Virtual Reality Application
Comfort Level Rating Evaluator

LYU2201 Final Year Project Term2 Presentation
Junjie XIE
Supervised by Prof. Michael R. Lyu
Review

- VR application comfort levels rating: Comfortable, Moderate, Intense
- Determining the comfort level of VR applications is time-consuming
- Provide a quick tool to evaluate the application comfort level
- Using description to make prediction
Text Encoder

- Count Vector
- TF-IDF Vector
- Word Embedding
- Transformer

Projection of the embedding vectors to 2D

Transformer encoder and decoder architecture
Model
- Bayes Classifier
- Shallow Neural Network
- Convolutional Neural Network
- Transformer-CNN

Convolutional Neural Network

Transformer-CNN
Experiment

Description - Count Vector → Text representation → Naive Bayes Classifier → Output

Description - TF-IDF Vector → Text representation → Naive Bayes Classifier → Output

Description-Bayes Classifier

Description - Count Vector → Text representation → Shallow Neural Network → Output

Description - TF-IDF Vector → Text representation → Shallow Neural Network → Output

Description-Shallow Neural Network
Experiment

Description-Convolutional Neural Network

Description-Transformers

Transformer 13 encoder layer hidden states
## Experiment Result

### Description-Bayes Classifier

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description Bayes Classifier (Count Vector)</td>
<td>0.7207</td>
</tr>
<tr>
<td>Description Bayes Classifier (Oversampled)</td>
<td>0.6723</td>
</tr>
<tr>
<td>Description Bayes Classifier (TF-IDF Vector)</td>
<td>0.6301</td>
</tr>
<tr>
<td>Description Bayes Classifier (TF-IDF Vector Oversampled)</td>
<td>0.5963</td>
</tr>
</tbody>
</table>

### Description-Shallow Neural Network

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description-Shallow Neural Network (Count Vector)</td>
<td>0.6483</td>
</tr>
<tr>
<td>Description-Shallow Neural Network (Oversampled)</td>
<td>0.6703</td>
</tr>
<tr>
<td>Description-Shallow Neural Network (TF-IDF Vector)</td>
<td>0.6192</td>
</tr>
<tr>
<td>Description-Shallow Neural Network (TF-IDF Vector Oversampled)</td>
<td>0.6321</td>
</tr>
</tbody>
</table>

### Description-Convolutional Neural Network

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description-Convolutional Neural Network</td>
<td>0.7074</td>
</tr>
<tr>
<td>Description-Convolutional Neural Network (Oversampled)</td>
<td>0.7665</td>
</tr>
</tbody>
</table>

### Description-Transformers

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description-Transformer</td>
<td>0.8316</td>
</tr>
<tr>
<td>Description-Transformer (Oversampled)</td>
<td>0.8525</td>
</tr>
<tr>
<td>Description-Transformer-CNN</td>
<td>0.8452</td>
</tr>
<tr>
<td>Description-Transformer-CNN (Oversampled)</td>
<td>0.8398</td>
</tr>
</tbody>
</table>
To improve

- Reviews
- Images
- Videos
Multimodal Machine Learning

- Different modality data
- Text, Image, Video, Audio
- Internal structure and Data transformation are different
Fusion scheme

- Goal: Obtain more comprehensive information
- Fusion: Combine features from different modalities
- Early fusion
- Mid fusion
- Late fusion
Early Fusion

- Features from different modalities are fused in the input layer.

- Early fusion methods directly integrate information from different modalities into a complete feature vector.

- Usually suitable for cases where the modality differences are small, and the features are relatively simple.
Mid Fusion

- Fuse the features of different modalities in an intermediate layer.
- Improve the representation of features.
- Suitable for cases where the modality differences are large and the features are complex.
Late Fusion

- The features of different modalities are firstly fed into different classifiers for processing, and finally the results of different classifiers are fused.

- Avoid the conflict of features of different modalities.

- Suitable for the cases where the modality differences of data sources are large, the features are complex, but their correlation are strong.
Text-Text Learning

**App descriptions**

**Blade & Sorcery: Nomad**

- 18+ Extreme Violence
- 32,303 Ratings

The era of the VR weightless, wiggle-sword combat is over. Blade & Sorcery: Nomad is a medieval fantasy sandbox like no other, focusing on melee, ranged and magic combat that fully utilizes a unique and realistic physics driven interaction and combat system. Built exclusively for VR, collisions are dictated by fine hitboxes, objects have weight and follow the laws of physics, creatures have full body physics and presence, and blades can be used to penetrate soft materials or deflect magic. In Blade & Sorcery: Nomad, the combat is limited only by your own creativity. Choose your weapon, choose your stance, choose your fighting style; Be the powerful warrior, ranger or sorcerer you always dreamed of becoming!

