Multi-Modal Game Recommendations on Steam

CS608-Recommender Systems – Project #2 Final Presentation

Group 6

Nayan Rajendra Rokhade (nayanr.2022@mitb.smu.edu.sg)
Neel Ketan Modha (neelmodha.2022@mitb.smu.edu.sg)
Nowshad Shaik (nowshads.2022@mitb.smu.edu.sg)
Siddharth Singh (siddharths.2022@mitb.smu.edu.sg)

Recap of our Problem Statement

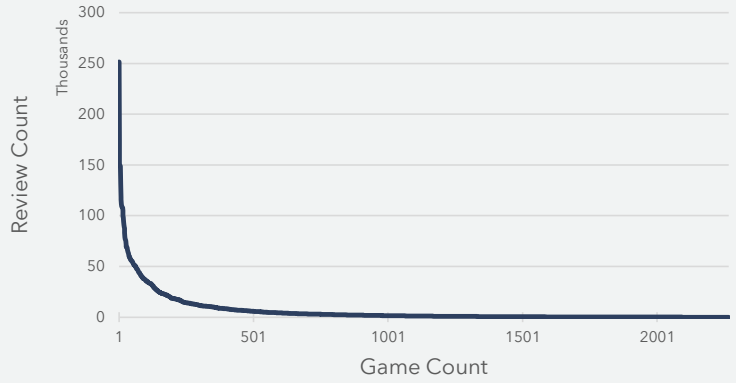
Comparing the effectiveness of traditional game recommendation systems to multi-model recommendations systems

- Develop a game recommendation system based on Steam's game library, which boasts >50,000 games. Steam collects a substantial amount of data on each user, game and users' interaction with games
- Create a recommendation system that goes **beyond rating-based suggestions**, to **incorporate visual elements** gamers consider in the adoption decisions
- Our proposed approach involves incorporating multiple modalities into the recommendation model. Specifically, we intend to utilize **game image data** to **enhance recommendations generated**

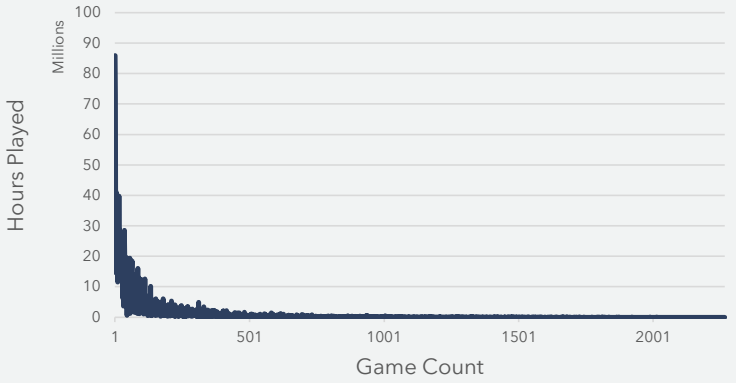
Refinement: Focusing on smaller, less popular titles

We refined our recommendation system with a greater focus on games which may not normally be recommended to users

Number of Reviews Published (desc.)



Number of Hours Played (desc.)



- Majority of game reviews published, and game hours played lie among the top 20% of games in scope (more on this during the data pre-processing section)
- By focusing on titles which are less popular, **our recommendation system aims to help smaller game developers gain traction on the Steam Marketplace**

Project Methodology

Data Pre-Processing

Collaborative Filtering:

Based on Users' Game Ratings – *EASE (cornac)*

Content-Based Filtering:

Based on Game Image Similarity – *Vector Similarity (Milvus)*

Collaborative + Content-Based Filtering:

Based on Users' Game Ratings + Game Image Features – *VBPR (cornac)*

Recommendation Post-Processing

Recommendation System Demo App. and Model Evaluation

Data Pre-Processing

Image scraping and data cleaning before training our models. *Dataset: Game Ratings on Steam (sourced from Kaggle[1])*

Image Scraping *(recap)*

- Game images were scraped from the Steam web store, and resized to 300x300px

Example:

Data Cleaning

- Dataset contained 14M game reviews across 2,264 games, despite there being >50K titles on the Steam web store
- A significant portion of the titles consisted of downloadable content/game addons which are not in scope for this analysis
- We **focused on these 2,264 games** and their reviews when training our recommendation models

Ground Truth Mapping

- "is_recommended" helps filter ratings into 1-2 and 3-5 stars.
- Higher ratings are given when "is_recommended" is True and more hours are played.
- Lower ratings are given when "is_recommended" is False and fewer hours are played.

