Modular Memorability:

Tiered Representations for Video Memorability Prediction

Théo Dumont[1,2], Juan S. Hevia[2], Camilo L. Fosco[2]

[1] Mines Paris, PSL [2] Memorable AI

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

memorable

WED-AM-240

Modular memorability: a 1’ summary

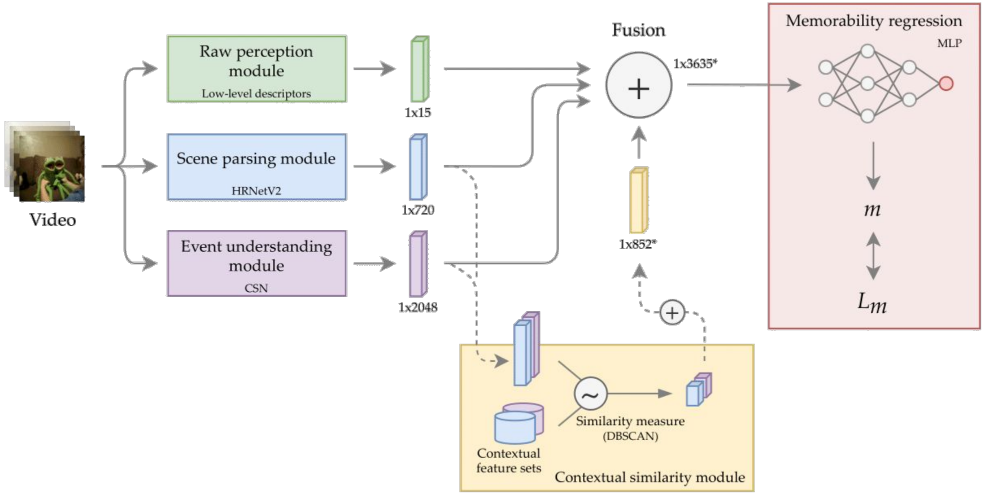
The probability one will remember this video: $m \in [0,1]$

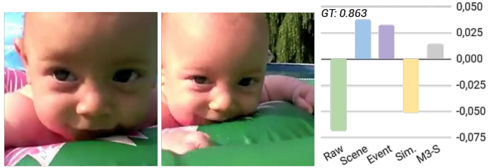
- **consistent** across people ⇒ predictable!
- highly **unintuitive** ⇒ *hard to predict…*

	Spearman RC $\rho \uparrow$	
Approach	Memento10k	VideoMem
MemNet baseline* [31]	0.485	0.425
Cohendet *et al.* (Semantic)* [13]	0.552	0.503
Cohendet *et al.* (ResNet3D)* [13]	0.574	0.508
SemanticMemNet† [40]	0.659	0.556
M3-S (ours)	**0.670**	**0.563**

Our contributions:

1. in-depth **analysis of factors of memorability** and **classification** in tiers
2. novel methodology using the classification and a measure of **distinctiveness**
3. leveraging the model’s structure to get insights on its **interpretability** and the features it learns





What is memorability?

The probability one will remember this video: $m \in [0,1]$

- **consistent** across people ⇒ predictable!
- highly **unintuitive** ⇒ *hard to predict...*

- harder than for images! a lot of additional factors (motion, emotions...)

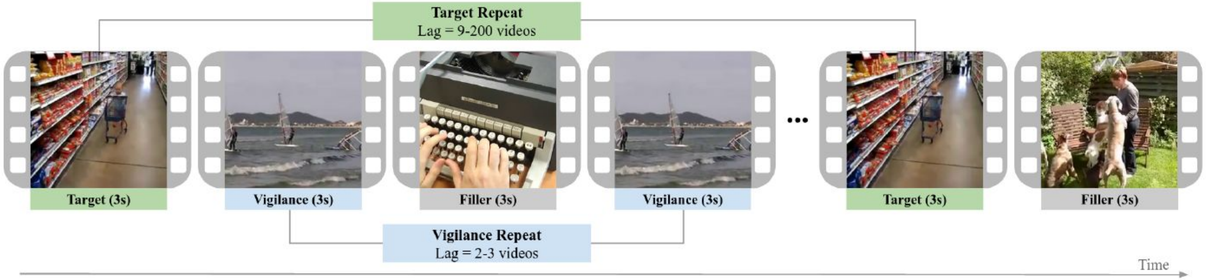
What is memorability?

The probability one will remember this video: $m \in [0,1]$

- **consistent** across people ⇒ predictable!
- highly **unintuitive** ⇒ *hard to predict…*
- harder than for images! a lot of additional factors (motion, emotions…)

How is the ground truth for video memorability obtained?

