RETHINKING GENRE CLASSIFICATION WITH FINE GRAINED SEMANTIC CLUSTERING

Edward Fish, Jon Weinbren, Andrew Gilbert (edward.fish, j.weinbren, a.gilbert)@surrey.ac.uk

Centre for Creative Arts and Technology, University of Surrey, Guildford, UK

ABSTRACT

Movie genre classification is an active research area in machine learning; however, the content of movies can vary widely within a single genre label. We expand these 'coarse' genre labels by identifying 'fine-grained' contextual relationships within the multi-modal content of videos. By leveraging pre-trained 'expert' networks, we learn the influence of different combinations of modes for multi-label genre classification. Then, we continue to fine-tune this 'coarse' genre classification network self-supervised to sub-divide the genres based on the multi-modal content of the videos. Our approach is demonstrated on a new multi-modal 37,866,450 frame, 8,800 movie trailer dataset, MMX-Trailer-20, which includes pre-computed audio, location, motion, and image embeddings.

1. INTRODUCTION

Genre labels are a useful device for concisely describing a movie's narrative, theme, and style. However, within a single genre, we can find a huge range of audio-visual diversity. Furthermore, in film theory, it has been shown that the semantics of specific genres has shifted throughout film history [23]. With this in mind, we propose that genre labels should be considered a weak labelling methodology and present a self-supervised clustering solution for identifying semantically similar information between videos that share similar genre labels.

To do so, we exploit expert knowledge in the form of semantic embedding 'experts', including scene understanding, image content analysis, motion, and audio. First, using collaborative gating as outlined in [17, 20], we train an encoder network for the task of multi-label genre classification to act as a weak proxy. Then inspired by [18, 23, 2], we attach a projection head and continue to train the model self-supervised to break apart genre clusters into sub-genres by fine-tuning the encoder network using a contrastive loss.

As in other works [24, 14, 30, 28, 26], we use movie trailers as they offer a condensed representation of a movies theme and content. First, we demonstrate the effectiveness of a multi-modal, collaboratively gated network for multi-label coarse genre classification of up to 20 genres. Then we implement fine-grained semantic clustering of genres via self-supervised learning for retrieval and exploration, demonstrating the results in a new large 37M frame multi-label genre dataset with pre-processed expert embeddings. The reader can find a detailed overview of the dataset and a description of the embedding pre-processing in the supplementary material.

Earlier techniques in this field pertain to extracting low-level audiovisual descriptors. Huang(H.Y.) et al. [13] used two features - scene transitions and lighting. In contrast, Jain & Jadon [16] applied a simple neural network with low-level image and audio features. Huang(Y.F.) & Wang [14] used the SAHS (Self Adaptive Harmony Search) algorithm in selecting features for different movie genres learnt using a Support Vector Machine with good results. Zhou et al. [31] predicted up to four genres with a BOVW clustering technique. Musical scores have also shown to offer a helpful mode for classification as in the work of Austin et al. [4] who predicted genre with spectral analysis using SVMs. More recent work has utilised deep learning and convolutional neural networks for genre classification. Wehrmann & Barros [28, 29] used convolutions to learn the spatial as well as temporal characteristic-based relationships of the entire movie trailer, studying both audio and video features. Shambharkar et al. [25] introduced a new video feature and three new audio features that proved useful in classifying genre, combining a CNN with audio features to provide promising results. While [26], employed 3D ConvNets to capture both the spatial and temporal information present in the trailer. The 'interestingness' of movies has also been predicted by audiovisual features [5]. Gating strategies for combining expert networks have been explored in [15, 27, 10, 19, 17].

2. METHODOLOGY

Fig. 1 presents an overview of our approach. First, using four multi-modal 'experts', we extract audio and visual features from the input video. Then, to enable genre classification, inspired by [17, 20], a collaborative gating model learns to emphasise or downplay combinations of these features. Finally, we train the network to develop fine-grained semantic clusters through self-supervised training, maximising the cosine similarity between sub-sequences within the trailers embedding vectors obtained from the same movie trailer (positive examples) while pushing negative sequence pairs apart.