Ground Truth Rating	is_recommended?	hours_played
5	True	>= 181
4	True	58 - 180
3	True	< 58
2	False	< 118
1	False	>= 118

STEAM

1: Game Recommendations on Steam: https://www.kaggle.com/datasets/antonkozyriev/game-recommendations-on-steam, more dataset details in Appendix slides

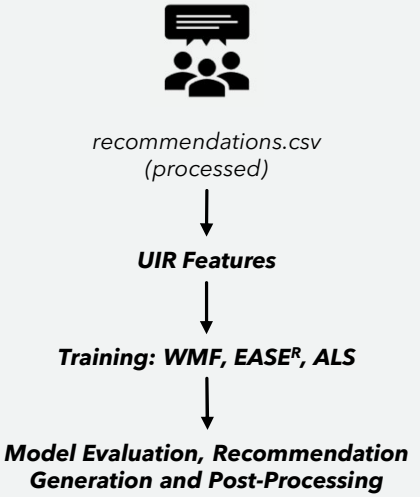
2: Threshold values calculated via percentile analysis conducted by the team

Collaborative Filtering

Best performing model was Embarrassingly Shallow Auto Encoder (EASE^R)

- Models tested: *WMF, EASE^R (cornac), ALS (Implicit)*
- UIR Features – From Game reviews, based on the defined ground truth
- EASE^R model performed better when compared to other models
  - Hyperparameter tuning was done on EASE^R, although we did not see much improvement to evaluation metrics[1]
- Snapshot of Results (model in **green** is the final chosen one):

Model	Params	AUC	Recall@20	NCRR@20	NDCG@20	Harmonic Mean
WMF	k = 50, LR = 0.01	0.5634	0.1806	0.0246	0.0572	0.0611
ALS	Default	0.5921	-	-	0.0695	0.1244
EASE^R	**Lamb = 1000**	**0.5308**	**0.2259**	**0.0751**	**0.1089**	**0.1388**
EASE^R	Lamb = 1250	0.5306	0.2257	0.0748	0.1082	0.1382

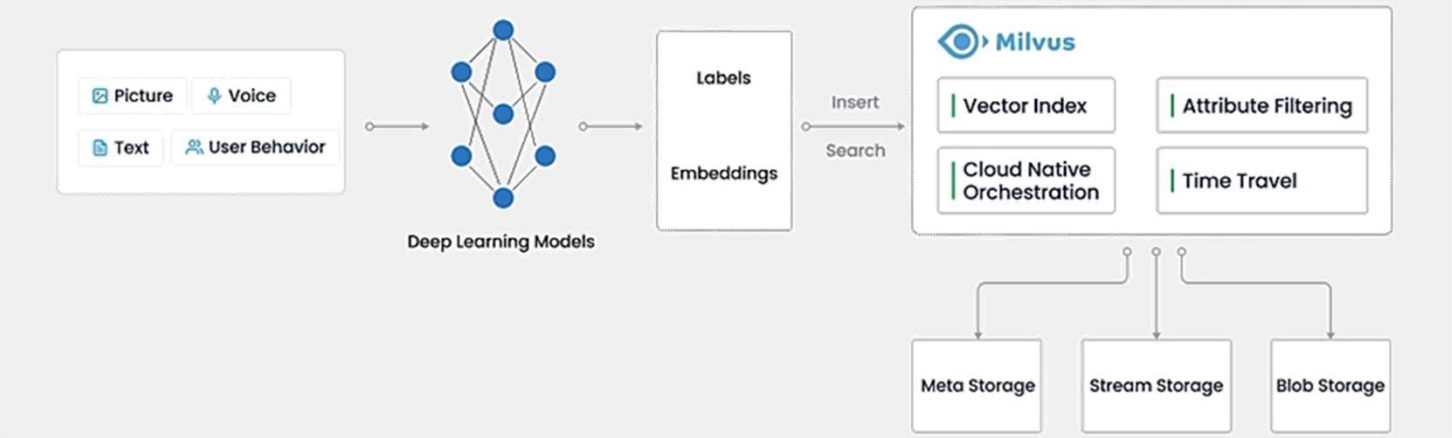


1: We hypothesize this is due to the Power Law impact – Discussed further in later slides

Content-based Filtering Model

We used Milvus to support calculations for image similarity

- Milvus is a open source vector database to store, index, and **manage massive embedding vectors** generated by deep neural networks and other machine learning (ML) models
- As a database specifically designed to handle queries over input vectors, it is **capable of indexing vectors on a trillion scale**
- Vector similarity search is the process of comparing a vector to a database to **find vectors that are most similar to the query vector**