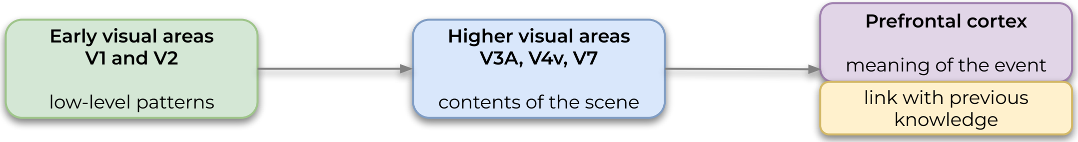
- 3-second videos shown to participants
- target video is **surrounded by filler videos**
- 2 main datasets: VideoMem and Memento10k



Limitations of existing works

Limitations:

- do not leverage underlying structure governing memorability in the brain



Limitations of existing works

Limitations:

- do not leverage underlying structure governing memorability in the brain

Early visual areas V1 and V2 — low-level patterns → **Higher visual areas V3A, V4v, V7** — contents of the scene → **Prefrontal cortex** — meaning of the event; link with previous knowledge

- videos that **stand out** are very memorable! distinctiveness hasn't been leveraged

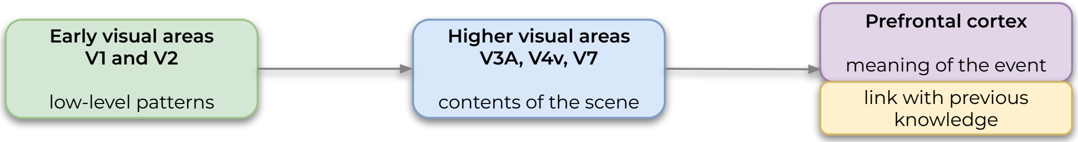
link with previous knowledge

≠

Limitations of existing works

Limitations:

- do not leverage underlying structure governing memorability in the brain

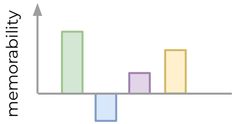


- videos that **stand out** are very memorable! distinctiveness hasn't been leveraged

link with previous knowledge

≠

- black-box models that **lack interpretability**



When is a video memorable?

- classification of factors into **tiers**

low-level
color,
brightness,
motion

mid-level
object semantics,
scene composition,
saliency

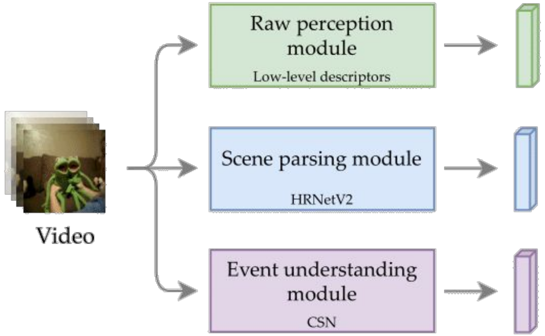
high-level
actions,
emotions

distinctiveness
standing out
from the corpus

Our modular approach

Our model separates between different tiers **by design**:

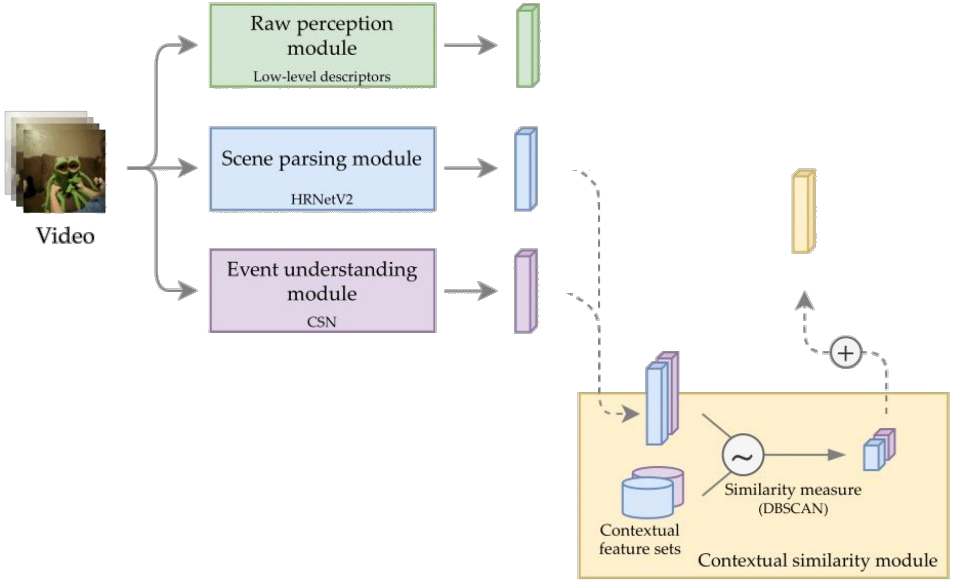
- low-level (green)
- mid-level (blue)
- high-level (purple)