Given a set of videos $\mathbf{v}$, each video is made up of a collection of sequences, $\mathbf{s}$, so $\mathbf{v} = \{s^1, s^2, ...s^m\}$, where there are m sequences in a video and each sequence is formed of n clips, giving $\mathbf{s} = \{c^1, c^2, ...c^n\}$. The aim of this work is to create a function Φ that can map a clip c from a video sequence $\mathbf{s}$, where $c \in \mathbf{s} \in \mathbf{v}$ to a joint feature space x_i that respects the difference between clips. To construct our function Φ, we rely on

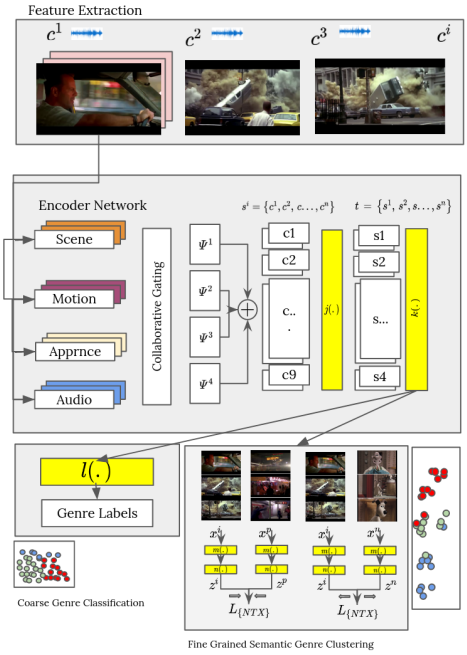


Fig. 1: An overview of the approach. Image and audio features from movie trailers are extracted and their influence is learnt via a collaborative weighting to classify broad genres such as Action, Adventure and Sci-Fi. A self-supervised network then compares these embeddings to generate contextually appropriate sub-genres.

several pre-trained single modality *experts*, $\{\Psi^1, \Psi^2, ...\Psi^E\}$, with E experts and Ψ^e is the e'th expert. These operate on the video or audio data and project the clip to an individual variable length embedding. Given that the embeddings Ψ are variable lengths, we aggregate the embeddings along their temporal component to form a standard vector size. You could use any temporal aggregation here, but we use average pooling for the video-based features. While for audio, we implement NetVlad [3], inspired by the vector of locally aggregated descriptors, commonly used in image retrieval. We apply linear projections to transform these task-specific embeddings to a standard dimensionality to enable their combination in the following collaborative gating phase.

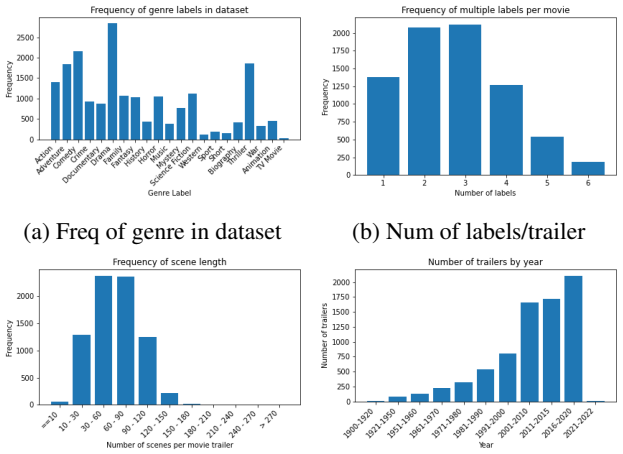
Collaborative Gating Unit: The Collaborative Gating Unit first proposed in [20] aims to achieve robustness to noise in the features through two mechanisms: **(i)** the use of information from a wide range of modalities; **(ii)** a module that aims to combine these modalities in a manner that is robust to noise. To learn the optimum combination of the expert embeddings we define a single attention vector for the e'th expert, then modulate the expert responses with the original data. To create the e'th expert's projection T^e, the attention vector of an expert projection will consider the potential relationships between all pairs associated with this expert, $T^e(\mathbf{v}) = \mathbf{h}(\sum_{\forall E}^{f \neq e} \mathbf{g}(\Psi^i(\mathbf{v}), \Psi^f(\mathbf{v})))$. This creates the projection between expert e and f, where $\mathbf{g}(.)$ is used to infer the pairwise task relationships while $\mathbf{h}(.)$ maps the sum of all pairwise relationships into a single attention vector T^e, and $\mathbf{v}$ is the set of sequences. Both $\mathbf{h}(.)$ and $\mathbf{g}(.)$ are defined as multi-layer perceptrons (MLPs). To modulate the result, we take the attention vectors $T = \{T^1(\mathbf{v}), T^2(\mathbf{v}), ..., T^e(\mathbf{v})\}$ and perform element wise multiplication with the initial expert embedding vector which results in a suppressed or amplified version of the original expert embedding. Each expert embedding is then passed through a Gated Embedding Module (GEM) [21] before being concatenated together into a single fixed length vector for the clip. We capture 9 clip embeddings before concatenating and passing through an MLP to obtain a sequence embedding. These sequence representations are then concatenated together before being passed through a bottleneck layer which learns a compact embedding for the whole trailer. We can train the trailer embedding obtained from the collaborative gating unit in conjunction with genre labels to enable classification. First, the sequence embeddings x are summed over a trailer and then projected via an MLP $\mathbf{k}(.)$ to produce a logits embedding. We then minimise a Binary Cross Entropy Logits Loss until convergence to make coarse genre predictions.

