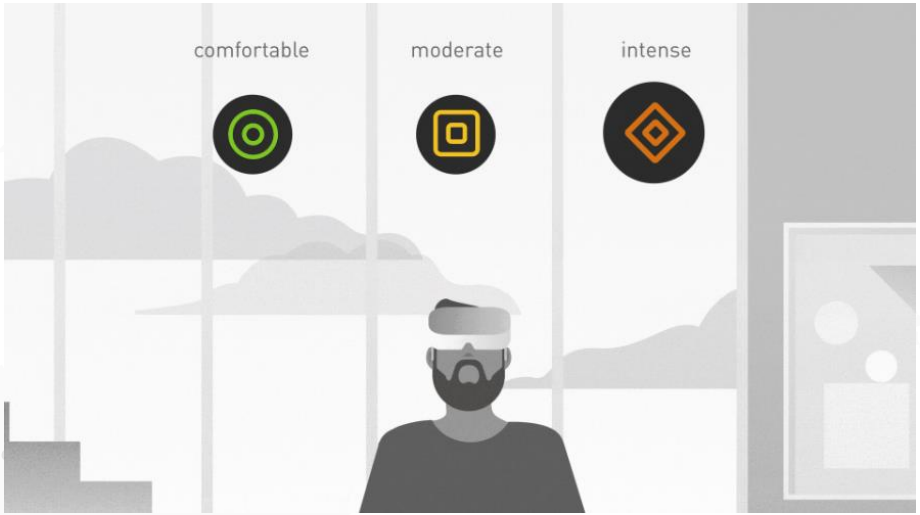
Virtual Reality Application Comfort Level Rating Evaluator

LYU2201 Final Year Project Term2 Presentation

Junjie XIE
Supervised by Prof. Michael R. Lyu

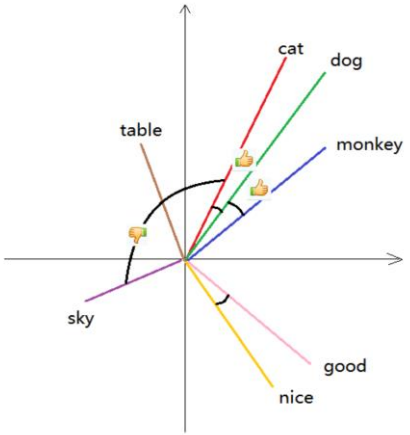
Review

- 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

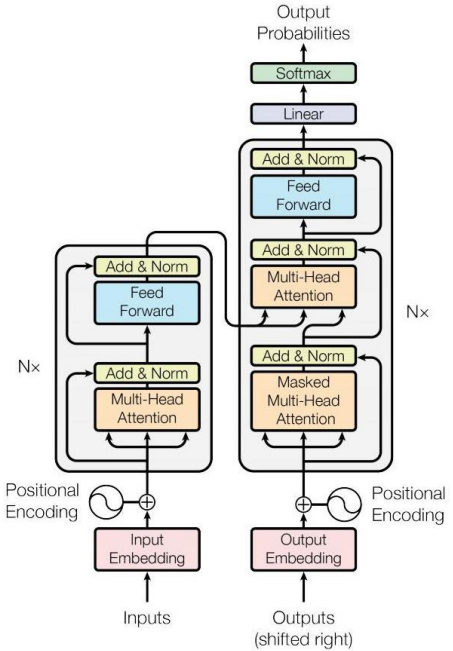


Text Encoder

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



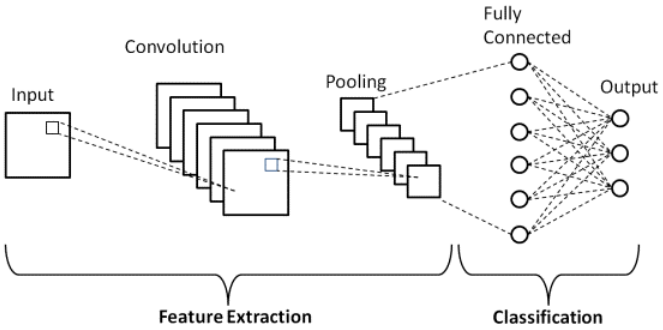
Projection of the embedding vectors to 2D



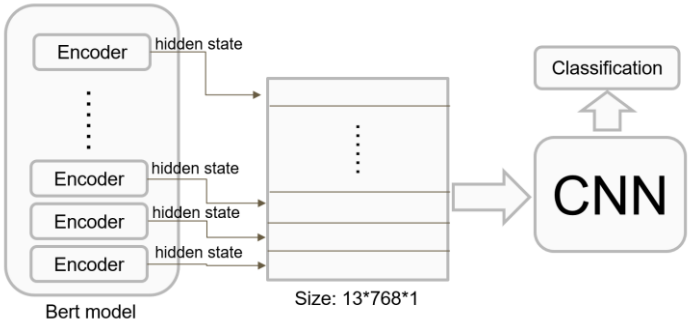
Transformer encoder and decoder architecture

Model

- Bayes Classifier
- Shallow Neural Network
- Convolutional Neural Network
- Transformer-CNN