Our modular approach

Our model separates between different tiers **by design**:

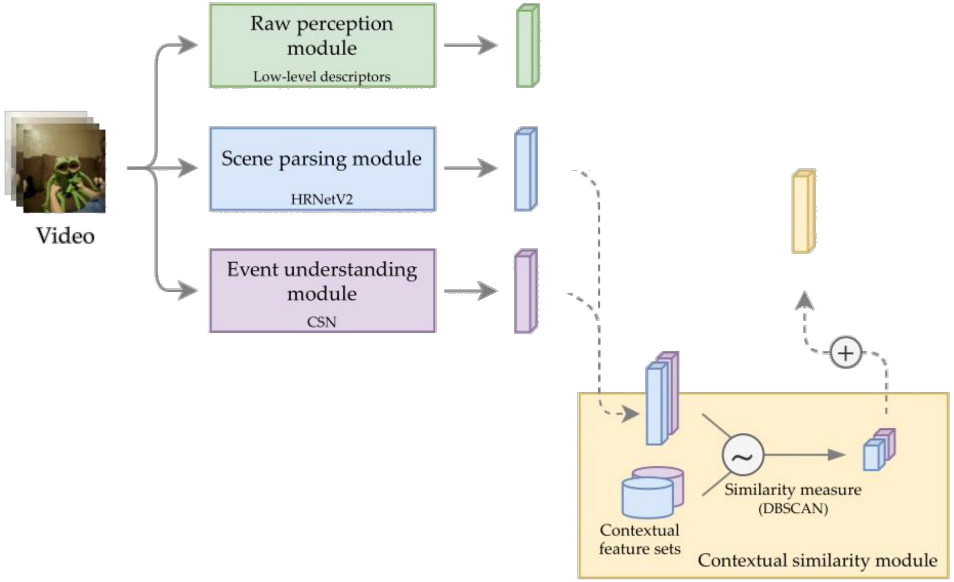
- low-level (green)
- mid-level (blue)
- high-level (purple)
- distinctiveness / similarity (yellow)



Our modular approach

Our model separates between different tiers **by design**:

- low-level (green)
- mid-level (blue)
- high-level (purple)
- distinctiveness / similarity (yellow)

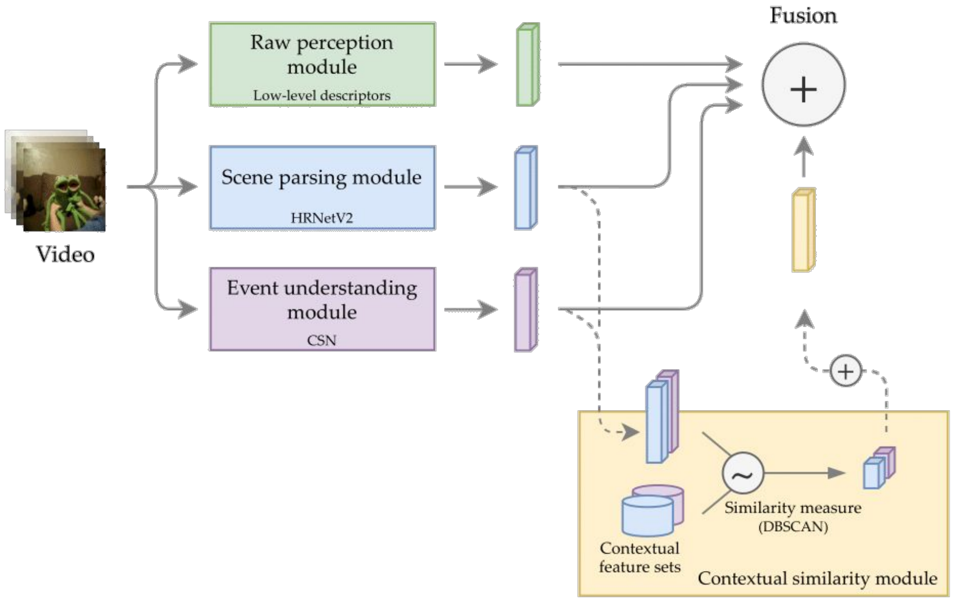


The model can be instantiated with **any combination of modules**!