Fine Grained Semantic Genre Clustering: As discussed in the introduction, discreet genre labels are restrictive and only offer a broad representation of a video's content. We aim to find finer-grained semantic content by identifying similarities in the sound, locations, objects, and motion within the videos. To achieve this, we extend the pre-trained coarse genre classification model with a self-supervised contrastive learning strategy using a normalised temperature-scaled cross-entropy loss [8], $\mathcal{L}_{\mathcal{NTX}}$. In [8], image augmentations are used as comparative features to fine-tune the embedding layer of a classification network to encourage more significant cosine similarity between augmentations obtained from the same image. We extend this method to video, splitting the video into two equal sequence lengths and then using these embeddings as the representation pairs x. When splitting the video, we sample sequences randomly from the trailer to obtain a good distribution of content.

$$\mathcal{L}_{\mathcal{NTX}}(x) = -\log \frac{exp(\text{sim}(\mathbf{m}(x_i), \mathbf{m}(x_p)/\tau)}{\sum_{k \neq j}^{2N} \mathbb{1}_{n \neq p} exp(\text{sim}(\mathbf{m}(x_i), \mathbf{m}(x_n))/\tau} \quad (1)$$

Here x_i, x_p and x_n are the feature representations and **m(.)** represents a projection head encoder formed from MLPs, $\tau > 0$ is a temperature parameter set at 0.5 and sim is the cosine similarity metric. x_i and x_p are two embedding vectors obtained from the same video as described above, while x_n is an embedding vector from another video. Here, the $\mathcal{L}_{NCE}$ loss will enforce x_i closer in cosine similarity to x_p but further from x_n. This process is illustrated in the overview Fig. 1.

After training, the MLP projection head ($\mathbf{m}(.)$) is removed, and we use the bottleneck layer of the collaborative gating model as a pre-trained embedding projection network to obtain feature embeddings for clustering and retrieval.



(a) Freq of genre in dataset (b) Num of labels/trailer

(c) Scene length (d) Year of distribution

Fig. 2: MMX-Trailer-20 Dataset statistics.

3. RESULTS AND DISCUSSION

MMX-Trailer-20: Multi-Model eXperts Trailer Dataset: There are several datasets upon which previous works test. However, to capture the scale and variability of a dataset is challenging, especially in terms of diversity of genre, size of dataset, and year of distribution. Tbl. 1 shows the comparison in size and labelling between recent works in genre classification.

Table 1: Movie genre datasets

Dataset	Video Source	Number Trailers	Frames	Label Source	Num. Genres	Genre/ Trailer
Rasheed [24]	Apple	101	-	-	4	1
Huang [14]	Apple	223	-	IMDb	7	1
Zhou [30]	IMDb+Apple	1239	4.5M	IMDb	4	3
LMTD-9 [28]	Apple	4000	12M	IMDb	9	3
Moviescope [7]	IMDb	5000	20M	IMDb	13	3
MMX-Trailer-20	Apple+YT	8803	37M	IMDb	20	6

Most datasets are small with limited numbers of genre labels, both in terms of variability and the number assigned to a single trailer. Moviescope [7] is the closest to the proposed dataset, with 3 genre labels and 5000 trailers; however, we increase the number of trailers, labels, and frames. Our dataset totals 8803 movie trailers drawn from Apple Trailers and YouTube, with 37,866,450 individual video frames. You can see the statistics of the dataset in Fig. 2 where we show that a wide range of genres exists, with each trailer featuring, on average, three labels. The distribution years of the trailers are also more diverse than current datasets, with MMX-Trailer-20 featuring movie trailers from the 1930s to the present day.

