УДК 004.89

Разработка мультимодальных признаков и интерфейсного решения для модели предсказания вовлечённости видеоконтента.

Коршунов И.К.
Научный руководитель – канд. техн. наук, доцент Ефимова В.А.
Университет ИТМО
I.Korshunov2004@yandex.ru

Введение

На практике создатели видеоконтента оценивают успешность роликов по тому, насколько долго зритель смотрит видео, однако без предиктивной модели оценить удержание до публикации невозможно. Вовлечённость определяется не только смысловой составляющей, но и совокупностью мультимодальных факторов: аудиодорожкой и речью, визуальным рядом, композицией кадра и другими компонентами. Извлечение таких характеристик и анализ вклада каждой из них позволяют повысить точность прогнозирования удержания и сделать модель более интерпретируемой. Поэтому актуальна разработка прикладного инструмента: метода, который обеспечивает формирование признаков, получение прогноза и интерпретацию результатов на основе методов машинного обучения.

Основная часть

В рамках работы рассматривается набор видеороликов с известными значениями метрики удержания аудитории. Предлагается построить пайплайн, который по входному видеоролику автоматически извлекает мультимодальные признаки и использует их для обучения модели прогнозирования удержания. Предполагается использовать рекуррентную модель для получения представлений временной динамики ролика. Основной набор признаков включает три группы:

1) визуальные признаки (характеристики динамики и композиции кадра, смена сцен и др.) [1];

2) аудиопризнаки (энергетические характеристики, темп, паузы и др.) [2];

3) текстовые и речевые признаки (транскрипт, характеристики речи).

Для корректного обучения и сопоставления вкладов предусмотрены процедуры нормализации и масштабирования признаков. Далее планируется экспериментальный анализ влияния отдельных признаков и их комбинаций на качество прогнозирования: сравнение моделей, обученных на каждой модальности отдельно, на попарных комбинациях и на полном мультимодальном наборе.

Качество наборов признаков оценивается по значениям метрик на моделях, обученных на этих признаках. Для сравнения используются следующие подходы: модель на признаках одной модальности, ансамблевые регрессионные методы, модель на предобученных эмбеддингах. Точность прогнозирования измеряется метриками MAE и RMSE.

Практическая часть включает разработку пользовательского интерфейса, обеспечивающего загрузку ролика, запуск вычисления признаков, получение прогноза удержания и визуализацию результатов, включая сравнительную аналитику по группам признаков. Это повышает прикладную применимость подхода для создателей видеоконтента.

Выводы

По итогам работы будет реализован модуль извлечения мультимодальных характеристик видеороликов и проведены эксперименты по оценке качества прогнозирования, включая анализ влияния отдельных групп признаков и их сочетаний. Результатом станет интерфейс, который помогает получать прогноз вовлекающего потенциала видео и наглядно интерпретировать факторы, связанные с удержанием аудитории. Предлагаемое решение может использоваться как вспомогательный инструмент для доработки видеоконтента.

Литература

1. YSDA course in Computer Vision Processing [Электронный ресурс]. – Режим доступа: https://code.mipt.ru/courses-public/cv/public, свободный. Яз. рус. (дата обращения 09.02.2026).

2. YSDA course in Speech Processing [Электронный ресурс]. – Режим доступа: https://github.com/yandexdataschool/speech_course, свободный. Яз. рус. (дата обращения 20.12.2025).