Our modular approach

Our model separates between different tiers **by design**:

- low-level (green)
- mid-level (blue)
- high-level (purple)
- distinctiveness / similarity (yellow)

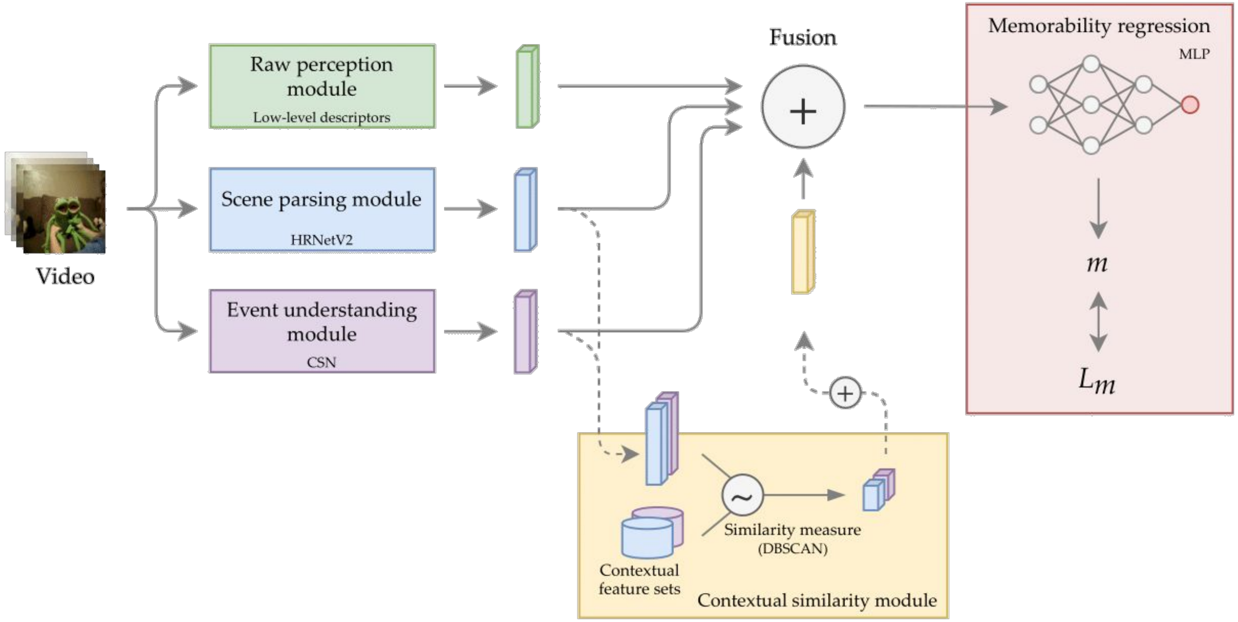


The model can be instantiated with **any combination of modules**!

Our modular approach

Our model separates between different tiers **by design**:

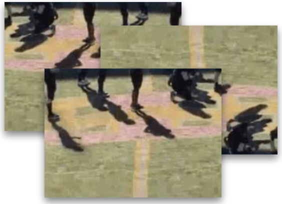
- low-level (green)
- mid-level (blue)
- high-level (purple)
- distinctiveness / similarity (yellow)



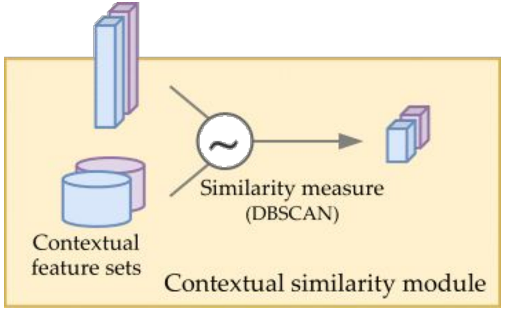
The model can be instantiated with **any combination of modules**!

The distinctiveness module

Idea: videos that **stand out** from a specific **context** are more memorable!

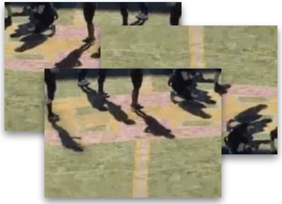


≠

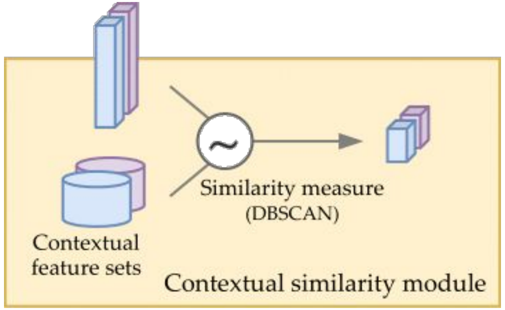


The distinctiveness module

Idea: videos that **stand out** from a specific **context** are more memorable!