Convolutional Neural Network



Transformer-CNN

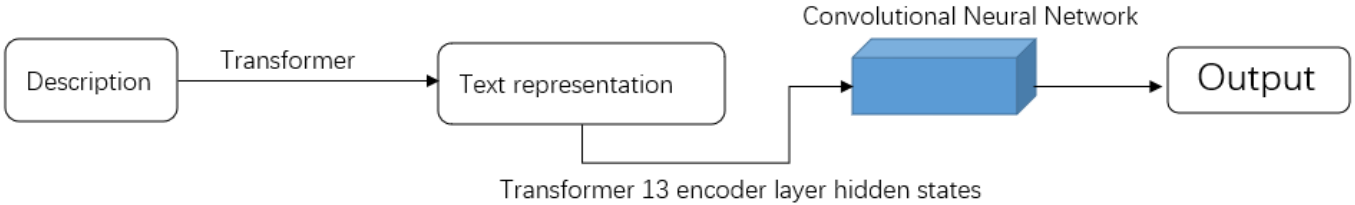
Experiment

Description-Bayes Classifier

Description-Shallow Neural Network

Experiment

Description-Convolutional Neural Network



Description-Transformers

Experiment Result

Model	Accuracy
Description Bayes Classifier (Count Vector)	0.7207
Description Bayes Classifier (Count Vector Oversampled)	0.6723
Description Bayes Classifier (TF-IDF Vector)	0.6301
Description Bayes Classifier (TF-IDF Vector Oversampled)	0.5963

Description-Bayes Classifier

Model	Accuracy
Description-Shallow Neural Network (Count Vector)	0.6483
Description-Shallow Neural Network (Count Vector Oversampled)	0.6703
Description-Shallow Neural Network (TF-IDF Vector)	0.6192
Description-Shallow Neural Network (TF-IDF Vector Oversampled)	0.6321

Description-Shallow Neural Network

Model	Accuracy
Description-Convolutional Neural Network	0.7074
Description-Convolutional Neural Network (Oversampled)	0.7665

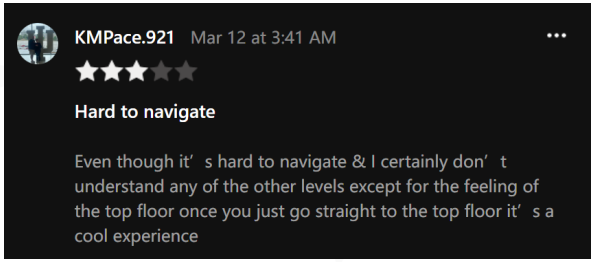
Description-Convolutional Neural Network

Model	Accuracy
Description-Transformer	0.8316
Description-Transformer (Oversampled)	0.8325
Description-Transformer-CNN	0.8432
Description-Transformer-CNN (Oversampled)	0.8398

Description-Transformers

To improve

- Reviews
- Images
- Videos



Review sample

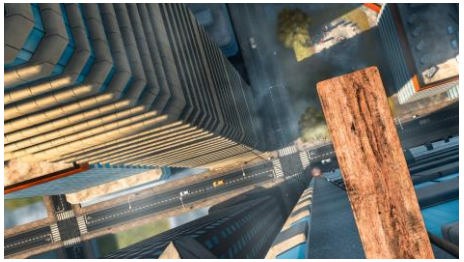
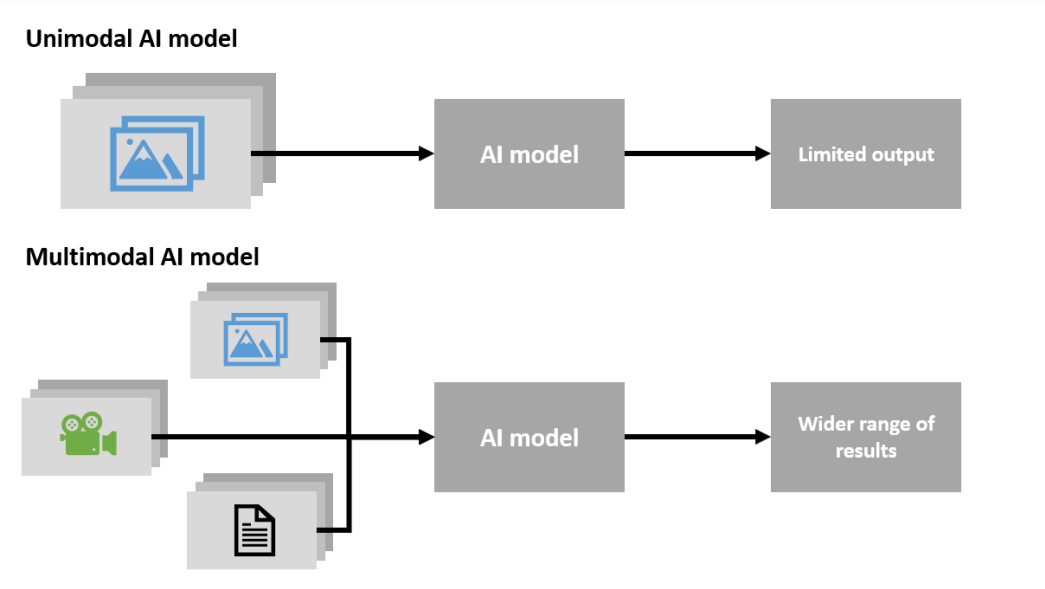


Image sample

Video sample

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.

