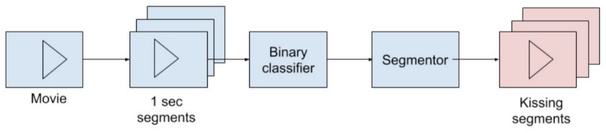
# Detecting Kissing Scenes in a Database of Hollywood Films

Amir Ziai (amirziai@stanford.edu)

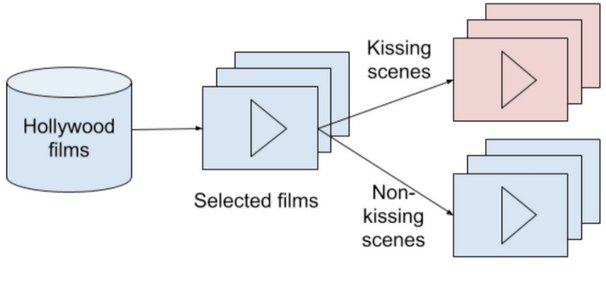
## Problem

Detecting scene types in a movie can be very useful for application such as video editing, ratings assignment, and personalization. We propose a system for **detecting kissing** scenes in a movie. This system consists of two components. The first component is a **binary classifier** that predicts a binary label (i.e. kissing or not) given features exctracted from both the **still frames** and **audio waves** of a one-second segment. The second component aggregates the binary labels for contiguous non-overlapping segments into a set of kissing scenes. We experimented with a variety of 2D and 3D **convolutional** architectures such as ResNet, DesnseNet, and VGGish[1], and developed a highly accurate kissing detector that achieves a validation F1 score of **0.95** on a diverse database of Hollywood films ranging many genres and spanning multiple decades.




## Dataset

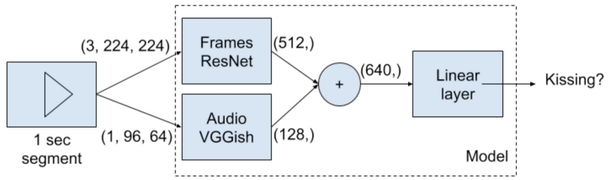
We have annotated kissing segments from a 2.3TB database of **600** Hollywood films spanning 1915 to 2016 and covering a broad range of genres and resolutions. We have produced a total of **263** kissing segments and 363 non-kissing segments ranging 10 to 120 seconds in length. The dataset is split into train, validation, and test partitions with 80%, 10% and 10% proportions respectively.




## Approach

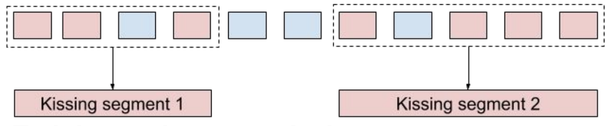
### Binary classifier

We extract a random crop of the last **still** and **audio** wave from each 1s segment. The audio is transformed into a 96x64 patch by applying a Fourier transform and integrating into mel-spaced frequency bins. Image is fed to **ResNet-18** and the audio to **VGGish**. Outputs are concatenated and passed to a linear layer for generating labels.



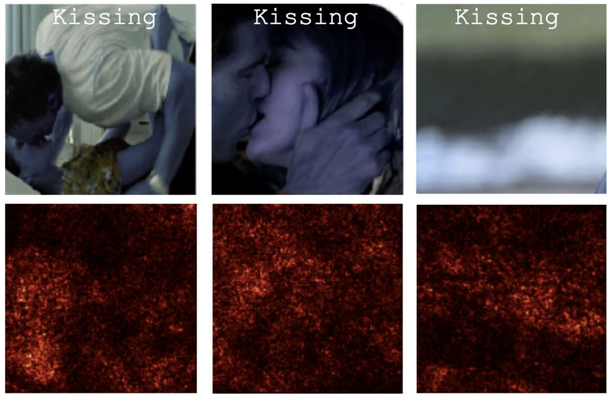

### Segmentor

The predicted labels for one-second segments are fed into an algorithm that finds long non-overlapping kissing segments that meet a minimum fraction of kissing segments.




### Training and error analysis

We trained the network for 10 epochs with cross-entropy loss using Adam with η=0.001 and achieved 0.95 validation F1. Pre-trained weights were used from **ImageNet** and **AudioSet**. **Saliency maps** suggest that the model has learnt to pay attention to faces. Random cropping at train-time has caused some confusion though.




## Experiments

### Ablation study

We experimented with excluding the audio and image feature extractor to understand their contribution.

| Trained params | VGGish included | ResNet-18 included | Validation F1 |
|---|---|---|---|
| Last layer only | Yes | Yes | 0.92 |
| Last layer only | No | Yes | 0.92 |
| Last layer only | Yes | No | 0.87 |
| All | Yes | Yes | **0.95** |
| All | No | Yes | 0.91 |
| All | Yes | No | 0.87 |

ResNet-18 plays a more important role but both extractors are important.

### Architecture and hyper-parameter search

We swept over {ResNet-18, DenseNet} architectures,learning rates {5e-4, 1e-3, 1e-2} and training either only the last layer (freezing the rest) or all parameters.

| Architecture | Trained params | Learning rate | Validation F1 |
|---|---|---|---|
| ResNet-18 | All | 0.001 | **0.95** |
| DenseNet | All | 0.001 | 0.88 |

The best configuration was using learning rate of 0.001 and training all parameters for both architectures. ResNet-18 performs substantially better.

### 3D ResNet-34

We tried a 3D ResNet-34[2] with the same configurations as before and a 16-frame temporal depth. Training time per epoch increased by 10x. We achieved F1 of **0.88**.

### Conclusions and future work

Our system can accurately detect kissing scenes in a wide range of movies. This methodology can be extended to detecting other types of scenes by annotating other scene types in the same database. We plan to explore training 3D CNNs over more iterations and with a higher temporal depth.

## References

[1] Hershey et al. CNN architectures for large-scale audio classification. IEEE international conference on acoustics, speech and signal processing, 2017

[2] Carreira et al. Quo vadis, action recognition? a new model and the kinetics dataset. IEEE Conference on Computer Vision and Pattern Recognition, 2017