≠

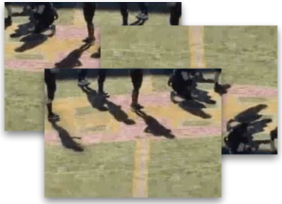


- comparing features to the corpus (here, training set) **in feature space** gives a measure of their distinctiveness

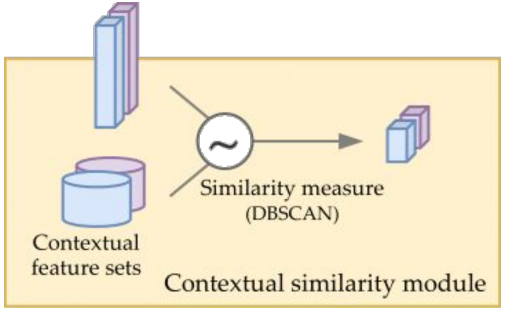
- a lot of ways to do this, some better than others:
  cosine similarity, Euclidean distance, kernel density estimation (KDE)...

The distinctiveness module

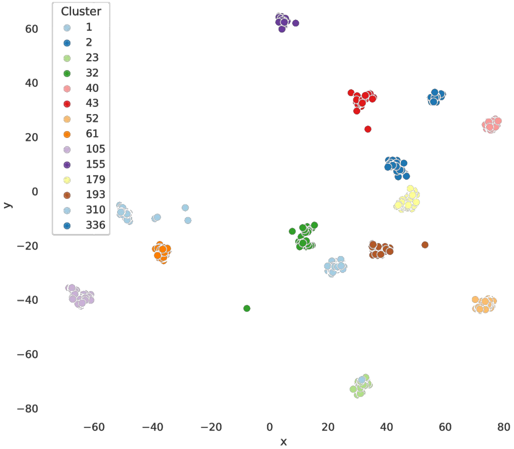
Idea: videos that **stand out** from a specific **context** are more memorable!



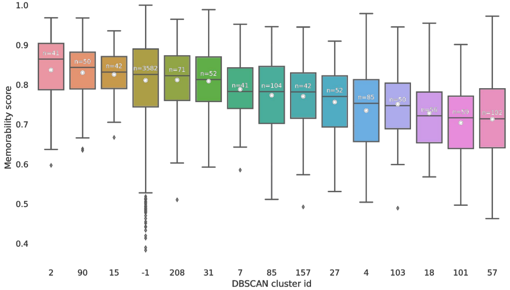
≠



- comparing features to the corpus (here, training set) **in feature space** gives a measure of their distinctiveness
- a lot of ways to do this, some better than others:
  cosine similarity, Euclidean distance, kernel density estimation (KDE)...
- we choose to **cluster** the corpus using **DBSCAN** and
  train a simple MLP classifier to **predict the labels** of videos

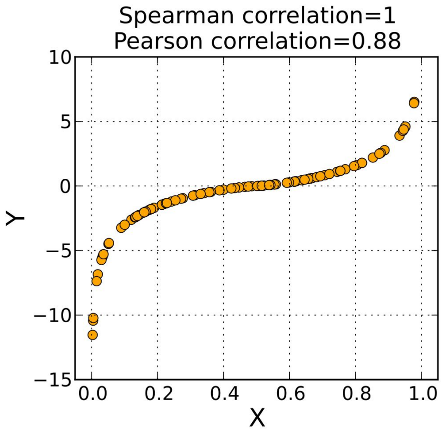


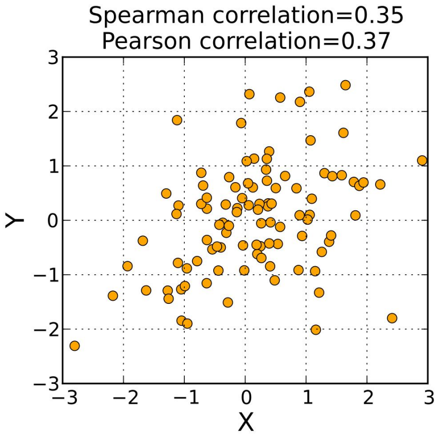
Result: a few clusters of videos that stand out from the rest of the corpus

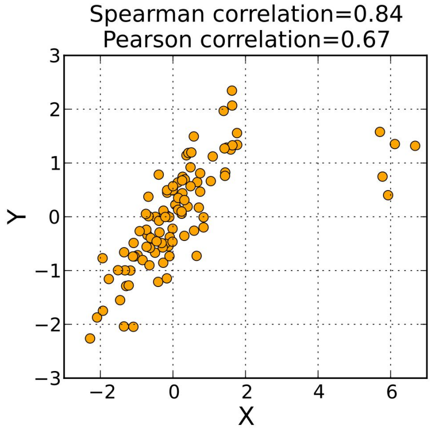


Training procedure

- evaluation metric: **Spearman rank correlation**
- training and evaluation on VideoMem and Memento10k (separately)