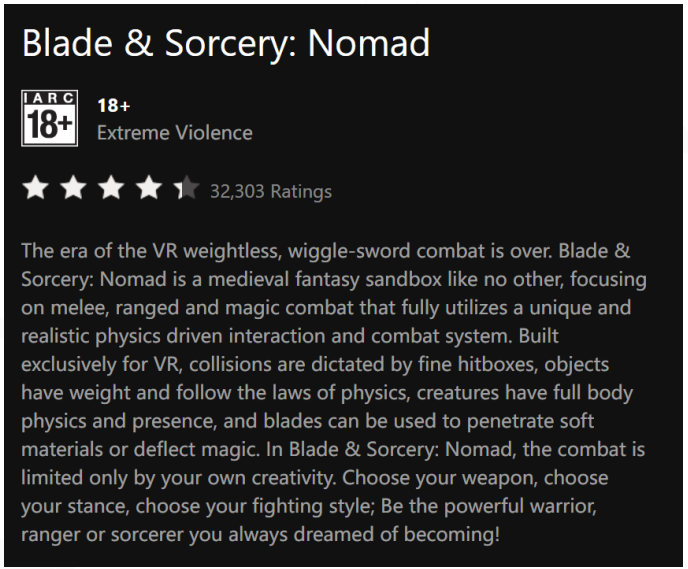
Mid Fusion

- 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.

