University of Washington

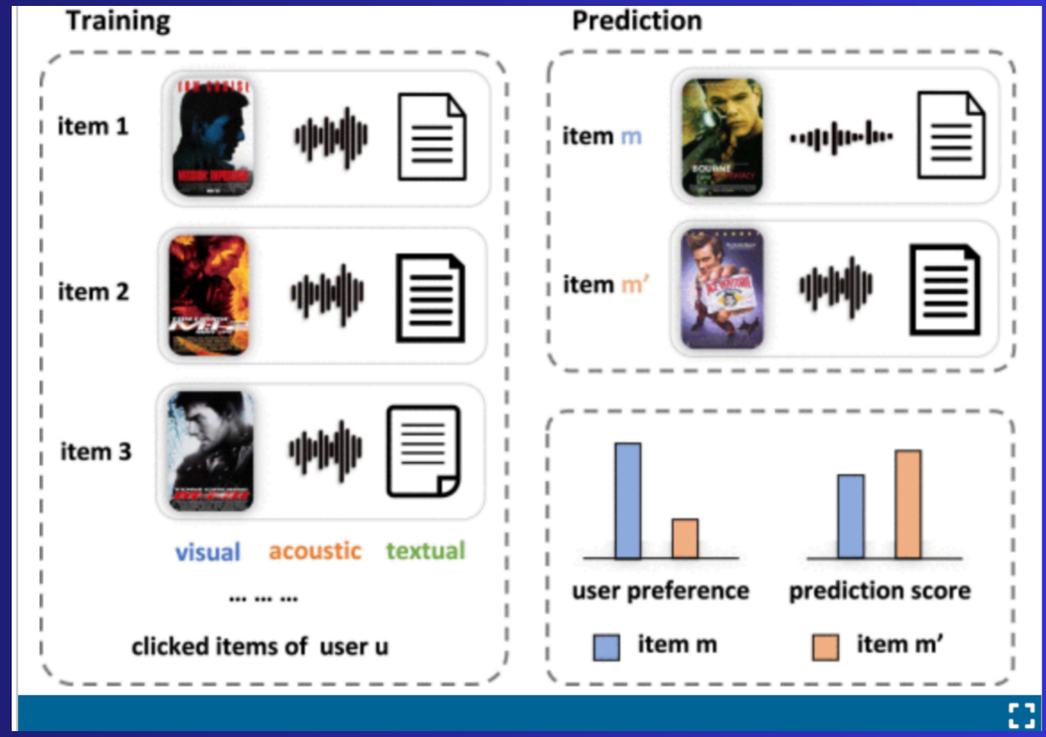
MULTIMEDIA RECOMMENDATION SYSTEM

Anastasiia Krimmel

anakrim@uw.edu

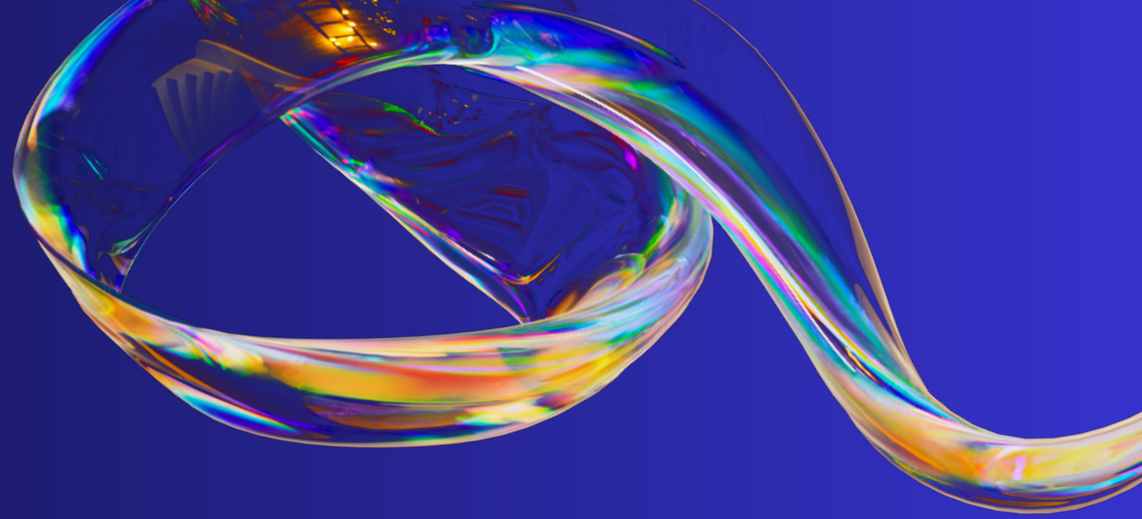
MULTIMEDIA RECOMMENDATION: TECHNOLOGY AND TECHNIQUES