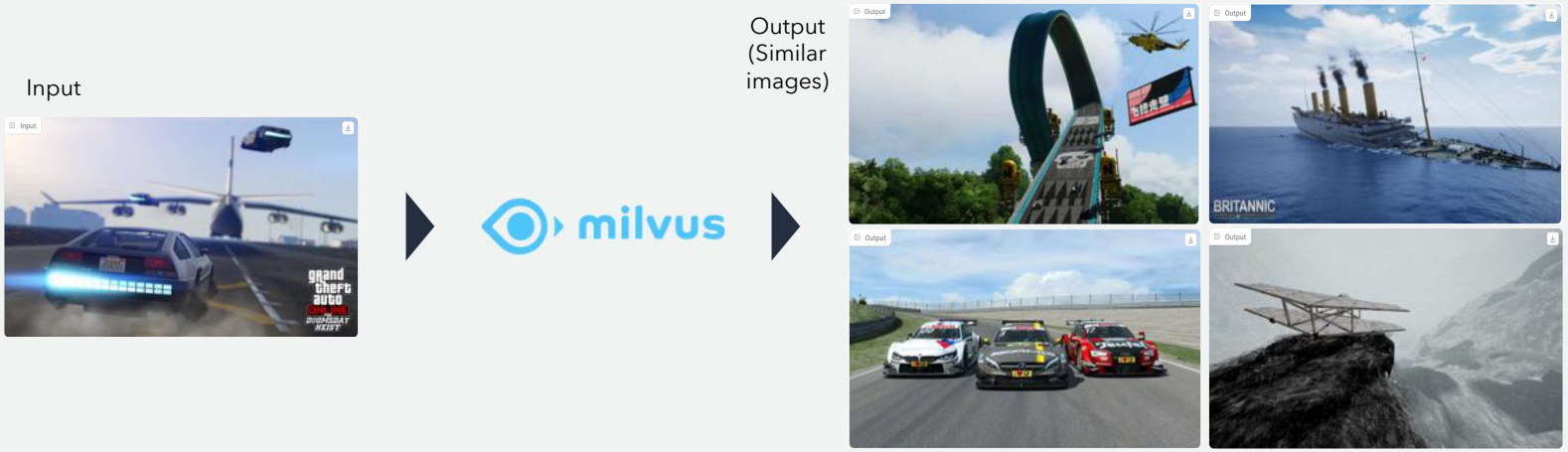


Content-based Filtering Model

Image Similarity using Vector Similarity calculated with Milvus

Workflow

Example

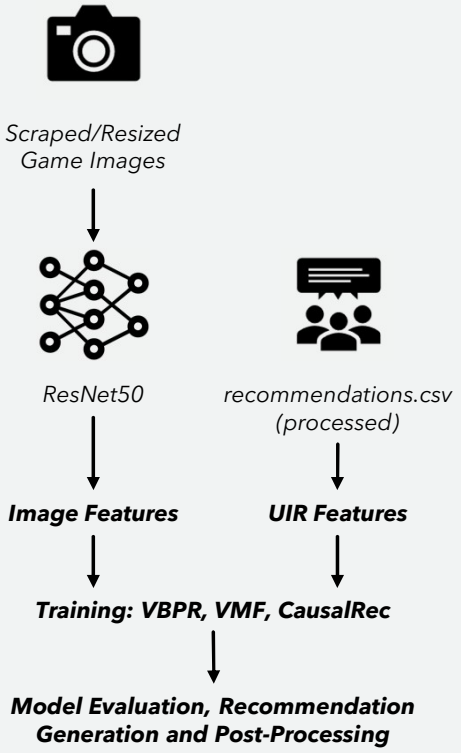


Content-based + Collaborative Filtering

Best performing model was Visual Bayesian Personalized Ranking (VBPR)

- Models tested: *VMF*, *CausalRec*, *VBPR (cornac)*
- UIR Features – same dataset as the one used in collaborating filtering
- Image Features – extracted using ResNet50 on for the 2,264 selected games' images
- VBPR performed better when compared to other models
  - Hyperparameter tuning was done on VBPR, although we did not see much improvement to evaluation metrics[1]
- Snapshot of Results (model in **green** is the final chosen one):

Model	Params	AUC	Recall@20	NCRR@20	NDCG@20	Harmonic Mean
VMF	Default	0.6572	0.0069	0.0016	0.0028	0.003544
CausalRec	Default	0.9278	0.2038	0.0370	0.0726	0.085496
VBPR	**Default**	**0.9282**	**0.2408**	**0.0629**	**0.1011**	**0.128951**
VBPR	$k = 10, k_2 = 10$	0.9281	0.2403	0.0629	0.1010	0.128874
VBPR	$k = 20, k_2 = 10$	0.9280	0.2405	0.0629	0.1010	0.128887



1: We hypothesize this is due to the Power Law impact – Discussed further in later slides

Recommendation Post-Processing

A heuristic approach to tailor our recommendations to focus on smaller, less popular games

- While we trained our models on all 2,264 games, we excluded some games from our final recommendations to better serve our intent to recommend less popular games
- Employing an approach called **Serendipitous Discovery**[1] – to ensure that games recommended to users would most likely never be recommended in a normal recommender system
  - Selectively removing certain items from the final recommendations based on defined rules:

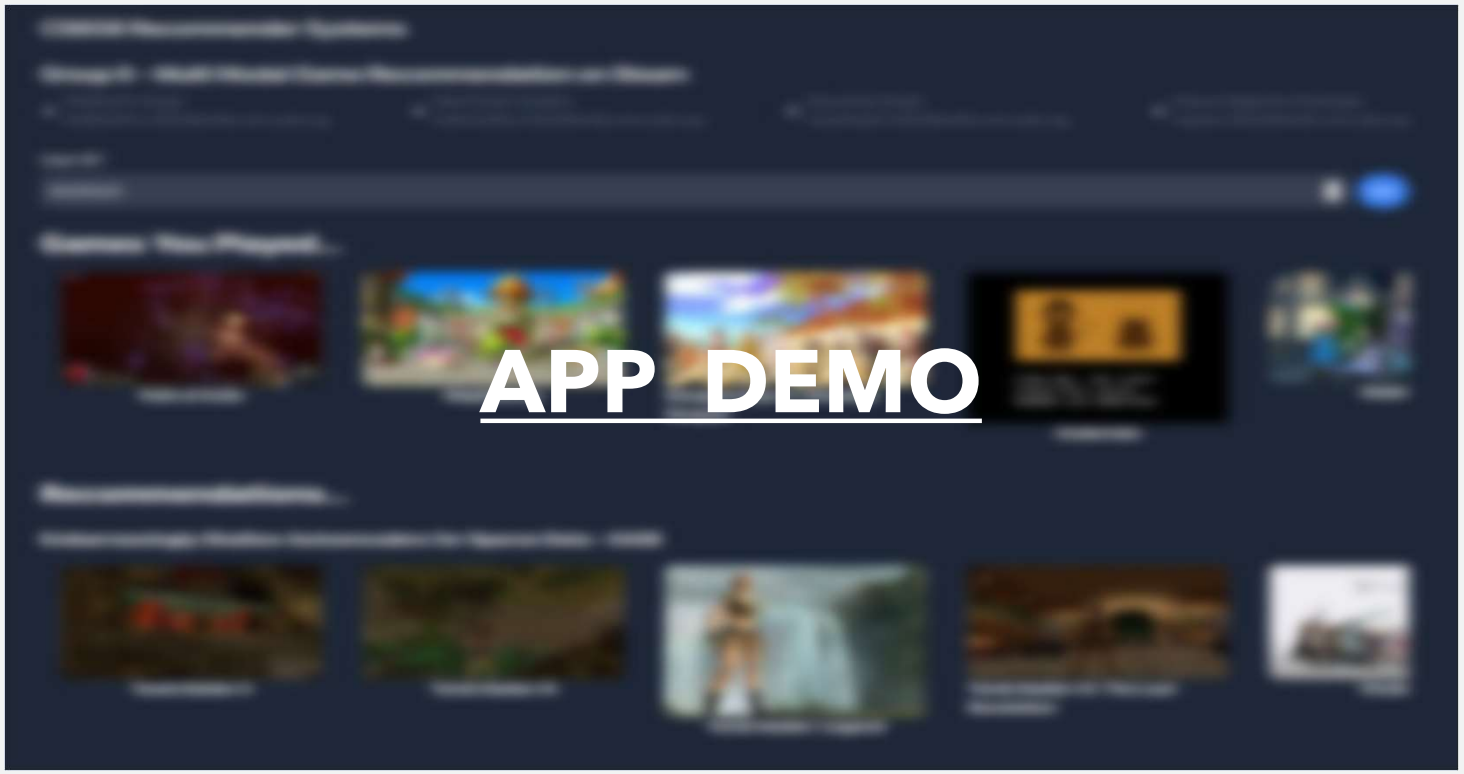
1. Removing Games that the user has already played before *(standard)*

2. Removing games whose number of reviews falls outside of a certain threshold
   - 2.1 Removing Games that have **< 100 reviews** (very new games, anomalies, etc.)
   - 2.2. Removing Games that have **> 7000 reviews** (80th percentile for number of reviews)

- ***Result:*** *Brought much needed diversity to initial recommendations, which were heavily biased towards the popular games due to the strong Power Law impact*

1: Refer to references 3 and 4 in References slide

Generating Recommendations



Evaluation of Models

Quantitative and Human Evaluation provided somewhat opposite conclusions

Quantitative Evaluation

- Based on evaluation metrics, VBPR performed very close to EASE^R on evaluated metrics
- Results:

Model	AUC	Recall@20	NCRR@20	NDCG@20
BPR	0.7837	0.1768	0.0352	0.0689
EASE^R	**0.5308**	**0.2259**	**0.0751**	**0.1089**
VBPR	**0.9282**	**0.2408**	**0.0629**	**0.1011**