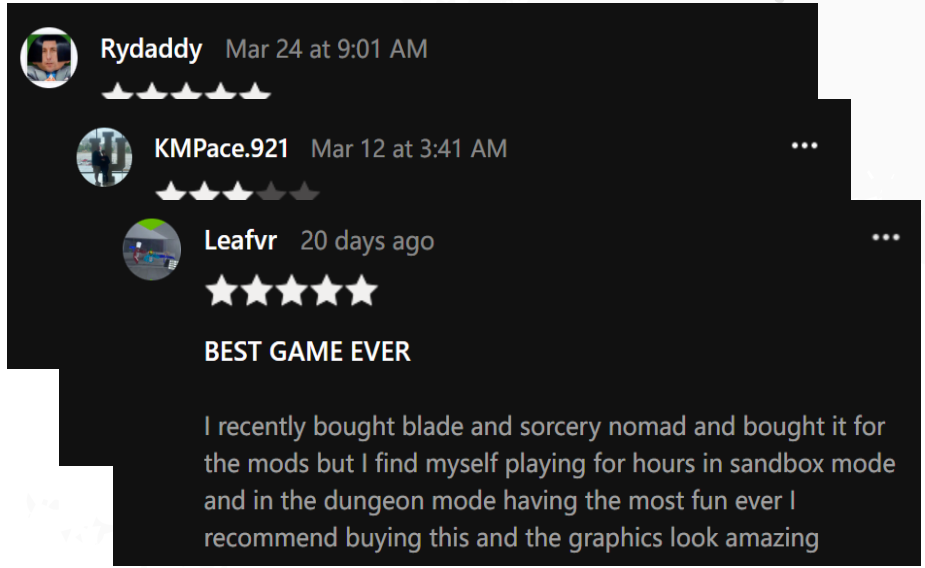
Late Fusion

- 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

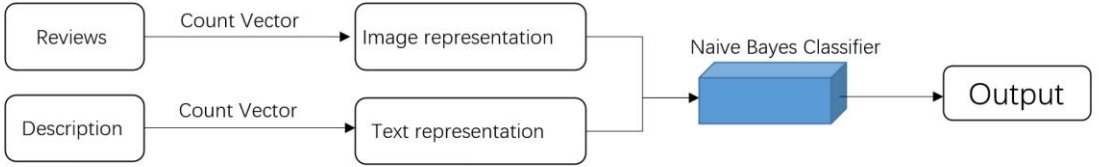


App descriptions

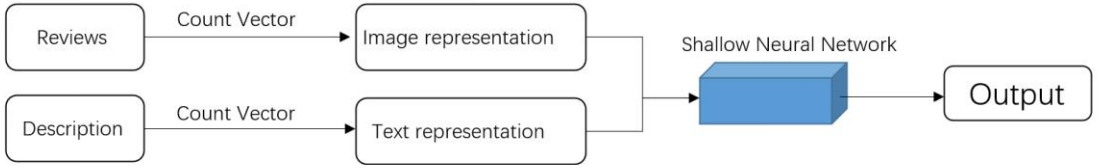


User reviews

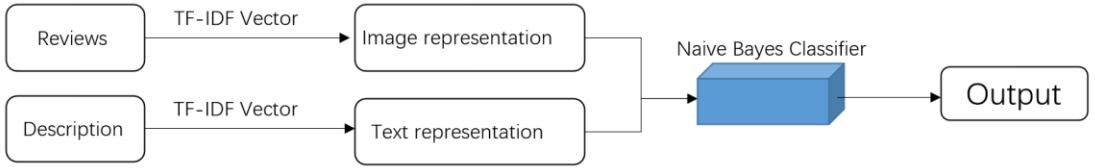
Text-Text Learning - early fusion



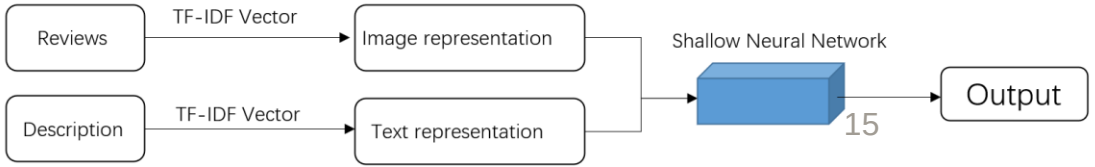
Description-Review Bayes Classifier (Count)



Description-Review Shallow Neural Network(Count)

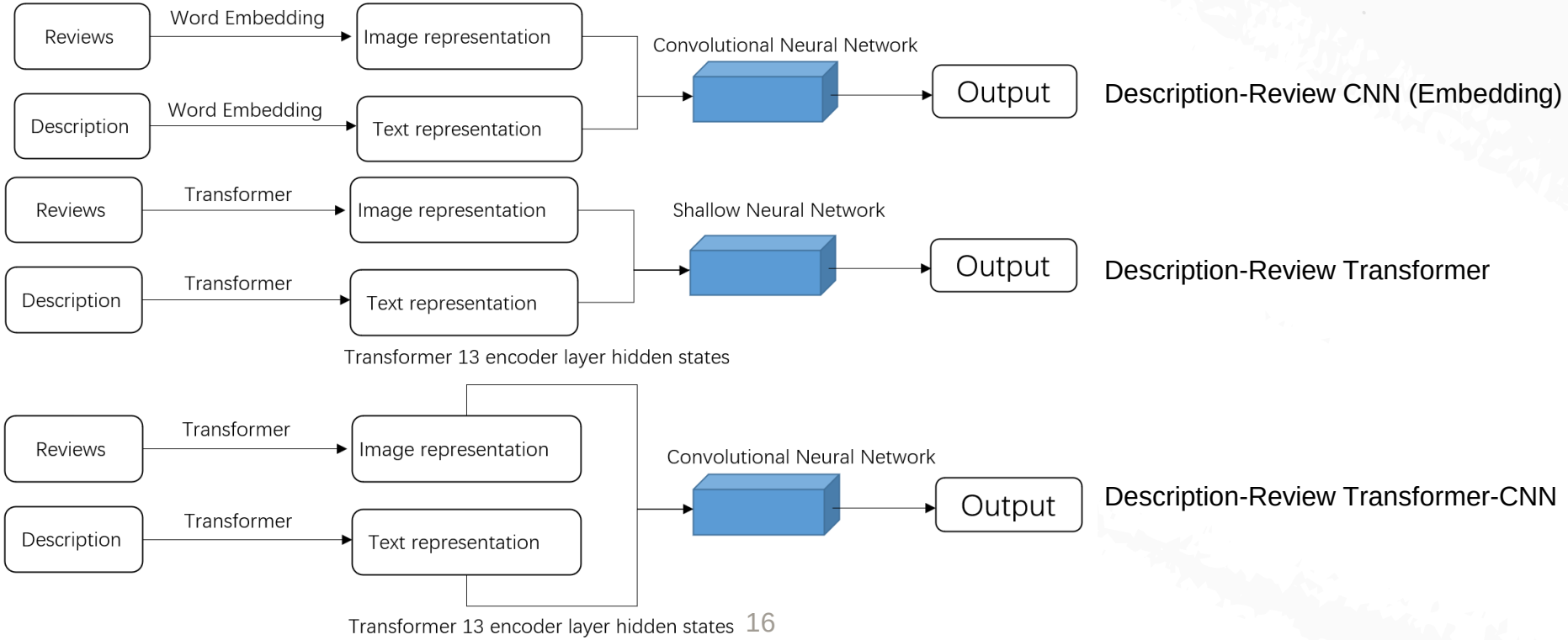


Description-Review Bayes Classifier (TFIDF)



Description-Review Shallow Neural Network(TFIDF)

Text-Text Learning – early fusion



Text-text learning result

Model	Accuracy
Description Bayes Classifier (Count Vector)	0.7207
Description Bayes Classifier (Count Vector Oversampled)	0.6723
Description Bayes Classifier (TF-IDF Vector)	0.6301
Description Bayes Classifier (TF-IDF Vector Oversampled)	0.5963

Model	Accuracy
Description-Review Bayes Classifier (Count Vector)	0.7103
Description-Review Bayes Classifier (Count Vector Oversampled)	0.6981
Description-Review Bayes Classifier (TF-IDF Vector)	0.6502
Description-Review Bayes Classifier (TF-IDF Vector Oversampled)	0.6233