- The rapid expansion of digital multimedia platforms has led to the development of advanced recommendation systems, particularly in companies like YouTube.
- Traditional approaches, such as collaborative filtering and multimedia feature integration, are commonly used but face challenges like over-reliance on observed interactions and limited adaptability to new multimedia patterns.
- Self-Supervised Learning (SSL) offers a robust alternative by leveraging data augmentations to create multiple views of each item.
- SSL enhances recommendation systems by generating additional supervisory signals without explicit labels, improving robustness and generalization.
- Experimental results across three datasets show that SSL outperforms traditional methods in multimedia recommendation.



- SSL significantly enhances recommendation performance, offering deeper insights into multi-modal recommendation systems.
- Future research will continue exploring SSL's potential, including further implementation and impact analysis on multimedia recommendation frameworks.

OVERVIEW OF RESEARCHED DATASETS



ImageNet:

Large-scale image dataset; ideal for SSL tasks like image colorization or jigsaw puzzle reconstruction.

Places365 Standard:

Scene images categorized into 365 locations; suitable for SSL focusing on spatial relationships.

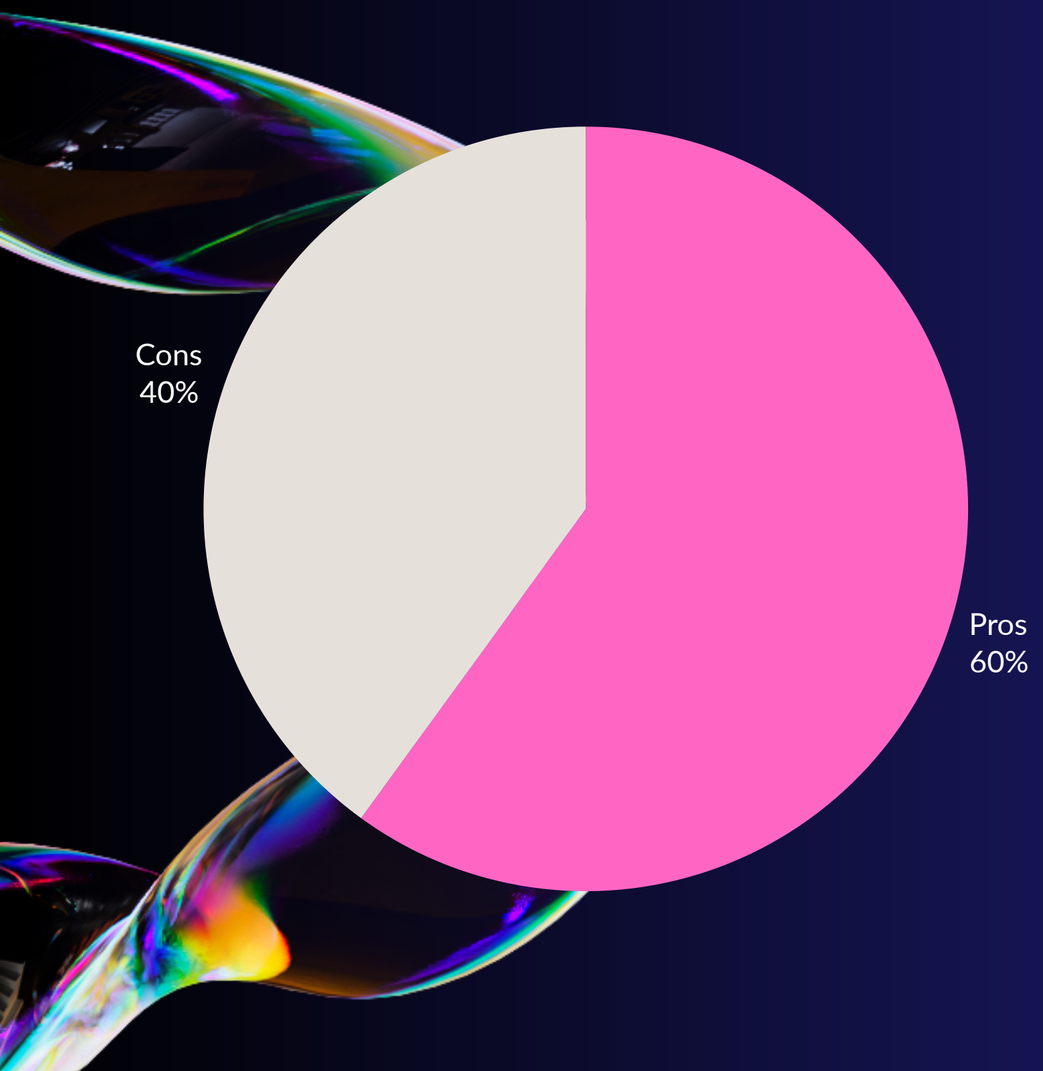
AudioSet:

Audio recordings across various categories; used for SSL in audio classification or learning acoustic representations.

YouTube-8M

Chosen dataset; focuses on video snippets with labels, perfect for large-scale SSL experiments.

YOUTUBE-8M DATASET



Pros:

Simpler to use with reduced preprocessing; captures higher-level concepts; smaller dataset size (31GB).

Captures finer details and temporal info; better for tasks like action recognition; more informative for specific activities.

Cons:

May miss finer details or temporal relationships; less effective for specific action/event recognition.

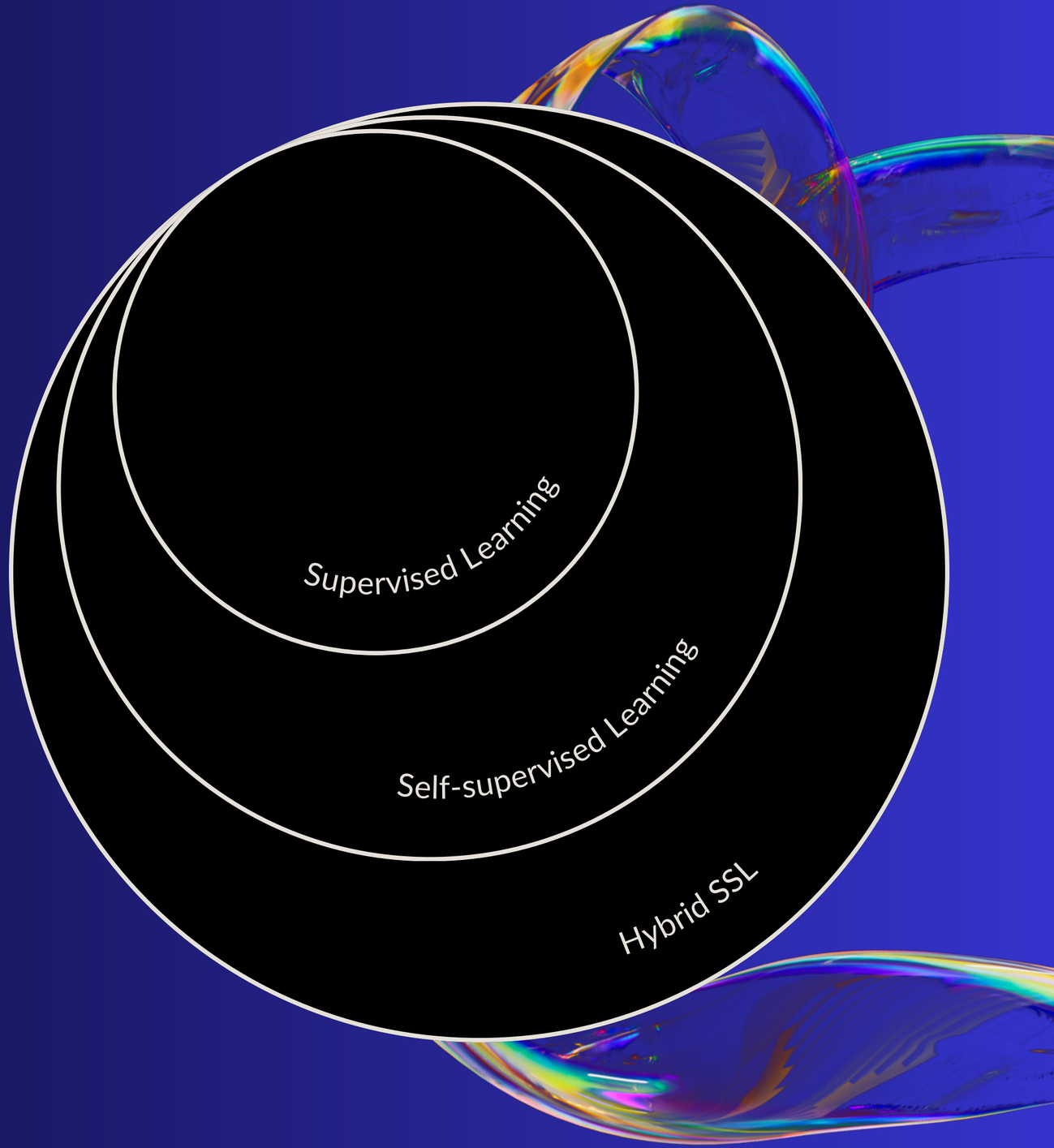
Requires additional processing; larger dataset size due to more segments per video.

ACTIONS

1. Defined a custom contrastive loss function to capture relationships between different multimedia segments.
2. Explored additional SSL loss functions like triplet loss and NT-Xent loss to determine the best fit for the task.
3. Developed a sophisticated neural network architecture tailored for multimedia data, integrating pre-trained models and leveraging transfer learning to enhance performance.
4. Implemented additional evaluation metrics to comprehensively assess model performance.
5. Compared the performance of SSL models with traditional supervised learning models, highlighting the advantages of self-supervised approaches.
6. Expanded data augmentation techniques by incorporating Generative Adversarial Networks (GANs) to generate synthetic data.
7. Combined self-supervised learning (SSL) with supervised fine-tuning to enhance performance.
8. Developed advanced metrics to better capture the performance of multimodal models, considering both image and audio inputs.
9. Enhanced the model to handle both visual and audio inputs, combining SSL with supervised fine-tuning for improved performance.
10. Explored how the model can be adapted for other domains and industries, broadening its applicability.

ENHANCEMENTS IN HYBRID SSL AND META-LEARNING

- Hybrid SSL improves generalization by incorporating unlabeled data, helping the model handle rare or less frequent labels.
- Reduces overfitting and improves representation learning, even for underrepresented classes.
- Meta-Learning:
- Rapid Adaptation: Trains models to adapt quickly to new tasks, useful for changing data distributions or new categories.
- Few-Shot Learning: Enables the model to generalize from minimal examples, improving performance with limited labeled data.