**User reviews**

**Rydaddy** Mar 24 at 9:01 AM

★★★★★

**KMPace.921** Mar 12 at 3:41 AM

★★★★★

**Leafvr** 20 days ago

★★★★★

**BEST GAME EVER**

I recently bought blade and sorcery nomad and bought it for the mods but I find myself playing for hours in sandbox mode and in the dungeon mode having the most fun ever! I recommend buying this and the graphics look amazing.
Text-Text Learning - early fusion

1. **Reviews** → **Count Vector** → **Image representation** → **Naive Bayes Classifier** → **Output**
   - Description-Review Bayes Classifier (Count)

2. **Reviews** → **Count Vector** → **Text representation** → **Shallow Neural Network** → **Output**
   - Description-Review Shallow Neural Network (Count)

3. **Reviews** → **TF-IDF Vector** → **Image representation** → **Naive Bayes Classifier** → **Output**
   - Description-Review Bayes Classifier (TFIDF)

4. **Reviews** → **TF-IDF Vector** → **Text representation** → **Shallow Neural Network** → **Output**
   - Description-Review Shallow Neural Network (TFIDF)
Text-Text Learning – early fusion

Description
- Review CNN (Embedding)
- Review Transformer
- Description-Review Transformer

Description
- Review CNN (Embedding)
- Review Transformer
- Description-Review Transformer

Reviews
- Word Embedding
- Transformer

Description
- Word Embedding
- Transformer

Reviews
- Transformer

Description
- Transformer
# Text-text learning result

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description-Review Bayes Classifier (Count Vector)</td>
<td>0.7103</td>
</tr>
<tr>
<td>Description-Review Bayes Classifier (Count Vector Oversampled)</td>
<td>0.6981</td>
</tr>
<tr>
<td>Description-Review Bayes Classifier (TF-IDF Vector)</td>
<td>0.6502</td>
</tr>
<tr>
<td>Description-Review Bayes Classifier (TF-IDF Vector Oversampled)</td>
<td>0.6233</td>
</tr>
</tbody>
</table>

**Description-Review Bayes Classifier**

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description-Review SNN (Count Vector)</td>
<td>0.6501</td>
</tr>
<tr>
<td>Description-Review SNN (Count Vector Oversampled)</td>
<td>0.6788</td>
</tr>
<tr>
<td>Description-Review SNN (TF-IDF Vector)</td>
<td>0.6455</td>
</tr>
<tr>
<td>Description-Review SNN (TF-IDF Vector Oversampled)</td>
<td>0.6333</td>
</tr>
</tbody>
</table>

**Description-Review Shallow Neural Network**

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description-Shallow Neural Network (Count Vector)</td>
<td>0.6483</td>
</tr>
<tr>
<td>Description-Shallow Neural Network (Count Vector Oversampled)</td>
<td>0.6703</td>
</tr>
<tr>
<td>Description-Shallow Neural Network (TF-IDF Vector)</td>
<td>0.6192</td>
</tr>
<tr>
<td>Description-Shallow Neural Network (TF-IDF Vector Oversampled)</td>
<td>0.6321</td>
</tr>
</tbody>
</table>
## Text-text learning result

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description-Review CNN (Word Embedding)</td>
<td>0.7221</td>
</tr>
<tr>
<td>Description-Review CNN (Word Embedding Oversampled)</td>
<td>0.7443</td>
</tr>
</tbody>
</table>

Description-Review Convolutional Neural Network

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description-Review Transformer</td>
<td>0.8401</td>
</tr>
<tr>
<td>Description-Review Transformer (Oversampled)</td>
<td>0.8356</td>
</tr>
<tr>
<td>Description-Review Transformer-CNN</td>
<td>0.8397</td>
</tr>
<tr>
<td>Description-Review Transformer-CNN (Oversampled)</td>
<td>0.8434</td>
</tr>
</tbody>
</table>

Description-Review Transformers
Text-Text-Image

App descriptions

User reviews

Images
Image Encoder - MobileNetV2

Convolutional neural network architecture built on an inverted residual structure
Text-Text-Image

Description-Review-Image (Early-fusion)

Description-Review-Image MobileNetV2 (Mid-fusion)
Text-Text-Image

Description-Review-Image MobileNetV2 FastText (Mid-fusion)

Description-Review-Image MobileNetV2 Transformer (Late-fusion)
## Text-Text-Image Result

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description-Review-Image (Early-Fusion)</td>
<td>0.7322</td>
</tr>
<tr>
<td>Description-Review-Image MobileNetV2 (Mid-Fusion)</td>
<td>0.7613</td>
</tr>
<tr>
<td>Description-Review-Image MobileNetV2 FastText (Late-Fusion)</td>
<td>0.7829</td>
</tr>
<tr>
<td>Description-Review-Image MobileNetV2 Transformer (Late-Fusion)</td>
<td>0.8475</td>
</tr>
</tbody>
</table>

Description-Review-Image Learning
Text-Text-Video

Blade & Sorcery: Nomad

App descriptions

User reviews

App video

24
Video Encoder
S3D (Separable 3D convolutions)

a single convolution can be divided into two or more convolutions to produce the same output
Text-Text-Video

Description-Review-Video (Early-fusion)

Description-Review-Video (Mid-fusion)
Text-Text-Video

Description-Review-Video (Late-fusion)
## Text-Text-Video learning result

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description-Review-Video Cross Encoder (Early-Fusion)</td>
<td>0.8398</td>
</tr>
<tr>
<td>Description-Review-Video Cross Encoder (Mid-Fusion)</td>
<td>0.8492</td>
</tr>
<tr>
<td>Description-Review-Video Transformer (Late-Fusion)</td>
<td><strong>0.8514</strong></td>
</tr>
</tbody>
</table>
Conclusion

- Multiple modality information fed to our model is effective.
- And the more modality information the model is fed, the more positively correlated the final accuracy of the models.
- The transformer model fed with review, description, and video using late fusion has the highest accuracy 85%.
Discussion

- Qualified comfort level rating evaluator to help the users, developers, and platforms.

- To further improve our model, utilizing information such as the category tags, the music from the application.
Acknowledgement

- I would like to express my deep gratitude to my supervisor, Professor Michael, and Ms. Shuqing, one of Michael's PhD students, for their guidance and invaluable advice in helping me throughout this final year project.
Thank You!
Q&A