Description-Review Bayes Classifier

Model	Accuracy
Description-Review SNN (Count Vector)	0.6501
Description-Review SNN (Count Vector Oversampled)	0.6788
Description-Review SNN (TF-IDF Vector)	0.6455
Description-Review SNN (TF-IDF Vector Oversampled)	0.6333

Description-Review Shallow Neural Network

Model	Accuracy
Description-Shallow Neural Network (Count Vector)	0.6483
Description-Shallow Neural Network (Count Vector Oversampled)	0.6703
Description-Shallow Neural Network (TF-IDF Vector)	0.6192
Description-Shallow Neural Network (TF-IDF Vector Oversampled)	0.6321

Text-text learning result

Model	Accuracy
Description-Convolutional Neural Network	0.7074
Description-Convolutional Neural Network (Oversampled)	0.7665

Model	Accuracy
Description-Review CNN (Word Embedding)	0.7221
Description-Review CNN (Word Embedding Oversampled)	0.7443

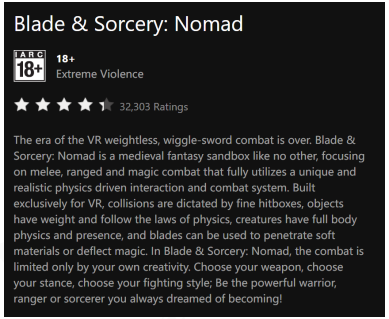
Description-Review Convolutional Neural Network

Model	Accuracy
Description-Review Transformer	0.8401
Description-Review Transformer (Oversampled)	0.8356
Description-Review Transformer-CNN	0.8397
Description-Review Transformer-CNN (Oversampled)	0.8434

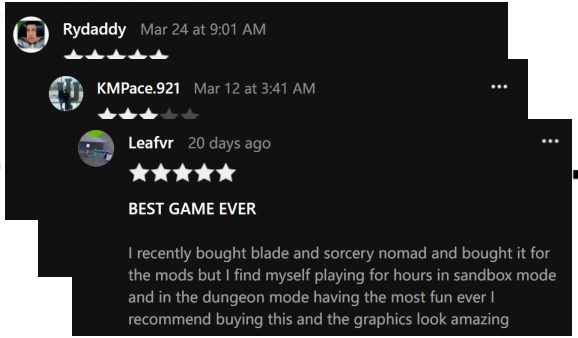
Description-Review Transformers

Model	Accuracy
Description-Transformer	0.8316
Description-Transformer (Oversampled)	0.8325
Description-Transformer-CNN	0.8432
Description-Transformer-CNN (Oversampled)	0.8398