Memento10k

VideoMem

Training procedure

- evaluation metric: **Spearman rank correlation**
- training and evaluation on VideoMem and Memento10k (separately)
- **loss on Memento10k:** MSE with **tail penalization**

$$\mathcal{L}_1(m,\hat{m}) = \big[1+p(m)\big]L_{\text{MSE}}(m,\hat{m}),$$

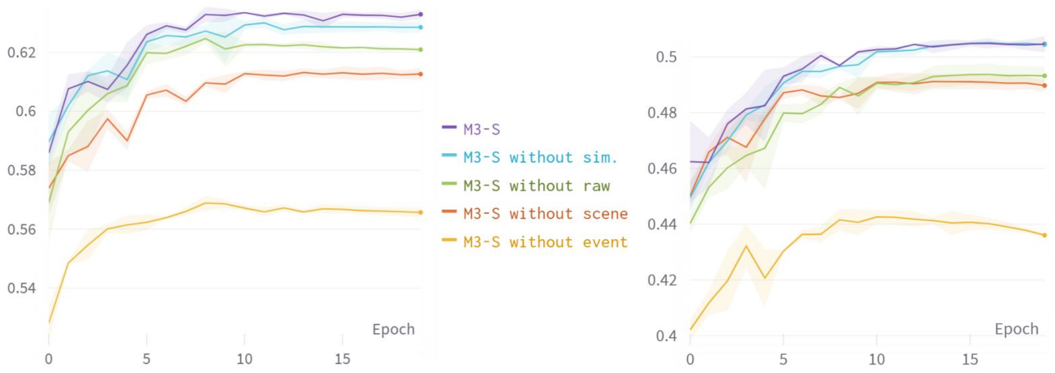
- **loss on VideoMem:** **weighted mean** between MSE and (smooth) Spearman RC

$$\mathcal{L}_2(m,\hat{m}) = (1-\alpha_{\text{ep}})L_{\text{MSE}}(m,\hat{m}) + \alpha_{\text{ep}}L_{\text{Spearman}}(m,\hat{m})$$

$$\alpha_{\text{ep}} = \frac{\text{ep}}{N_{\text{ep}}-1}$$

$$\text{ep} \in \{0,\ldots,N_{\text{ep}}-1\}$$

Parameter	Memento10k [6]	VideoMem [4]
Hidden channels	$[512, 64, 1]$	–
Batch size	32	–
Learning rate	10^{-3}	–
Scheduler	StepLR, $\gamma = 0.2$, step size $= 5$	–
Epochs	20	–
Loss	MSE (tails)	MSE + Spearman RC
Weight decay	10^{-5}	–
Optimizer	Adam	–
Normalizing raw	✓	–
Normalizing sim	✓	–



Memento10k VideoMem

Results

- our model **outperforms** existing approaches...

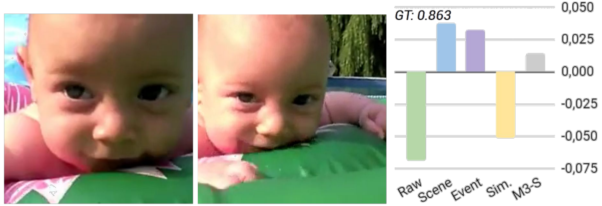
Approach	Spearman RC $\rho \uparrow$ Memento10k	VideoMem
MemNet baseline* [31]	0.485	0.425
Cohendet *et al*. (Semantic)* [13]	0.552	0.503
Cohendet *et al*. (ResNet3D)* [13]	0.574	0.508
SemanticMemNet† [40]	0.659	0.556
M3-S (ours)	**0.670**	**0.563**

Results

- our model **outperforms** existing approaches…

Approach	Spearman RC $\rho\uparrow$ Memento10k	Spearman RC $\rho\uparrow$ VideoMem
MemNet baseline* [31]	0.485	0.425
Cohendet *et al*. (Semantic)* [13]	0.552	0.503
Cohendet *et al*. (ResNet3D)* [13]	0.574	0.508
SemanticMemNet† [40]	0.659	0.556
M3-S (ours)	**0.670**	**0.563**

- … while keeping a degree of **interpretability**



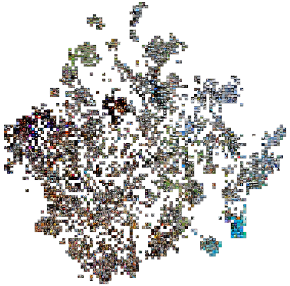
Memorable semantics, non-memorable motion, low distinctiveness

Memorable motion, non-memorable event, low distinctiveness.

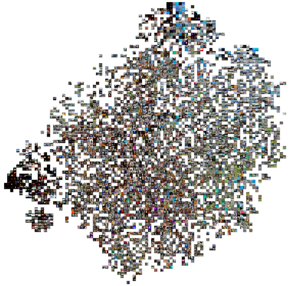
⇒ each module contributes to memorability prediction!

