Understanding of emotion perception from art

Digbalay Bose[*,o], Krishna Somandepalli[*,o], Souvik Kundu[o], Rimita Lahiri[*,o], Jonathan Gratch[ǂ,], Shrikanth Narayanan[*,o]**

[*] Signal Analysis and Interpretation Laboratory

o Department of Electrical and Computer Engineering

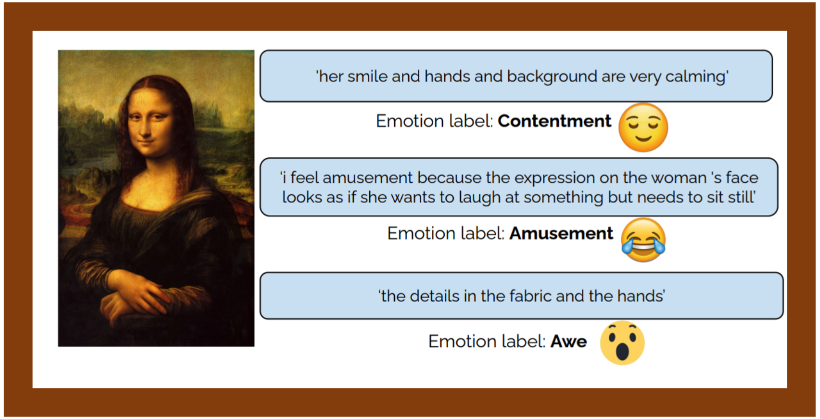
[**] Department of Computer Science

ǂ USC Institute of Creative Technologies

Art and Emotions

"A work of art which did not begin in emotion is not art" – Paul Cézanne

- Evoked emotion in viewers highly **subjective**.
- Variations in individual aesthetic experiences studied for observers using combination of fMRI and behavioral analysis [1].
- Art pieces from Wikiart annotated for 20 emotions and likeability [2].
- Subjectivity can be handled by explanation of *why* certain emotion was felt by a viewer [3]



Different captions and emotions associated with Monalisa painting from Artemis dataset [3]. Image source: Wikiart link

[1]: **Vessel et.al** : The brain on art: intense aesthetic experience activates the default mode network, Link
[2] **Mohammad et.al**: WikiArt Emotions: An Annotated Dataset of Emotions Evoked by Art, Link
[3] **Achiloptas et.al**: ArtEmis: Affective Language for Visual Art, **Link**

Explanations-only vs Multimodal cues

- Complementary cues present in text and image:
  - Perceptually affective cues in images
  - Direct signal about felt emotion in text caption.

- Emotion prediction using BERT based text classifier:
  "*sadness*"

- The artwork image when taken into context along with the caption evokes a feeling of "**contentment**" .

Sea-coast Crimean coast near Ai-petri painting by Ivan Aivazovsky. Image source: [Wikiart link]

Image-centric Affective cues

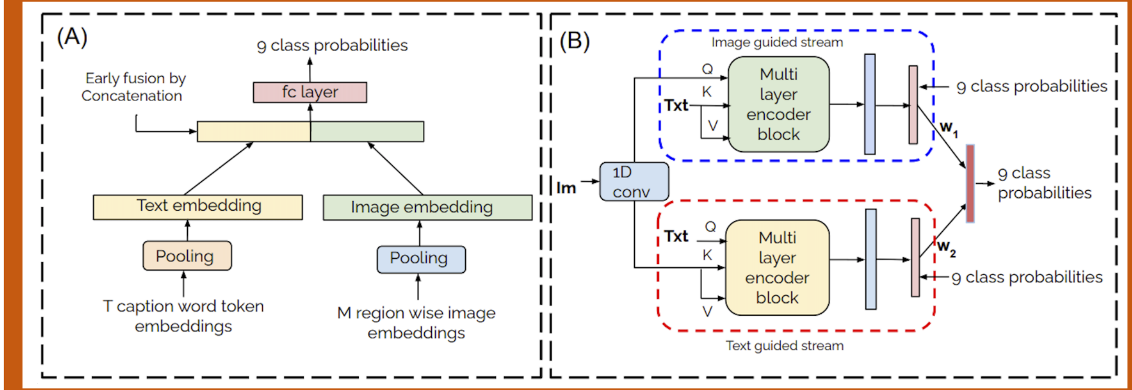
Objects | Expression | Subject | Style | Technique | Color

Text caption

It makes me think of how things used to be and how simple life once was

Caption source: Artemis dataset [Link]

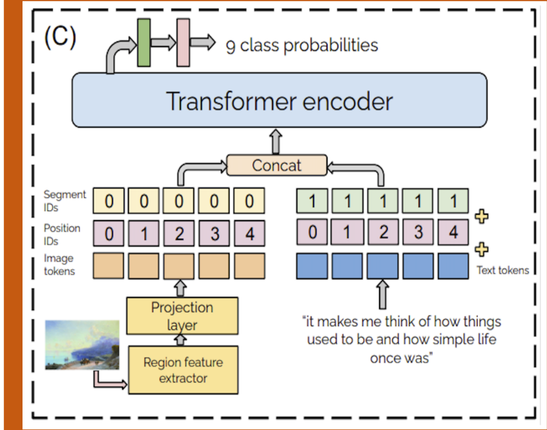
Multimodal Model adaptations



Dual stream models:

(A) Early fusion average pool / Early fusion first token

(B) Weighted late fusion (5 encoder layers, 8 heads, $w_1 = 0.76$ and $w_2 = 0.24$)



