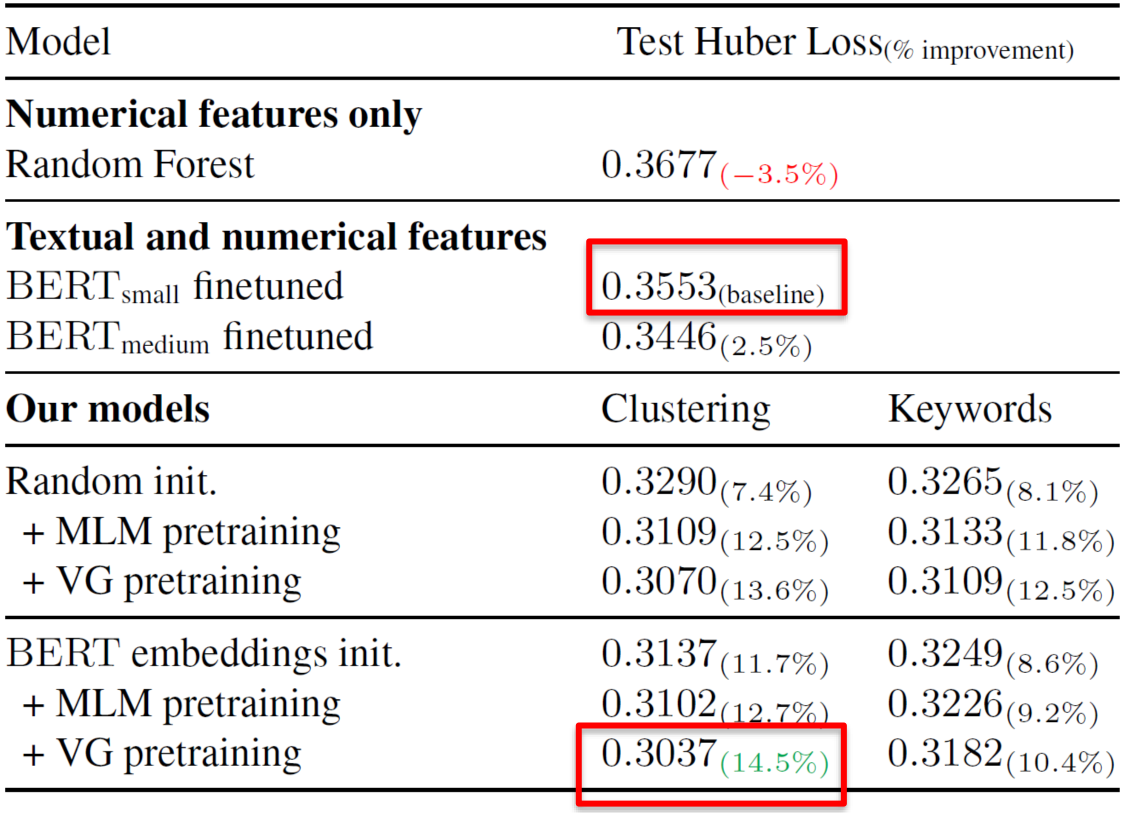
[CLS][PG-13]1.5678[Genres][Action][Sci-Fi][Keywords][shield][superhero][Directors][Joss Whedon][Actors][Chris Evans][SEP]



$$\mathcal{L}_{huber} = \begin{cases} 0.5\,(y-\hat{y})^2\,, & \text{if } |y-\hat{y}| < 1 \\ |y-\hat{y}| - 0.5, & \text{otherwise} \end{cases}$$