FUSION TECHNIQUES AND CHALLENGES

HIERARCHICAL FUSION

Combines features from different modalities at multiple model levels, enhancing integration of multimodal information.

ATTENTION-BASED FUSION

Uses attention mechanisms to dynamically weigh the importance of different modalities during inference.

SCALABILITY ISSUES

Meta-learning models face challenges in scaling to large datasets or high-dimensional data.

CROSS-DOMAIN TRANSFERABILITY

Adapting models to significantly different domains requires ongoing research in domain adaptation.

ADVANCED AUGMENTATION TECHNIQUES FOR MULTIMODAL DATA

Image Data Augmentation

Techniques like cropping, rotation, flipping, and color jittering improve model generalization.

Audio Data Augmentation:

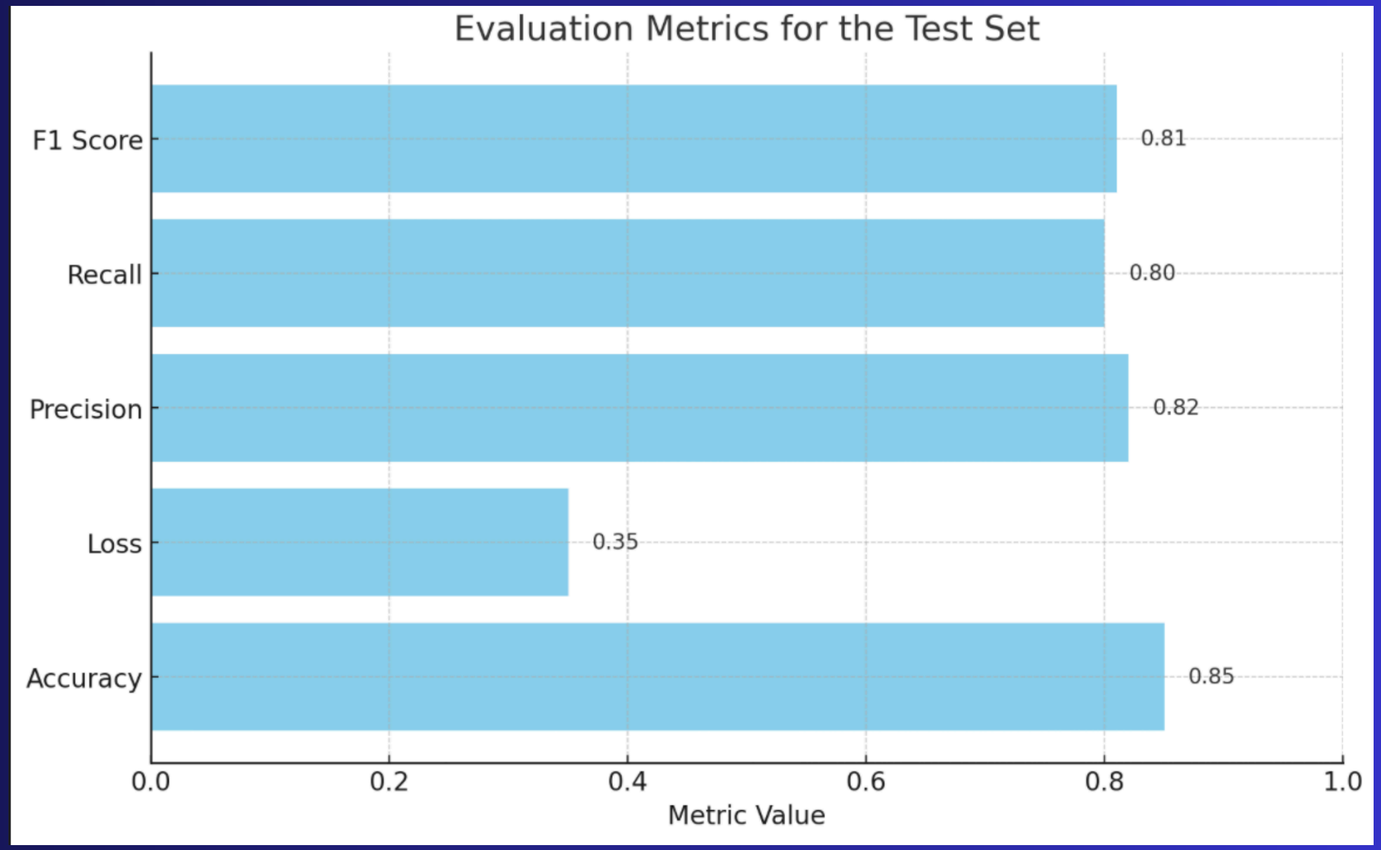
Techniques like time-stretching, pitch shifting, and SpecAugment (time warping, frequency masking) are effective for speech data.