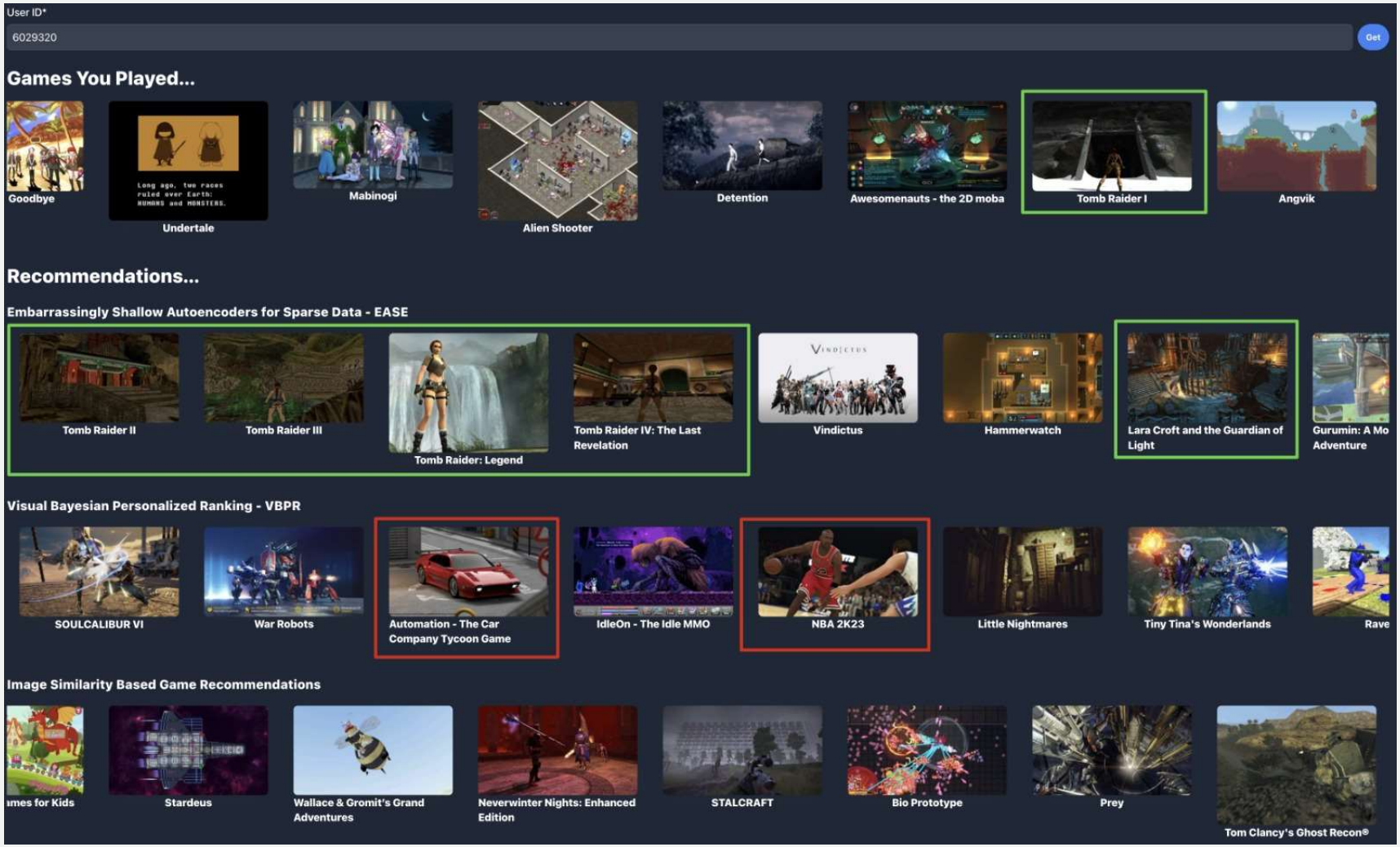
- With EASE^R being a much more powerful model than standard BPR, we hypothesize that the presence of visual features in VBPR helped to bridge the gap

Human Evaluation

- Based on analysis of sample recommendations generated after post-processing, we felt EASE provided better, more intuitive recommendations compared to VBPR
- Example:

On the next slide

- One potential reason for this is the quality of image features based on ResNet50, which is trained on real-world items instead of game images


