Emotion-Aware Music Recommendation

Hieu Tran*, Tuan Le*, Anh Do*, Tram Vu, Steven Bogaerts, Brian Howard
Department of Computer Science
DePauw University
Greencastle, IN, USA

(* Equal contribution)

Abstract

People often listen to songs that match their mood. Thus, an AI music recommendation system that is aware of the user's emotions is likely to provide a superior user experience to one that is unaware. In this work, we present an emotion-aware music recommendation system. Multiple models are discussed and evaluated for affect identification from a live image of the user. We propose two models: DRViT, which applies dynamic routing to vision transformers, and InvNet50, which uses involution. All considered models are trained and evaluated on the AffectNet dataset. Each model outputs the user's estimated valence and arousal under the circumplex model of affect. These values are compared to the valence and arousal values for songs in a Spotify dataset, and the top-five closest-matching songs are presented to the user. Experimental results of the models and user testing are presented.

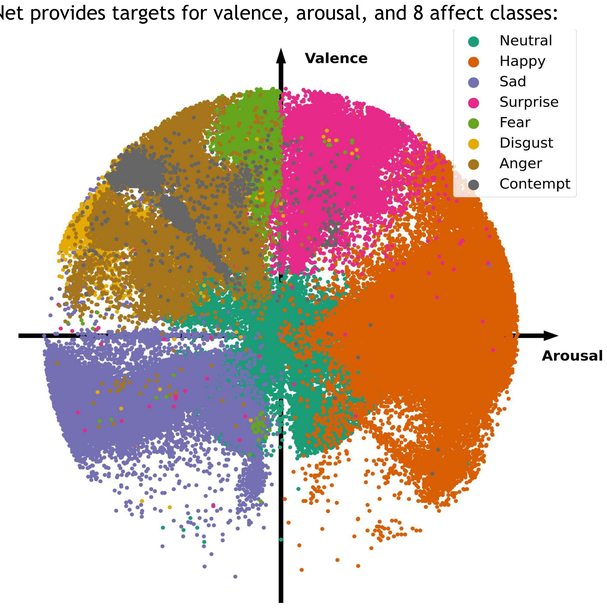
System Overview

Image Acquisition and Processing

In the web application, the user provides a picture of their face using the front-facing camera of their device.

Affect Identification

- The system runs the image through an affect identification model to predict the valence and arousal values, two continuous ranges representing human emotion:
  - Valence: level of positivity or negativity
  - Arousal: level of energy
- Our system randomly chooses one of the two following models trained on the AffectNet dataset (Mollahosseini, Hasani, and Mahoor 2017):
  - Dynamic Routing for Vision Transformers (DRViT)
  - Involution Residual Network with 50 layers (InvNet50)
- AffectNet provides targets for valence, arousal, and 8 affect classes:

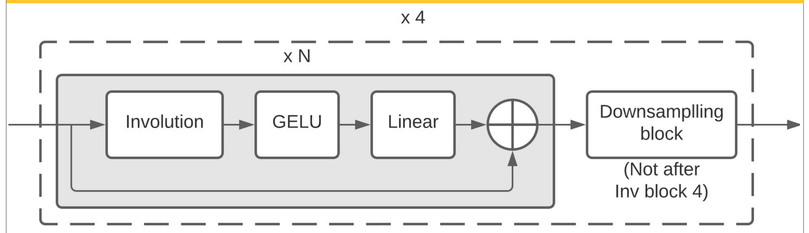


- After generating the valence-arousal values, the system passes them to a decision tree classifier to interpret the values as English words.

Music Recommendation

- The system then uses a 600k-song Spotify dataset to match the songs' normalized valence and arousal values with the pairs generated from the model.
- The top-five songs (according to nearest neighbor on the valence-arousal plane) are recommended to the user. Users may listen to 30-second clips or follow a link to the song on Spotify.

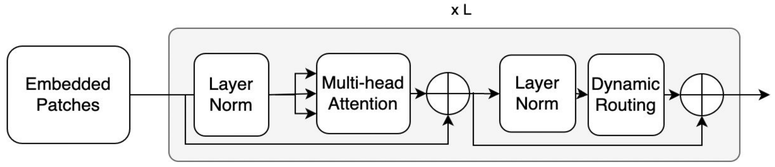
Affect Identification Architecture #1: InvNet50



- Involution kernel of Li et al. (2021):
  - Spatial-specific: aware of spatial relationships.
  - Channel-agnostic: ignorant of channel-specific features.
- InvNet50 also adopts skip-connection, dropout, layer, and batch normalization to reduce overfitting.

Layer Name	Configuration
Stem layer	(Conv) 7×7, 64, stride 2 (Max Pooling) 2×2, stride 2
Inv block no.1 ($N = 3$)	(Inv) 7×7, 64, stride 1
Downsample block no.1	(Conv) 3×3, 128, stride 2
Inv block no.2 ($N = 4$)	(Inv) 7×7, 128, stride 1
Downsample block no.2	(Conv) 3×3, 256, stride 2
Inv block no.3 ($N = 6$)	(Inv) 7×7, 256, stride 1
Downsample block no.3	(Conv) 3×3, 512, stride 2
Inv block no.4 ($N = 2$)	(Inv) 7×7, 512, stride 1
	Average pool, 2-d fc or 1-d fc

Affect Identification Architecture #2: DRViT



We replace the feed-forward neural network layers in each encoder block of Vision Transformers of Dosovitskiy et al. (2020) with Dynamic Routing proposed by Sabour, Frosst, and Hinton (2017).

- L=3 encoder blocks
- Number of heads: 8
- Dimension of embedding layers: 256
- Last layer: feed-forward perceptron.

