MIT MBAn 2020
CAPSTONE PROJECT

ADVISORS: Prof. Karen Zheng, Michelle Li
LOCATION: Copenhagen, Denmark
GROUPM: Kristjan Brøderskift, Andreas Christensen, Danni Dromi, Martin Olesen, Fredrik Skeppstedt, Mathias Stuhaug

PEGGY OLSON 2.0
CREATIVE AI

Alexandru Socolov
Girish Kishen Govindarajan

groupm
MIT MANAGEMENT BUSINESS ANALYTICS

PROBLEM STATEMENT

WHY?
Video production is expensive. Creatives spend hundreds of hours researching ads of competitors & related products.

WHAT?
The idea is to provide inspiration to creatives in the early stages of video production. This will also free up a lot of time for other tasks.

HOW?
A data-driven ad search platform where creatives can filter on features of ads and clearly communicate ideas to clients.

Peggy 2.0 Features

CUT DETECTION
Creatives would like to mix and match shots from different videos on to their storyboard. Hence, we have to come up with a way to split videos based on camera cuts.

AD PURPOSE
Ads are branding or tactical in nature. How can we leverage the call-to-action text, brands and logos to quantify the purpose of an ad?

MOOD DETECTION
The same visual elements and objects can appear in multiple videos, but in different contexts. Can we utilize the background music to infer the mood or tone of the ad?

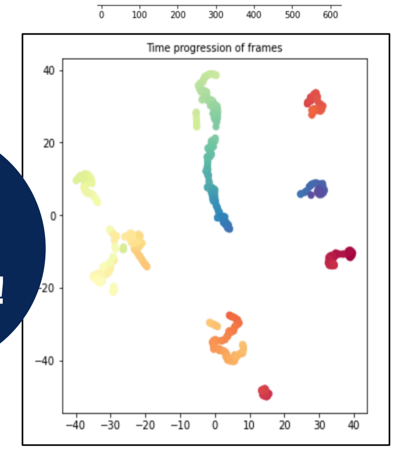
SIMILARITY SEARCH
When a user searches for objects that are not present in our videos, we would still like to show them related cuts. We need a versatile algorithm that can search for similar ideas.

CUT DETECTION

DATA
~500 videos with annotated cuts

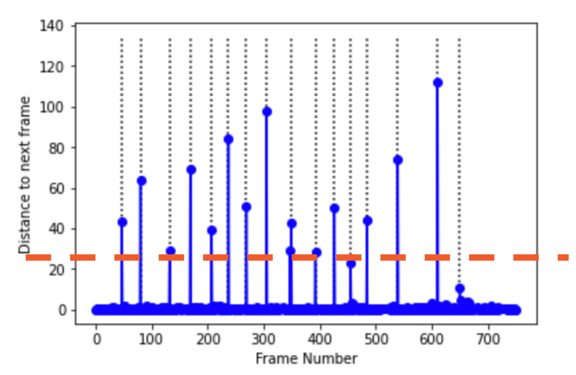
METHOD
video frames → inception v3 imagenet network → t-SNE 2D representation of video frames → $\Delta = || X_{current} - X_{previous} ||$ If Δ > threshold, then cut!

frames



shots cluster together!

2D t-SNE embedding of penultimate layer activations of Inception v3 output for the frames of a video



threshold above which frame is a cut

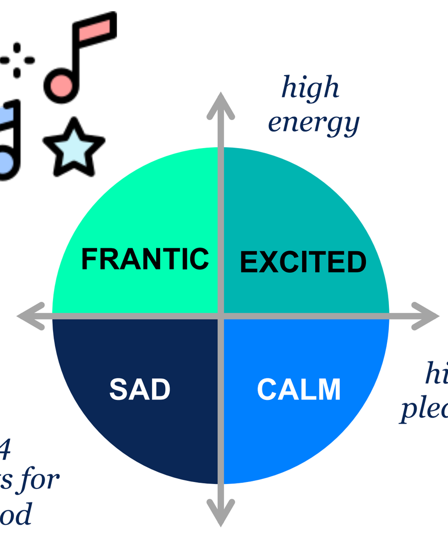
WHERE TO PLACE THE THRESHOLD?
We search for an optimal threshold that maximizes the F2 score across 300+ annotated videos with 6000+ cuts!