User ID*
6029320
Get
Games You Played...
Goodbye
Long ago, two races ruled over Earth: HUMANS and MONSTERS.
Undertale
Mabinogi
Alien Shooter
Detention
Awesomenauts - the 2D moba
Tomb Raider I
Angvik
Recommendations...
Embarrassingly Shallow Autoencoders for Sparse Data - EASE
Tomb Raider II
Tomb Raider III
Tomb Raider: Legend
Tomb Raider IV: The Last Revelation
Vindictus
Hammerwatch
Lara Croft and the Guardian of Light
Gurumin: A Mo Adventure
Visual Bayesian Personalized Ranking - VBPR
SOULCALIBUR VI
War Robots
Automation - The Car Company Tycoon Game
IdleOn - The Idle MMO
NBA 2K23
Little Nightmares
Tiny Tina's Wonderlands
Rave
Image Similarity Based Game Recommendations
ımes for Kids
Stardeus
Wallace & Gromit's Grand Adventures
Neverwinter Nights: Enhanced Edition
STALCRAFT
Bio Prototype
Prey
Tom Clancy's Ghost Recon®

Further Insights on using Image Modality (1/2)

Visual BPR does show greater contribution to ratings via Visual factors as compared to CF factors for top-rated games

Using User 19 as an example:

All Games Rated >= 3

Title	Genre
Dying Light 2 Stay Human	**FPS/RPG**
Red Dead Redemption 2	**RPG**
Papers Please	Strategy
STAR WARS Jedi: Fallen Order Deluxe Edition	**RPG**
Zombie Army Trilogy	**RPG**
God of War	**RPG**
S.T.A.L.K.E.R.: Shadow of Chernobyl	**FPS**
Viscera Cleanup Detail	Simulation
Labyrinthine	Puzzle

Top 10 Recommendations

Rating	CF Contribution	Visual Contribution	Title	Genre
0.798	-0.564	**1.362**	**Ravenfield**	**FPS**
0.672	-0.400	1.073	SOULCALIBUR VI	Fighting
0.546	-0.489	**1.036**	**Tiny Tina's Wonderlands**	**FPS/RPG**
0.394	-1.137	**1.531**	**Mabinogi**	**RPG**
0.380	-0.737	**1.117**	**War Robots**	**FPS**
0.244	-0.620	0.865	NBA 2K23	Sports
0.228	-0.446	0.675	Call to Arms	Strategy
0.224	-0.981	**1.205**	**Freedom Planet**	**RPG**
0.118	-0.471	0.590	Little Nightmares	Horror
0.108	-0.786	**0.894**	**Call of Juarez: Gunslinger**	**FPS**

- User rated RPG/FPS games highly
- Based on this, our VBPR model was also able to recommend several games in the FPS/RPG genre, which is greatly influenced by visual factors vs. collaborative filtering factors

Further Insights on using Image Modality (2/2)

Games recommended using image similarity were in reality, a mixed-bag of similar and dissimilar games

Similar games

Input - Cities Skyline (City Simulation)

Output – Age of Empires (Empire Builder)

Not so similar games

Input - Little Nightmares (Horror)

Output – Battle Chasers (Fantasy RPG)

- Single image does not represent the actual gameplay
- When given an input from Cities Skyline, Age of Empires is output as the most similar looking game – Positive example, as both games have some sort of infrastructure building gameplay
- Conversely, when given an input image Little Nightmares, which is a horror game, Battle Chasers is output which is a Fantasy RPG – Negative example, as both are completely different genres, but may be associated due to the "darkness" of the games
- Based on the input image chosen out of all the available images the recommendations will vary

Areas for further Exploration

- Additional Modalities we could not explore
  - Text Modality – Using valuable game genre/description data to supplement content-based filtering
  - Social Network Modality – Access to user-user friend relationship data could make recommendations even more personalized to a users' social setting
- Ensemble Vectors to better characterize a game
  - Instead of just a single image vector for games, an ensemble vector would be able to quantify multiple images for a single game + descriptive, genre (and more) features (similar to diffusion models for generating new images)
- Exploring other ways to mitigate impact of Power Law
  - Personalization of ratings by collecting and referencing user-preference data
  - Selective imputation of ratings for games which have very few ratings (or conversely, removing ratings for games which have too many)

Key Learnings from this Project

- Image Feature extraction using ResNet50 showed decent performance, but could be improved
  - Fine tuning an existing pre-trained model for game images or using a custom CNN trained on game images would provide more relevant features for VBPR
- Power Law heavily skewed our initial set of recommendations
  - Initially, our recommended games for each user seemed very similar for the top *k* games
  - We realized that due to the presence of several games with a large amount of ratings, our recommendations were very biased
  - Explored multiple options to "un-bias" the recommendations, eventually ending up un-biasing via our post-processing steps
- Milvus - Storing the embeddings of the content (images/text) in vector databases can greatly speed up vector similarity calculations