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}$$

Capsule layer output:

$$\hat{\mathbf{u}}_{j|i} = \sum_i \mathbf{W}_{ij} \times \mathbf{u}_i$$

$$\mathbf{s}_j = \sum_i c_{ij}\hat{\mathbf{u}}_{j|i}$$

$$\mathbf{v}_j = \frac{||\mathbf{s}_j||^2}{1 + ||\mathbf{s}_j||^2}\frac{\mathbf{s}_j}{||\mathbf{s}_j||}$$

Experimental Setup

Augmentation

- Both plans:
  - 20,000 images selected per class (or fewer if class size < 20,000)
  - Random application of: no augmentation, Gaussian blur, horizon flip, color jitter, random erasing
  - Performed online (during training)
- Plan A:
  - Image randomly selected from among the dataset up to 20,000 per class.
  - Augmentation (or no augmentation) is applied as above.
  - Result: 160,000 images with perfect balance between classes
- Plan B:
  - Considers each previously selected image once.
  - Augmentation (or no augmentation) is applied as above.
  - Result: 108,021 images with imbalance between classes

Models x outputs

- 2 x 1: two models with the same architecture and one output, one trained on valence, one on arousal
- 1 x 2: one model with two outputs trained on valence and arousal

Experiment Results and Analysis

ID	Arch	M × O	Params (M)	Aug	Valence RMSE (<)	Valence CORR (>)	Valence CCC (>)	Valence SAGR (>)	Arousal RMSE (<)	Arousal CORR (>)	Arousal CCC (>)	Arousal SAGR (>)
1	AlexNet	2×1	2×58.2	No	0.37	0.66	0.60	0.74	0.41	0.54	0.34	0.65
2	ResNet50	2×1	2×25.0	No	0.41	0.58	**0.53**	**0.68**	0.43	0.46	**0.47**	0.65
3	ResNet50	2×1	2×25.0	A	**0.39**	**0.59**	**0.53**	0.67	**0.40**	**0.48**	0.41	**0.66**
4	ViT	2×1	2×85.0	No	0.40	**0.58**	0.55	**0.66**	0.42	0.50	**0.46**	0.62
5	ViT	2×1	2×85.0	A	**0.39**	0.57	**0.56**	0.65	**0.39**	**0.52**	0.41	**0.68**
6	InvNet50	2×1	2×10.5	No	0.43	0.57	0.53	0.72	0.36	0.50	0.43	0.75
7	InvNet50	2×1	2×10.5	B	0.37	0.63	0.61	0.76	0.34	0.53	0.49	0.78
8	InvNet50	1×2	10.5	No	0.42	0.59	0.55	0.73	0.36	0.51	0.45	0.74
9	InvNet50	1×2	10.5	A	**0.36**	0.62	0.57	**0.77**	**0.33**	0.51	0.42	0.79
10	InvNet50	1×2	10.5	B	0.37	**0.65**	**0.63**	**0.77**	**0.33**	**0.55**	**0.52**	**0.80**
11	DRViT	1×2	13.0	No	**0.36**	**0.68**	**0.66**	0.78	0.36	**0.67**	**0.53**	0.75
12	DRViT	1×2	13.0	A	0.37	0.66	0.63	**0.79**	**0.35**	0.65	0.48	**0.77**
13	DRViT	1×2	13.0	B	0.39	0.61	0.57	0.72	0.37	0.56	0.48	0.63

- ResNet50 and ViT fail to significantly improve upon AlexNet.
- 1x2 versus 2x1 shows little difference in performance, but 1x2 is more efficient.
- DRViT and InvNet50 give better results than AlexNet, ResNet50 and ViT.
- DRViT prefers no augmentation, while InvNet50 prefers augmentation plan B.
  - Due to a full attention mechanism and dynamic routing, DRViT may require more data than InvNet50, hence the preference for no augmentation over random selection and augmentation plans.

Best InvNet50: 1×2 with augmentation B (row 10).
Best overall: DRViT 1×2 with no augmentation (row 11).

Application Overview

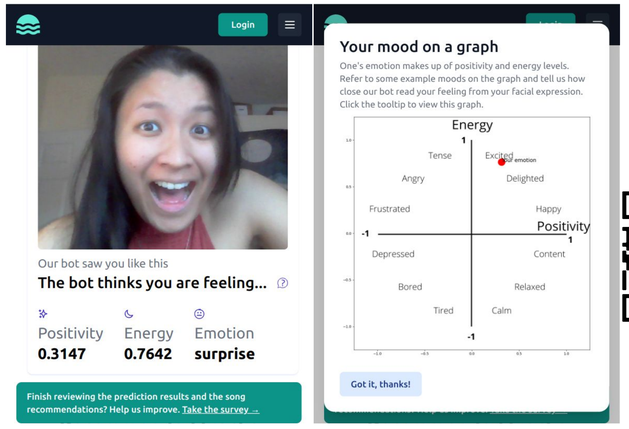


Image acquisition

Affect identification

Scan the QR code below to try it out.

Not shown: Music rec.

Acknowledgement

We are grateful for the support of the Tenzer Center, the J. William Asher and Melanie J. Norton Endowed Fund in the Sciences, and the Kranbuehl, Roberts, and Hillger Endowed Fund for Faculty Summer Research.

References

- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. CoRR, abs/2010.11929.
- Li, D.; Hu, J.; Wang, C.; Li, X.; She, Q.; Zhu, L.; Zhang, T.; and Chen, Q. 2021a. Involution: Inverting the Inherence of Convolution for Visual Recognition. CoRR, abs/2103.06255.
- Mollahosseini, A.; Hasani, B.; and Mahoor, M. H. 2017. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. IEEE Transactions on Affective Computing, PP(99): 1-1.
- Sabour, S.; Frosst, N.; and Hinton, G. E. 2017. Dynamic routing between capsules. Advances in neural information processing systems, 30.