Results – feature representations

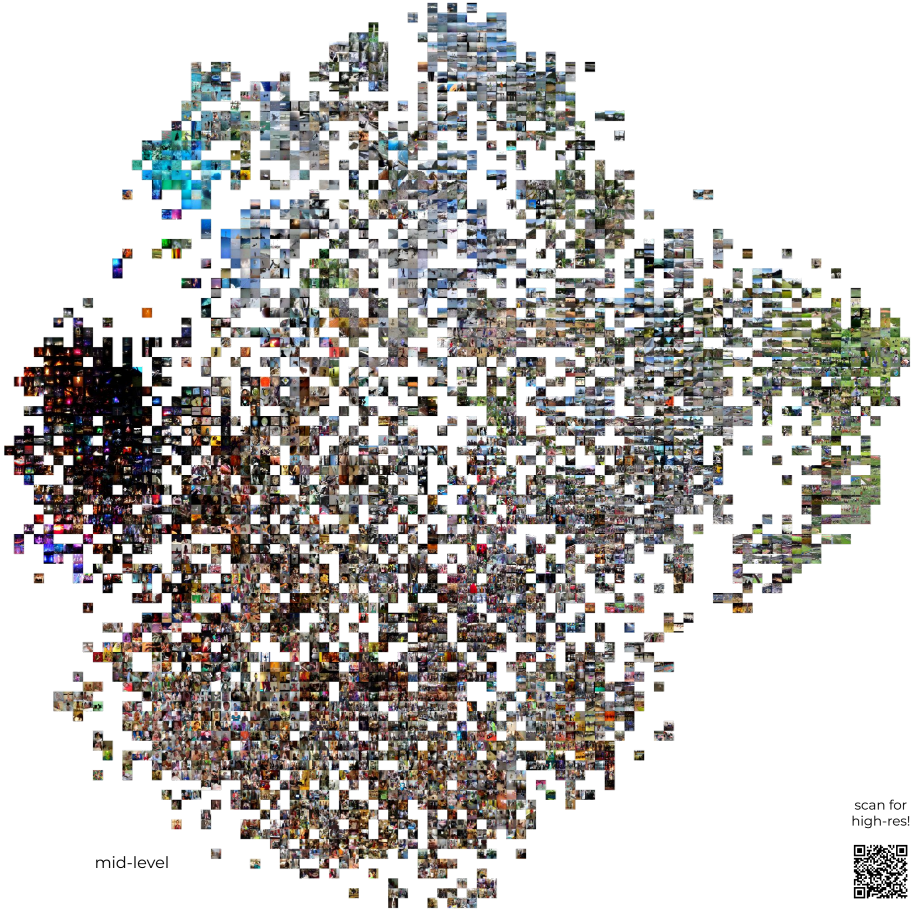
- each module learns representations that are **meaningful** and substantially **different from each other**



low-level



high-level



mid-level

scan for
high-res!

Limits

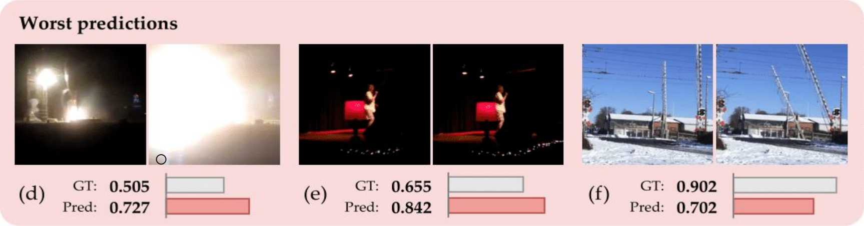
Our model...

- ... **is good** with
  - specific actions/objects **(a,b)**
  - peculiar semantic context **(c)**

Limits

Our model...

- ... **is good** with
  - specific actions/objects **(a,b)**
  - peculiar semantic context **(c)**
- ... **has trouble** with
  - strong variations **(d)**
  - blurriness **(e)**
  - complex semantics **(f)**

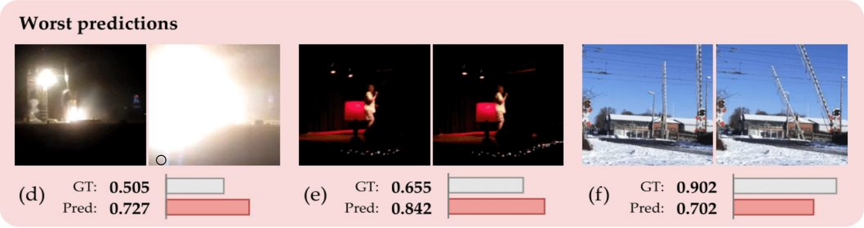


Limits

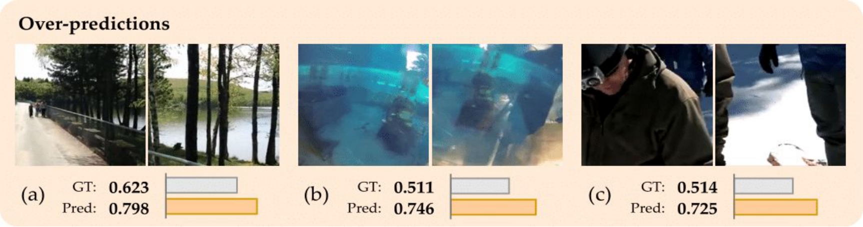
Our model...