RESULTS
We benchmark the performance of our model against the Google Video Intelligence Shot Change Detector API on 200 videos.

METRIC	WE WIN	WE TIE	WE LOSE
PRECISION	70%	0%	30%
RECALL	85%	11%	4%
F1	81%	0%	19%
F2	89%	0%	11%

MOOD CLASSIFICATION

DATA
225 audio tracks per mood
30 sec each (chop into 5-sec clips)

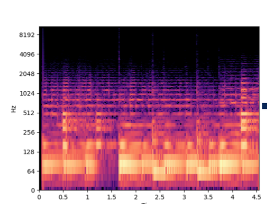


Russell's 4 quadrants for audio mood

FEATURE EXTRACTION
Raw audio → Background Music → MIR toolbox → **43 low-level features**: rhythmic, spectral, tonal, dynamic

filter out voice-over & singing

Select 5 seconds around each cut



Spectrogram image of audio → Pre-trained network (MusicNN) → **50 high-level features**

MODEL
XGBOOST → mood prediction

RESULTS

MODEL	AUC	ACCURACY
Manual Feature Extraction (Feature Packages)	0.61	36%
MusicNN Features v1	0.70	43%
MusicNN Features v2	0.73	29%
MusicNN Features v3	0.76	45%
MusicNN Features v3 + Manual Features	0.76	45%
Convolutional Neural Network	0.64	22%
MusicNN Features v4	0.85	60%
MusicNN Features v4 + Manual Features from MATLAB	0.87*	64%*

**close to state-of-the art performance. We cannot leverage articulation or lyrical features since ads have voice-over.*

AD PURPOSE CLASSIFICATION

DATA
370 ads scored along tactical-branding axis.
Label types:
- 2-class ("Branding" & "Tactical")
- Ordinal Score 3-level
- Ordinal Score 5-level

FEATURE EXTRACTION
video frames → Google Cloud Vision API →
Call-to-action text
Logos
Bounding box
- area
- position

Total: 25+ features

METHOD

Ordinal Classification
Transform an ordered k-class problem into k-1 binary classification problems.

video	X (features)	Y (score)
1	...	1
2	...	3
:	:	:
n	...	2

original data with 3 ordered classes

Train
$f_1(x)$ and $f_2(x)$ are the pdfs for the 2 binary classifiers.

$f_1(x)$ = P(score >1) = P(score = 2 or 3)

video	x	Y (score > 1)
1	...	0
2	...	1
:	:	:
n	...	1

$f_2(x)$ = P(score > 2) = P(score = 3)

video	x	Y (score > 2)
1	...	0
2	...	1
:	:	:
n	...	0

Predict
let's say for video j:
$f_1(x_j) = 0.8$
$f_2(x_j) = 0.6$

$P(score = 1) = 1 - f_1(x_j) = 0.2$
$P(score = 2) = f_2(x_j) - f_1(x_j) = 0.2$
$\mathbf{P(score = 3) = f_2(x_j) = 0.6}$

Since P(score = 3) is maximum, we predict a score of 3 for video j

RESULTS

MODEL	LABELING	METRIC	VALUE
Logistic Classifier	2-class	Accuracy AUC	80% 86%
Ordinal Logistic Classifier	1-3 Scoring	MAE	0.45
Ordinal Logistic Classifier	1-5 Scoring	MAE	0.67
Ordinal XGBoost Classifier	1-5 Scoring	MAE	0.74

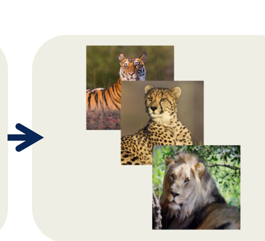
SIMILARITY SEARCH

DATA
5,600+ labels detected in videos and 20,000 most common English words

MOTIVATION
End product is a user interface where the user inputs keywords and we show similar videos/cuts in that context e.g. 'tiger' should return 'lion' and 'cheetah' too.