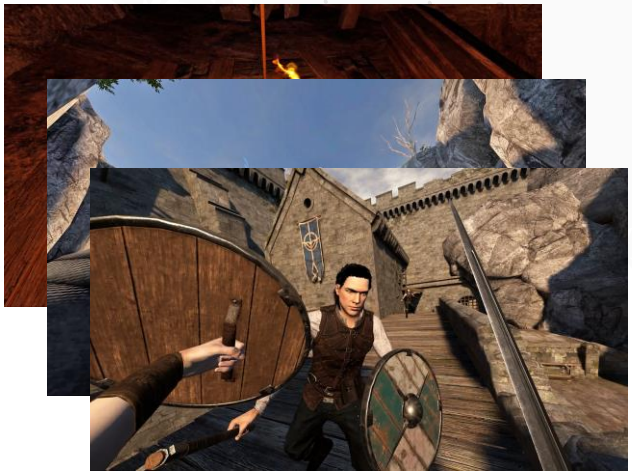
Text-Text-Image



App descriptions

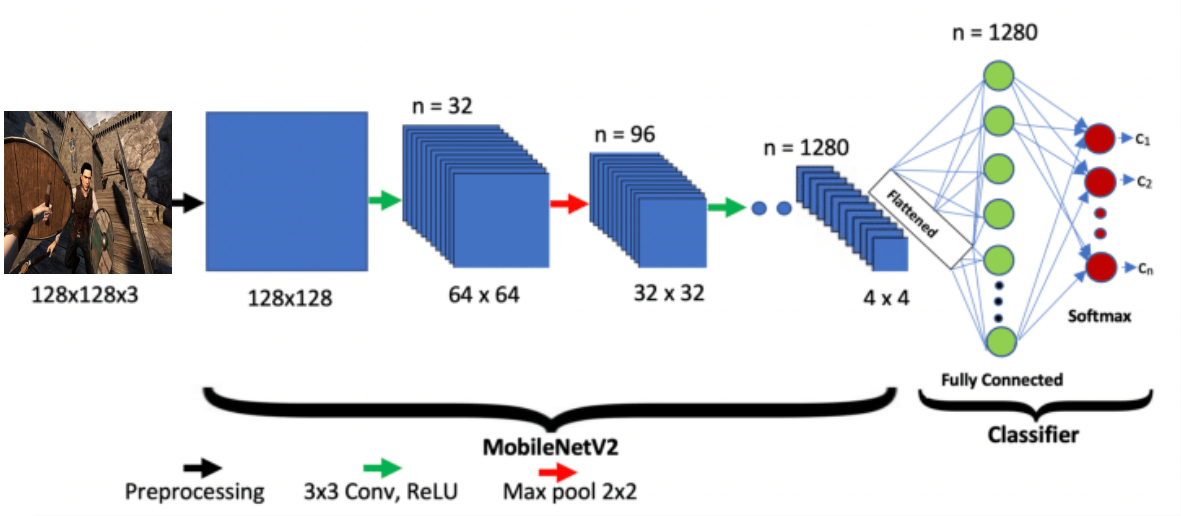


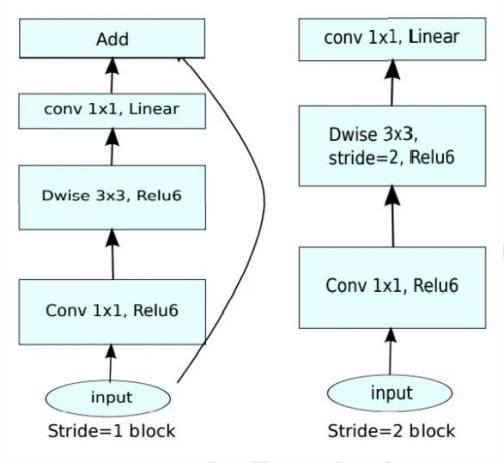
User reviews



Images

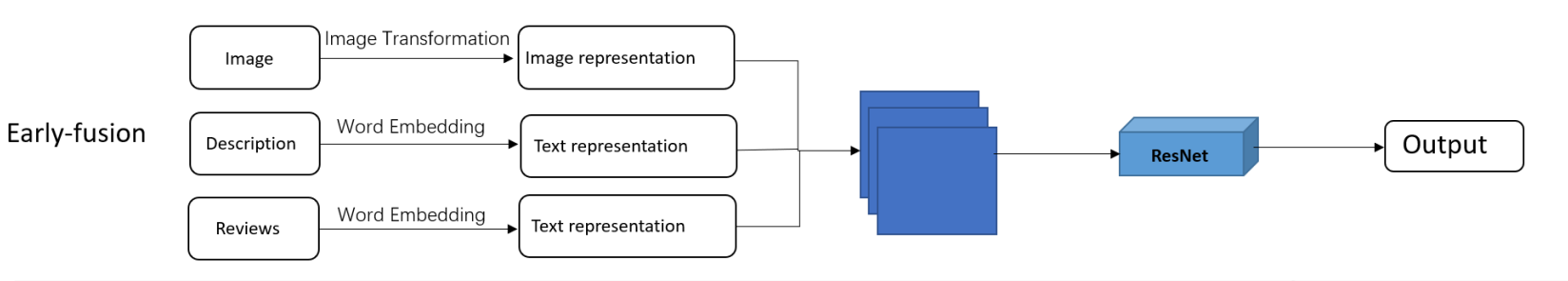
Image Encoder - MobileNetV2



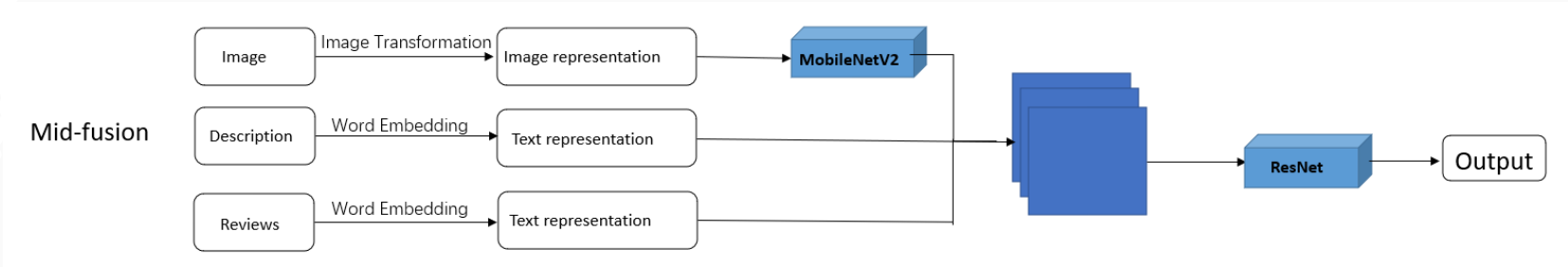


Convolutional neural network architecture built on an inverted residual structure