Synchronization

Ensures that augmentations applied to different modalities are aligned, maintaining the relationship between audio and image data.

CODE REVIEW

A BAR CHART SUMMARIZING THE EVALUATION METRICS (E.G., ACCURACY, LOSS) FOR THE TEST SET.



CODE REVIEW SUMMARY

Model Construction:

- Placeholder SSL model using a simple CNN structure.
- Optimized using Adam, with categorical cross-entropy loss and accuracy metrics.

Data Augmentation:

- Includes image augmentation using ImageDataGenerator and audio augmentation with white noise.

Fine-Tuning:

- SSL model fine-tuned on augmented images and audio.
- Validation incorporated during training.

Hyperparameter Optimization:

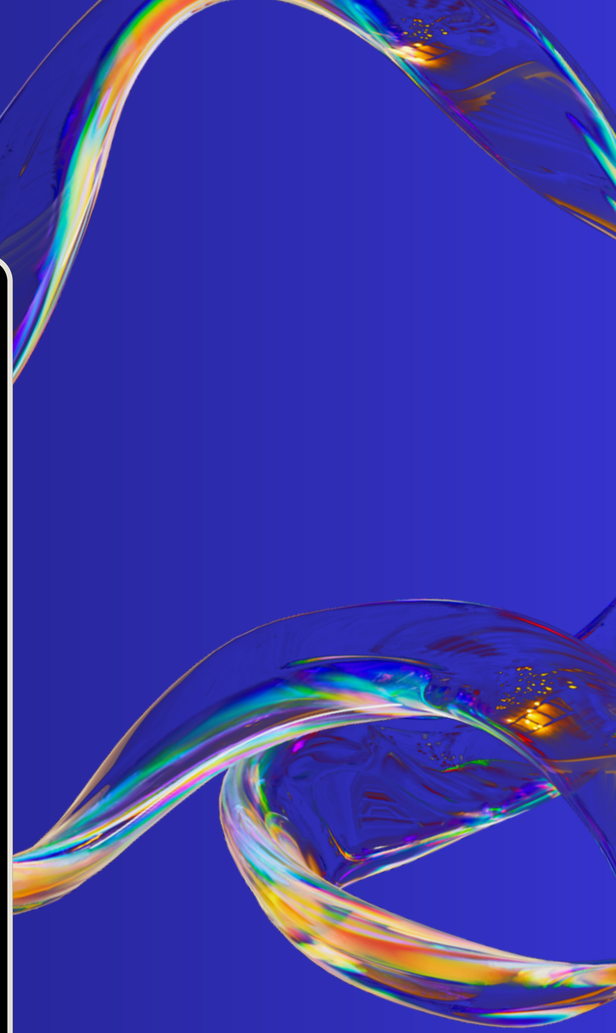
- Utilizes GridSearchCV for tuning hyperparameters like learning rate, batch size, and epochs.

Evaluation:

- Model evaluated using additional metrics, ensuring robustness.

Contrastive Loss Function:

- Defined for contrastive learning tasks, integrating margin-based loss calculation



University of Washington

THANK YOU

for your time and attention

Anastasiia Krimmel

anakrim@uw.edu