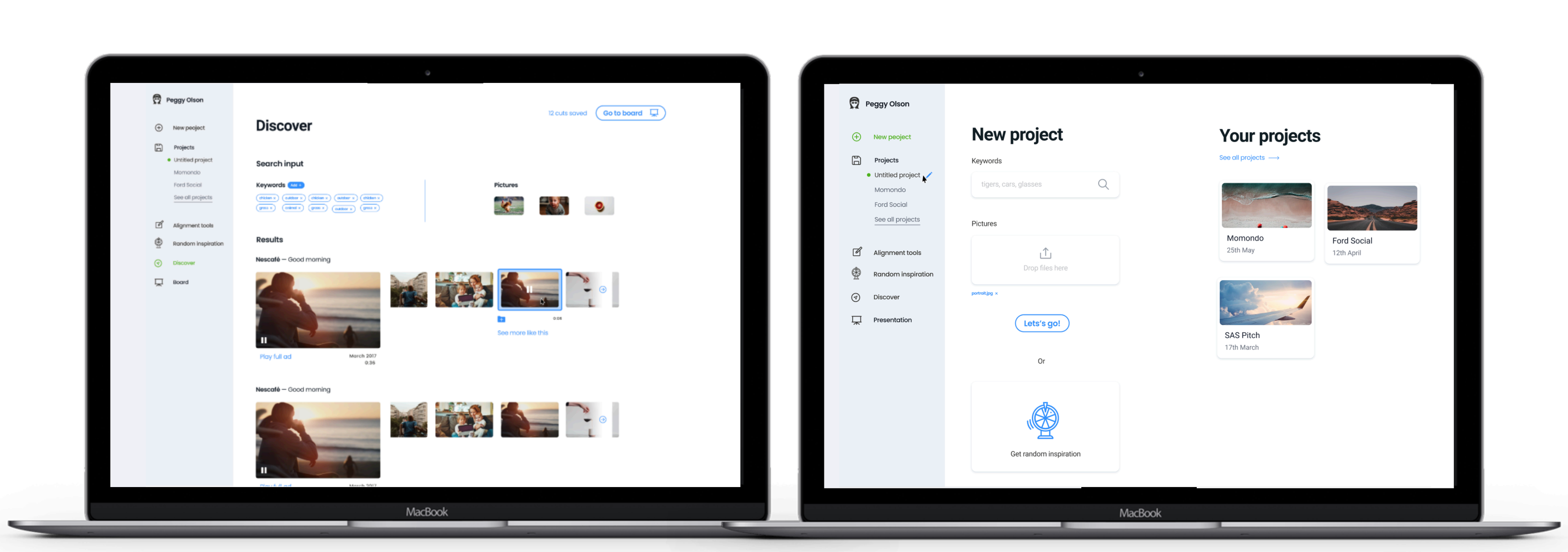
EMBEDDING LABELS
Video frames → Google Cloud Vision API → Raw label (tiger) → pass through an NLP model to get an embedding → 1 x 300 vector representing the label (tiger)

ONLINE USER SEARCH
User keywords (tiger, nature) → 1 x 300 vectors representing the keywords (tiger, nature) → compare with vectors in 7,000 videos using cosine similarity (forest, tree, lion) → return sorted cuts by similarity



TAKEAWAYS
- Given user's input, we return shots that have similar labels detected within them.
- A huge advantage of this approach is that, if no cuts contain a keyword, we still return something relevant. For example, if we do not have the label "cabriolet" anywhere in the cuts, we would still return cuts with "car", "convertible" and "coupe".

DASHBOARD



IMPACT

Scalability
Language versatility allows the tool to be marketed to all global clients of GroupM.

Speed
Online search happens in seconds, and the user is met with a set of similar scenes, their moods and purpose.

Efficiency
As per GroupM's estimate, this will save the creatives 375+ hours of work (Nordic region).

NEXT STEPS
- **Front-end**: Right now, Peggy accepts only keyword searches, taking images or sketches as input will enhance functionality.
- **Modeling:** Voiceover or narration can be leveraged to infer mood and ad purpose.
- **Maintenance**: GroupM can benefit from recording the details of all new ads produced in terms of the scenes, music, etc.