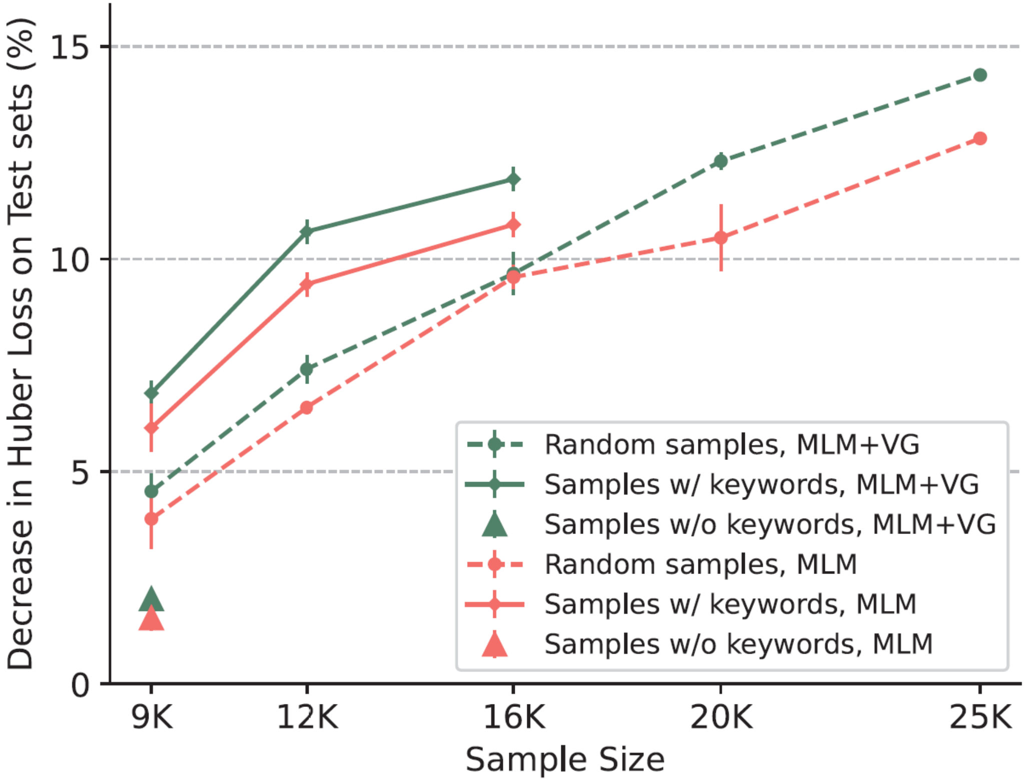
Main Results

1. Our best model shows a 14.5% of accuracy improvement compared to BERTsmall.



Model	Test Huber Loss(% improvement)	
Numerical features only		
Random Forest	0.3677(−3.5%)	
Textual and numerical features		
BERT$_{small}$ finetuned	0.3553(baseline)	
BERT$_{medium}$ finetuned	0.3446(2.5%)	
Our models	Clustering	Keywords
Random init.	0.3290(7.4%)	0.3265(8.1%)
+ MLM pretraining	0.3109(12.5%)	0.3133(11.8%)
+ VG pretraining	0.3070(13.6%)	0.3109(12.5%)
BERT embeddings init.	0.3137(11.7%)	0.3249(8.6%)
+ MLM pretraining	0.3102(12.7%)	0.3226(9.2%)
+ VG pretraining	0.3037(14.5%)	0.3182(10.4%)