Text-Text-Image

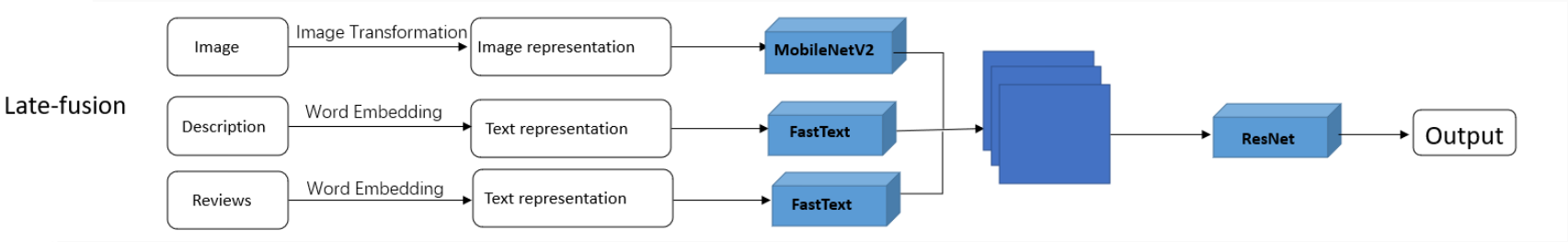


Description-Review-Image (Early-fusion)

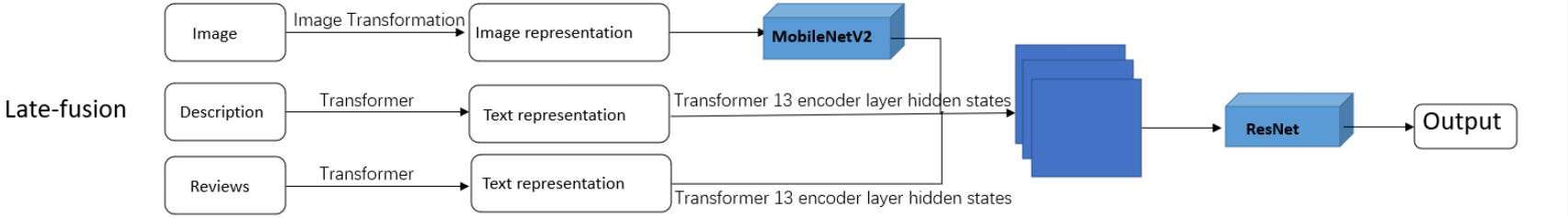


Description-Review-Image MobileNetV2 (Mid-fusion)

Text-Text-Image



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



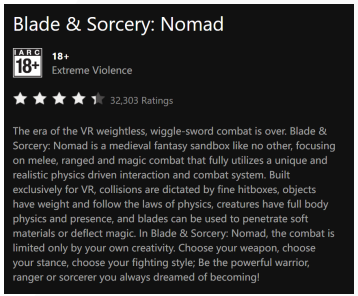
Description-Review-Image MobileNetV2 Transformer (Late-fusion)

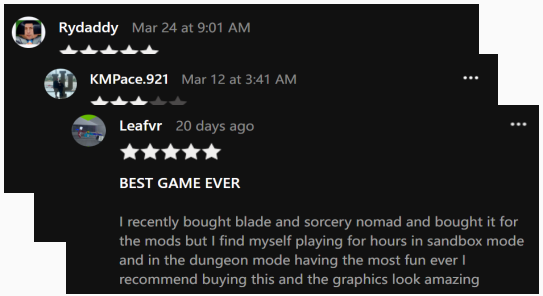
Text-Text-Image Result

Model	Accuracy
Description-Review-Image (Early-Fusion)	0.7322
Description-Review-Image MobiletNetV2 (Mid-Fusion)	0.7613
Description-Review-Image MobiletNetV2 FastText (Late-Fusion)	0.7829
Description-Review-Image MobiletNetV2 Transformer (Late-Fusion)	0.8475