References

1. Jung, KY. (2006). Content-Based Image Filtering for Recommendation, Foundations of Intelligent Systems, ISMIS 2006, Volume 4203. https://doi.org/10.1007/11875604_36
2. Wang, D. Moh, M. Moh, T.S. (2020). Using Deep Learning and Steam User Data for Better Video Game Recommendations, Proceedings of the 2020 ACM Southeast Conference (ACM SE '20). Association for Computing Machinery, pg. 154–159. https://doi.org/10.1145/3374135.3385283
3. Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001). Stereotype and Diversify: Learning to Route in Recommender Systems. In Proceedings of the 2001 ACM Conference on Information and Knowledge Management (CIKM '01) (pp. 355-362). ACM. https://dl.acm.org/doi/pdf/10.1145/371920.372071
4. Kamishima, T., Akaho, S., Asoh, H., Sakuma, J., & Ono, S. (2014). Serendipitous Recommendations with Personalized Diversity. In Proceedings of the IEEE International Conference on Data Mining (ICDM) (pp. 477-486). IEEE. http://proceedings.mlr.press/v81/kamishima18a.html
5. Packages and Software used:
    a. Cornac: https://github.com/PreferredAI/cornac
    b. Implicit: https://github.com/benfred/implicit/
    c. Milvus: https://milvus.io/
6. Data Sources:
    a. Kaggle Game Recommendations: https://www.kaggle.com/datasets/antonkozyriev/game-recommendations-on-steam
    b. Steam Store: https://store.steampowered.com/

Why Multi-Modal Game Recommendations?

Our motivation to address this problem

- When it comes to game recommendations, traditional approaches have **relied primarily on user ratings and individual player histories**.
- However, these methods may fail to capture **other crucial factors that are important to players when selecting or playing games** (e.g., how visually appealing a game may be, or how interesting a game sounds based on its description)
- Different modalities' impact to game recommendation systems has been **underexplored in previous research**, and has hence motivated our decision to focus on this area

Problem Statement

Comparing the effectiveness of traditional game recommendation systems to multi-model recommendations systems

- Our objective is to address this challenge by developing a game recommendation system based on a gaming library. One such library is Steam, which boasts >50,000 games. Steam **collects a substantial amount of data on each user, game and users' interaction with games**
- This presents an opportunity to leverage this information to create a recommendation system that **goes beyond simple rating-based suggestions**, to **incorporate more elements gamers consider in the adoption decisions**
- Our proposed approach involves incorporating multiple modalities into the recommendation model. Specifically, we intend to **utilize game images and potentially text metadata to enhance the recommendations generated**

Dataset Used

Game Ratings on Steam (sourced from Kaggle[1]) augmented with scraped game images

- The dataset contains **48,675 games** and **~13.8m reviews** from **~6.8m users** across 4 files:
  - **games.csv** – Game information: app_id, release date, available platform, overall ratings, price, discounts etc.
  - **games_metadata.json** – Game metadata: title, description, tags
  - **recommendations.csv** – Published reviews: app_id, user_id, hours played, is review helpful, is review funny, etc.
  - **users.csv** – User information: no. of games user owns, how many reviews published
- Game images were scraped from the Steam web store, capturing all screenshots provided by the game publishers for marketing purposes:
  - app_id used to navigate to the game's web storefront and retrieve HTML data
  - HTML processed to retrieve image links which are then used to download screenshots (1080p), forming a ~200GB dataset
  - Images are then resized to 300x300px, reducing dataset size to ~10GB

1: Game Recommendations on Steam: https://www.kaggle.com/datasets/antonkozyriev/game-recommendations-on-steam

A Snapshot of our Data

Example:

app_id	title	date_release	win	mac	linux	rating	positive_ratio	user_reviews	price_final	price_original	discount	steam_deck
13500	Prince of Persia: Warrior Within	11/21/2008	TRUE	FALSE	FALSE	Very Positive	84	951	9.99	9.99	0	TRUE

Game Metadata

{

"app_id" : 13500,

"description" : "Enter the dark underworld of Prince of Persia Warrior Within, the sword-slashing sequel to the critically acclaimed Prince of Persia: The Sands of Time™. Hunted by Dahaka, an immortal incarnation of Fate seeking divine retribution, the Prince embarks upon a path of both carnage and mystery to defy his preordained death.",

"tags" : ["Action", "Adventure", "Parkour", "Third Person", "Great Soundtrack", "Singleplayer", "Platformer", "Time Travel", "Atmospheric", "Classic", "Hack and Slash", "Time Manipulation", "Gore", "Fantasy", "Story Rich", "Dark", "Open World", "Controller", "Dark Fantasy", "Puzzle"]

}

Game Reviews

app_id	helpful	funny	date	is_recommended	hours	user_id	review_id
13500	2	0	2015-05-30	True	0.4	1564069	13207266
13500	4	0	2014-07-25	False	15.2	5298176	13207274
13500	0	0	2021-12-15	True	19.4	5575321	13207284
13500	13	0	2022-08-11	True	10.8	1963218	13207302
13500	0	0	2021-05-18	False	10.9	4086771	13207328
...	...	...	...	...	...	...	...
13500	0	0	2021-11-27	True	17.2	3849626	13280812
13500	0	0	2020-06-12	True	8.8	24655	13280819
13500	2	0	2020-02-26	True	10.8	6660202	13280822
13500	0	0	2022-12-28	True	9.5	6781281	13281076
13500	0	0	2020-11-27	True	2.3	3003072	13281171