Single stream models:
(C) Single stream configurable MMBT [1] model

[1] Kiela e.al . Supervised Multimodal Bitransformers for Classifying Images and Text. Link

Results

Model	Acc	F1	Feat
Image (N = 79327)			
VGG-16	47.36	27.04	
ResNet-50	44.98	21.31	
Text (N = 429431)			
BERT	66.2	61.42	
Multimodal (N = 429431)			
Early fusion avg pool	56.35	46.72	BU+Bert
Early fusion first token	56.98	48.34	BU+Bert
Weighted late fusion	65.14	60.27	BU+Bert
MMBT	66.33	62.24	BU+Bert
VisualBERT	66.03	61.47	VinVL+Bert

❖ **Experiments conducted on Artemis :**

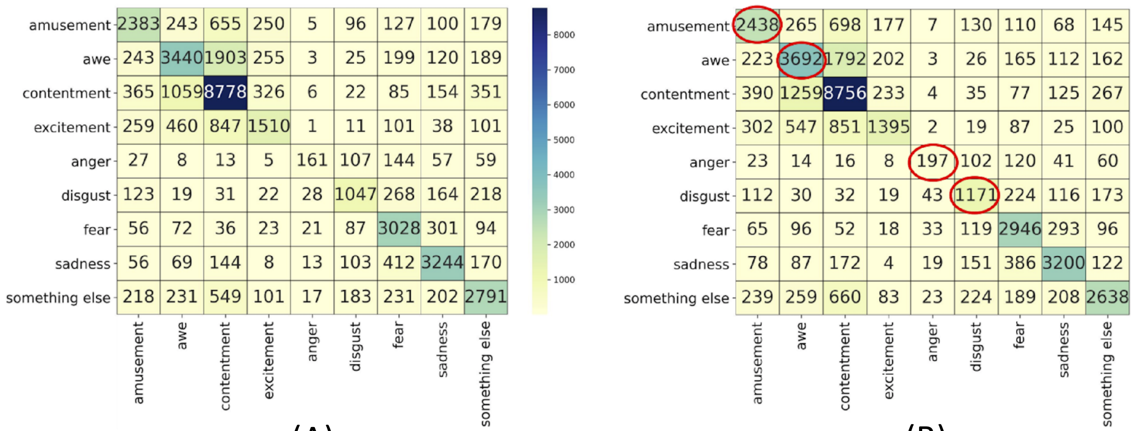
- 81446 art-work from Wikiart.
- 27 art styles from 15^{th} to 21^{st} century.
- 9 emotion classes.
- 429k textual captions.
- Train/val/test split same as [1].

❖ **Settings:**

- **BU[2]:** 2048 dim region features from top-50 salient regions using FasterRCNN with ResNet101 backbone.
- **VinVL[3]**: 2048 dim region features from top-50 salient regions using ResNeXt-152 C4 model.
- **Bert**: 768 dim token representations from pretrained BERT-base uncased model.
- KL-Divergence loss used for image-only models between network outputs and per-image distribution of emotions.
- Categorical cross-entropy with label smoothing used for training the multimodal models.

[1] **Achiloptas et.al**: ArtEmis: Affective Language for Visual Art, Link
[2] **Anderson et.al**: Bottom-up and top-down attention for image captioning and visual question answering, Link
[3] **Zhang et.al**: Vinvl: Revisiting visual representations in vision-language models, Link

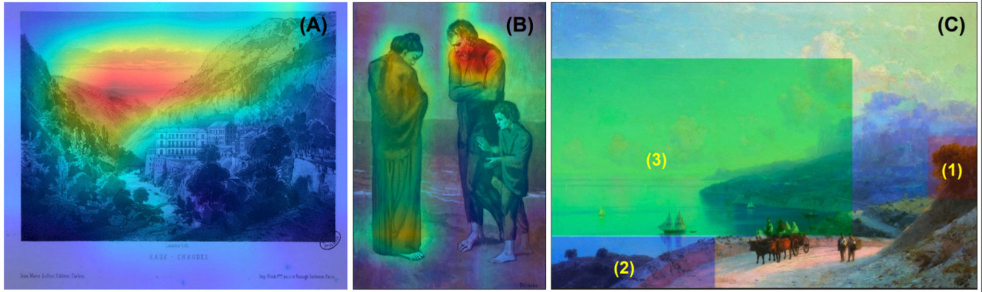
Visualizations



(A) (B)

(A) Confusion matrix of BERT (text-based classification)

(B) Confusion matrix of MMBT (red circles indicate classes where MMBT performs better)



Majority GT class: awe
Predicted class: awe

Majority GT class: sadness
Predicted class: sadness

Caption: "it makes me think of how things used to be and how simple life once was"
GT class: contentment **Predicted class:** contentment

(A) : VGG-16: Grad cam visualization for correctly predicted class "**awe**".

(B) : VGG-16: Grad cam visualization for VGG-16 for correctly predicted class "**sadness**".

(C): MMBT: Top-3 image regions in t... s of gradient based attributions (1)-(3) for correctly predicted class "**contentment**"

Summary

- Single stream multimodal models like MMBT and VisualBERT perform better when compared with dual-stream multimodal models and image-only models.
- Predicting a single emotion label from an art-work image is difficult due to multiple interpretations.
- On the visual side, **art-style** based and holistic image features along the lines of **color, lighting** can improve emotion understanding from art-work.

Thank you