- ... **is good** with
  - specific actions/objects **(a,b)**
  - peculiar semantic context **(c)**

- ... **has trouble** with
  - strong variations **(d)**
  - blurriness **(e)**
  - complex semantics **(f)**



- ... **overestimates** the memorability of scenes that
  - are semantically bland with humans **(a)**
  - are very dynamic with no clear action **(b)**
  - contain memorable elements, such as humans or faces, but that are very shaky **(c)**, cluttered or blurry.

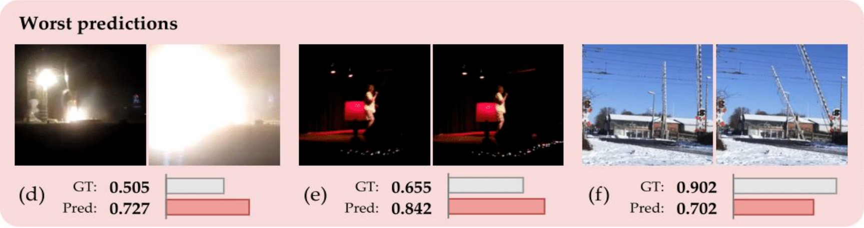


Limits

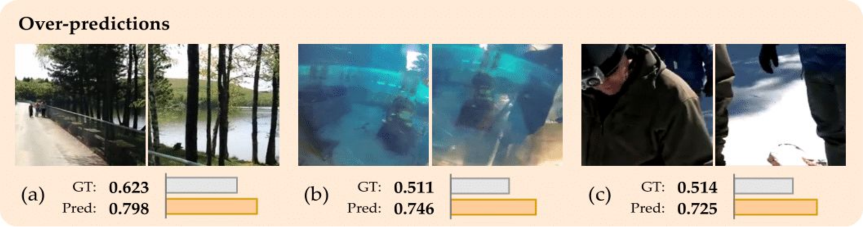
Our model...

- ... **is good** with
  - specific actions/objects **(a,b)**
  - peculiar semantic context **(c)**

- ... **has trouble** with
  - strong variations **(d)**
  - blurriness **(e)**
  - complex semantics **(f)**



- ... **overestimates** the memorability of scenes that
  - are semantically bland with humans **(a)**
  - are very dynamic with no clear action **(b)**
  - contain memorable elements, such as humans or faces, but that are very shaky **(c)**, cluttered or blurry.



- ... **underestimates** the memorability of scenes that are
  - emotionally salient (scary **(d)**, funny **(e)**)
  - bland with hard to grasp semantic content **(f)**.

Conclusion

Our contributions:

1. in-depth **analysis of factors of memorability** and **classification** in tiers
2. novel methodology using the classification and a measure of **distinctiveness**
3. leveraging the model’s structure to get insights on its **interpretability**

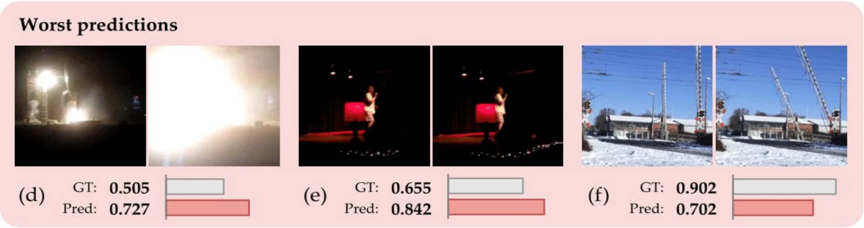
Conclusion

Our contributions:

1. in-depth **analysis of factors of memorability** and **classification** in tiers
2. novel methodology using the classification and a measure of **distinctiveness**
3. leveraging the model's structure to get insights on its **interpretability**

Future directions: video memorability remains an open problem!

- model often fails because of **complex semantic**, extreme **pixel intensity** or extreme **motion**
- room for understanding how to research each module
- overhaul high-level module through **emotion prediction**
  (bottleneck: no competitive model or dataset for video emotion prediction)



Modular Memorability:

Tiered Representations for Video Memorability Prediction

Théo Dumont[1,2], Juan S. Hevia[2], Camilo L. Fosco[2]

[1] Mines Paris, PSL   [2] Memorable AI

code and models available at
https://github.com/tekal-ai/modular-memorability

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

memorable

WED-AM-240