Description-Review-Image Learning

Text-Text-Video



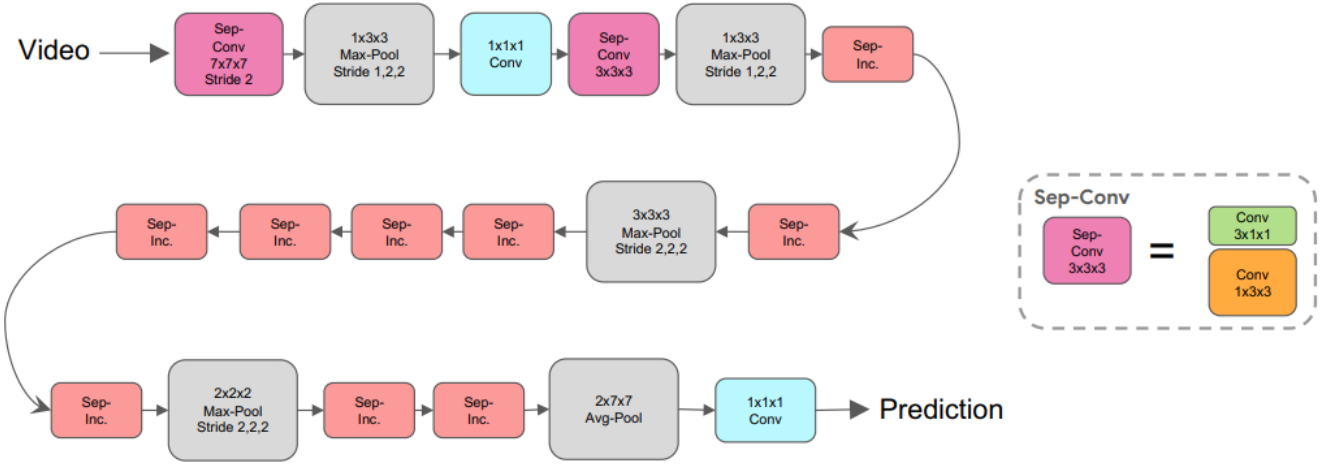


App descriptions

User reviews

App video

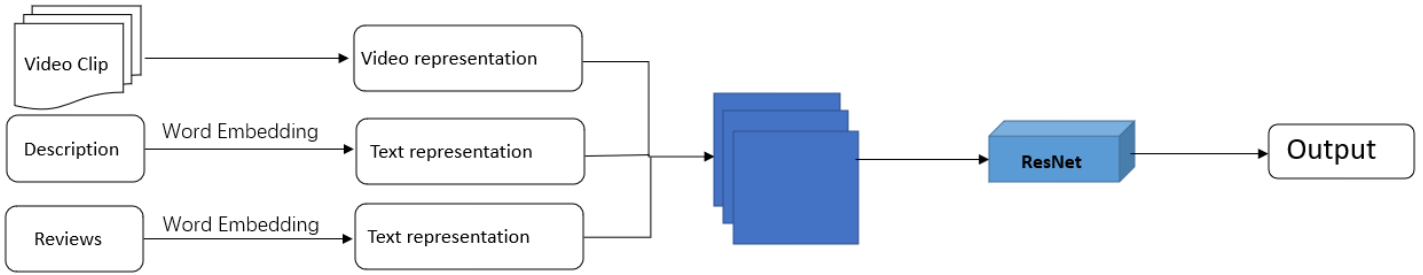
Video Encoder
S3D (Separable 3D convolutions)



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

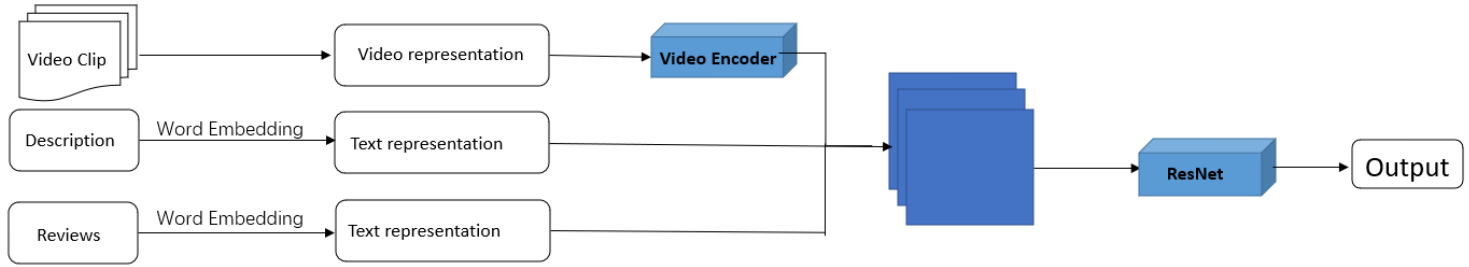
Text-Text-Video

Early-fusion



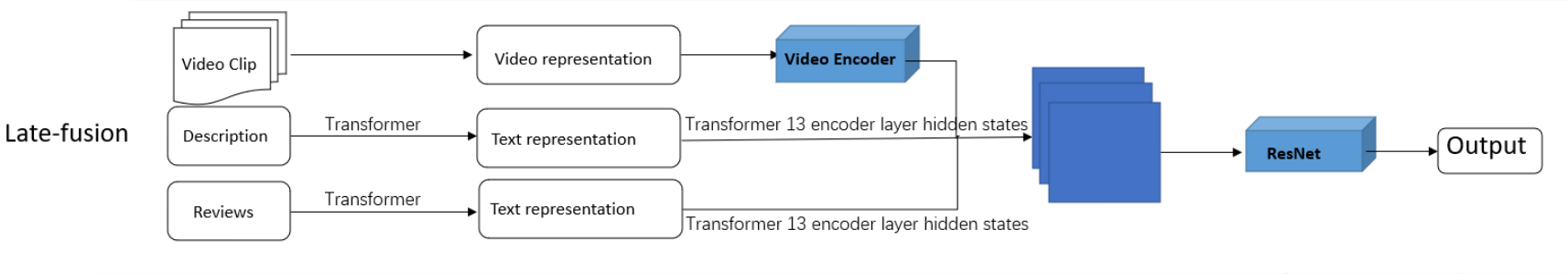
Description-Review-Video (Early-fusion)

Mid-fusion



Description-Review-Video (Mid-fusion)

Text-Text-Video



Description-Review-Video (Late-fusion)

Text-Text-Video learning result

Model	Accuracy
Description-Review-Video Cross Encoder (Early-Fusion)	0.8398
Description-Review-Video Cross Encoder (Mid-Fusion)	0.8492
Description-Review-Video Transformer (Late-Fusion)	**0.8514**

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