2. Independent to the sample size, VG method consistently improve the result.



Qualitative Eval – Image Retrieval

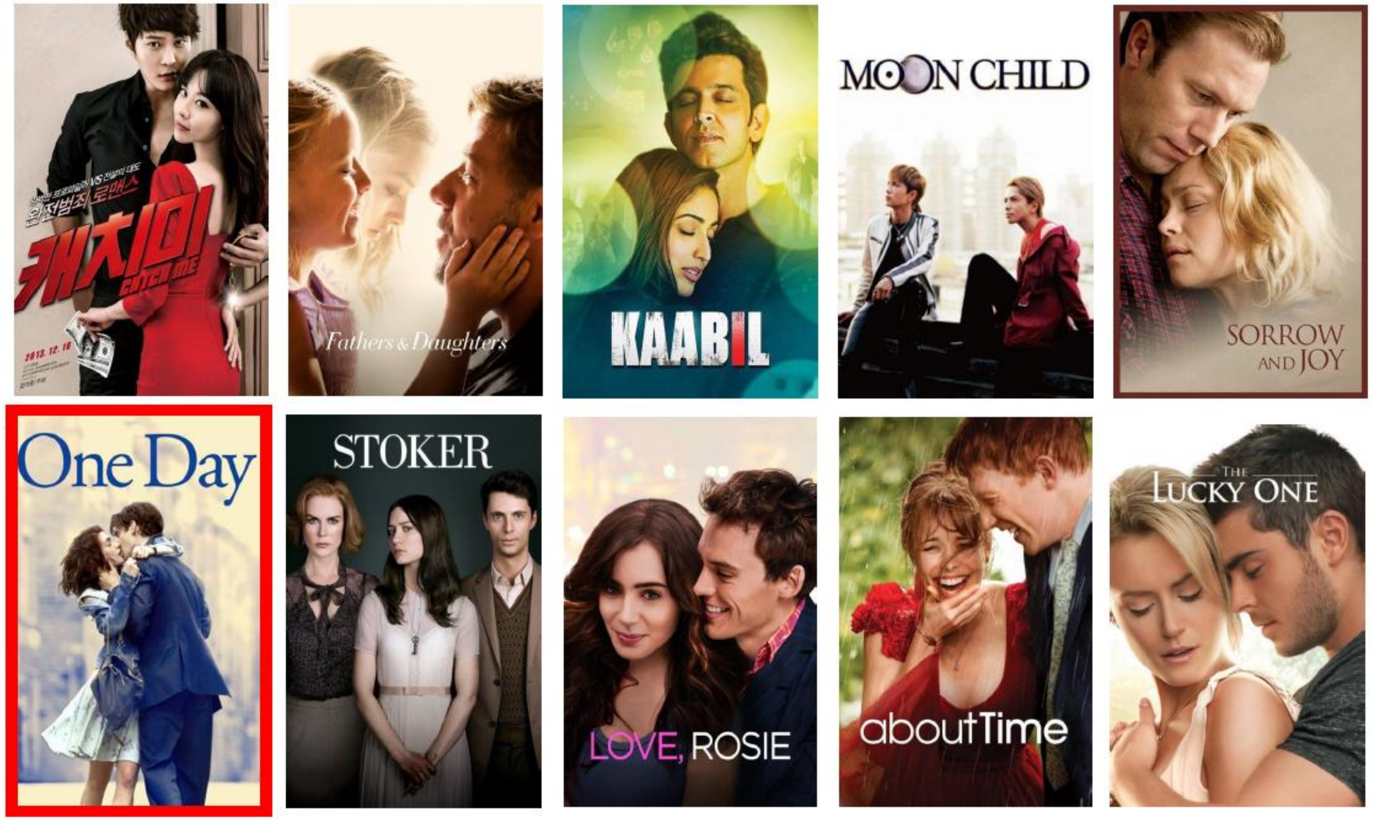


Fig 4a: Use the contextualized word embedding of the keyword ‘*love*’ in the context of a romantic movie *One Day (2009)* to retrieval movie posters. The ground truth shows up as the top 6th .

Conclusion

- We propose to pretrain a transformer network with masked language modeling and visual grounding objectives tailored to the film industry context.

- Compared to BERT embedding, the contextualized and visual grounded representation improve the box office prediction accuracy.

- We constructed a large dataset for community to continue exploring the movie box office prediction task.

Paper:

GitHub:

Contact information:
Qin Chao chao0009@e.ntu.edu.sg
Eunsoo Kim eunsoo@ntu.edu.sg
Boyang Li boyang.li@ntu.edu.sg