Game Images (Scraped)

Model Training Strategy

Training models that are using image modality in some form versus models that are only using ratings

- We propose training the following categories of models, with examples of specific model types are planning to experiment with:
  - **Ratings Modality ("Control" model)** – e.g., Weighted Matrix Factorization, Bayesian Personalized Ranking (*cornac*)
  - **Image Modality** – e.g., Convolutional Neural Network to capture features, to compute Image similarity based on Cosine similarity (*Tensorflow*)
    - The idea here is that users may be attracted to how a particular game and its gameplay looks
  - **Image + Ratings Modality** – e.g., Convolutional Matrix Factorization, Visual Bayesian Personalized Ranking (*cornac*)
  - **Ensemble Models** – Combining models results across the individual Ratings and Image modalities to generate a single result
    - One possible way: **WMF ratings added to the product of WMF ratings and image similarity scores**, to ensure game recommendations with gameplay (images) similar to what a user already likes will be given a higher ranking (inspired by Decoupled Collaborative Ranking[1])

1: Jun Hu and Ping Li. 2017. Decoupled Collaborative Ranking. In Proceedings of the 26th International Conference on World Wide Web (WWW '17): https://doi.org/10.1145/3038912.3052685

Model Evaluation Strategy

Testing the output of our models, which will be the top *k* recommended games that the user has not played

- With ***k*** yet to be determined, we plan to compare our models using all available/taught ranking metrics:
  - **AUC, Precision@*k*, Recall@*k*, F1 Score, MAP, NCRR@*k*, NDCG@*k***
- The idea here is to find out whether certain types of models are performing better on certain metrics
- We also want to determine which is the best overall performing model
  - Possibly using a **Weighted Average**, or **Harmonic Mean** of the ranking metrics assessed above
- To establish the **Ground Truth** on our validation/test datasets, we will determine the confidence-level based on the "hours" column of the dataset (number of hours each user has played a particular game)

Confidence

for example: $$c_{ij} = \begin{cases} a, \text{if } r_{ij} > 5 \text{ hours of gameplay} \\ b, \text{otherwise} \end{cases}$$

Potential Experiments

Additional areas we aim to explore, and possibly integrate into our final output

- Exploring Text Modality – more models to test and evaluate:
  - **Text Modality only** – e.g., performing Topic modelling and ranking (*Gensim*)
  - **Text + Ratings Modality** – e.g., using Collaborative Topic Modelling (*cornac*)
  - **Text + Images + Ratings Modality** – *Still exploring feasibility of this approach*
  - **Ensemble** – Adding Text Modality model outputs into our Ensemble described earlier
  - Game text would be based off the provided game "title", "description" and "tags" – with the idea being that certain users may be attracted to the publisher's game marketing messages or tagged genres
- Recommending the top *k* games which the user has **not played** and are in **new genres** that (we feel) the user would like, based on ratings, image features and potentially text features:
  - Will require the evaluation of ranking metrics as described earlier, and additional **novelty metrics** (e.g., Novelty Score[1])
- Variations in calculating the Ground Truth on our validation/test datasets:
  - Using the "is_recommended" and "helpful" columns provided, to augment the ground truth
  - Alternative formulas for calculating c_{ij}
- Speeding up recommendation generation for similar images/documents:
  - Using Vector Databases (e.g., Milvus[2]) for large image datasets

1: Novelty and Diversity in Recommender Systems" by Gediminas Adomavicius and YoungOk Kwon (2008): https://ieeexplore.ieee.org/abstract/document/4781121

2: Milvus; Vector database built for scalable similarity search: https://milvus.io/

References

- Hu, J., & Li, P. (2017, April). Decoupled collaborative ranking. In Proceedings of the 26th International Conference on World Wide Web (pp. 1321-1329)
- Jung, KY. (2006). Content-Based Image Filtering for Recommendation. In: Esposito, F., Raś, Z.W., Malerba, D., Semeraro, G. (eds) Foundations of Intelligent Systems. ISMIS 2006. Lecture Notes in Computer Science(), vol 4203. Springer, Berlin, Heidelberg.
- Dylan Wang, Melody Moh, and Teng-Sheng Moh (2020). Using Deep Learning and Steam User Data for Better Video Game Recommendations. In Proceedings of the 2020 ACM Southeast Conference (ACM SE '20). Association for Computing Machinery, New York, NY, USA, 154–159.
- Fasiha Ikram, Humera Farooq, "Multimedia Recommendation System for Video Game Based on High-Level Visual Semantic Features", Scientific Programming, vol. 2022,