The dataset has 20 genres - Action, Adventure, Animation, Comedy, Crime, Documentary, Drama, Family, Fantasy, History, Horror, Music, Mystery, Science-Fiction, Western, Sport, Short, Biography, Thriller and War, with up to six genre labels for each trailer. Every trailer is a compact encapsulation of the full movie through a short 2 to 3 minute video, and we collect a weak proxy of genre classification by matching the trailer to its user-generated entry on the website `imdb.com`.

Evaluation Metrics: We use the standard retrieval metrics as proposed in prior work [9, 20, 22]. The $\overline{AU(PRC)}$ (micro average), $AU(\overline{PRC})$ (macro average), and $AU(\overline{PRC})_w$ (weighted average). We also show weighted Precision (P_w), weighted Recall (R_w), and weighted F1-Score ($F1_w$). For all metrics, higher is better.

Coarse Grained Genre Classification Results: Tbl. 2 illustrates the quantitative performance of the coarse genre prediction model *MMX-Trailer-20* and the global metrics. The table also shows the random baseline, which varies according to the frequency of the genre in the dataset. Finally, the table explores each experts' influence via an MLP on the coarse genre classification task. Using collaborative gating yields a 10% increase in basic fusion through concatenation. Unfortunately, audio and scene are the weakest experts for the classification task, which could be due to features that are not genre-specific, such as dialogue and external environments. We find all visual experts perform best on Animation, most likely due to its unique style compared with the other trailers. In contrast, the audio expert performs better in Comedy and Sport. To identify the importance of the collaborative gating unit, we compute a naive concatenation of the feature embeddings from the experts passed through an MLP layer (Naive Concat), with a 10 point reduction which illustrates the importance of the learnt collaborative gating framework.

Tbl. 3 shows the best performance of other approaches on different datasets. Our model, *MMX-Trailer-20* uses up to 6 genre labels per sample from 20 genres, double most other approaches and will affect the random baseline, which is nearly half that of the 9 genre datasets. To contextualise our method with others we compare previous approaches including low level video features (**VLLF**) [24], audio-visual features (**AV**) [14, 7], and audio-visual features with convolutions over time **CTT-MMC** [28]. From the results in Tbl. 3 we show that our model performs better than low-level features. We do not improve performance on other audiovisual approaches that fine-tune pre-trained networks in an end to end manner as we only train the collaborative gating layers and generate 'expert' embeddings offline for efficient retrieval and publication.

Fined Grained Genre Exploration: We evaluate the self-supervised model by comparing the cosine similarity between embedding vectors obtained from the encoder after training the classification network and following self-supervised fine-tuning. This is visualised in Fig. 3(a), where the T-SNE plot shows the learnt embedding for the coarse genre classification. Fig. 3(b) is after the model's self-supervised training, where we can see how the clusters have broken up into an overlapping distribution as genres are separated depending on the multi-modal content. Three trailers (Cleopatra, Braveheart, and Darkest Hour) share the triple genre classification of *Drama, Biography, History* and the coarse genre encoder correctly clusters and labels these together despite the significant differences in their content. In Fig. 3(b), we find that Cleopatra is drawn closer to Adventure films featuring deserts and orchestral scores (Lawrence of Arabia is one example). Braveheart shares a similarity with medieval and mythological trailers featuring large scale battles, while Darkest Hour moves

Table 2: Coarse genre classification of the MMX-Trailer-20 dataset. Across differing expert features and combinations methods

Model	Actn	Advnt	Animtn	Bio	Cmdy	Crme	Doc	Drma	Famly	Fntsy	Hstry	Hrror	Mystry	Music	SciFi	Wstrn	Sprt	Shrt	Thrll	War	$F1_w$	$AU(\overline{PRC})_w$	P_w	R_w
Support	130	197	46	13	224	102	87	267	117	115	44	104	41	86	107	181	30	45	12	21	-	-	-	-
Random	0.29	0.41	0.11	0.03	0.46	0.24	0.21	0.52	0.27	0.26	0.11	0.24	0.1	0.2	0.25	0.39	0.08	0.11	0.03	0.05	0.318	0.134	0.19	1
Scene [11]	0.43	0.55	0.74	0	0.49	0.38	0.63	0.55	0.51	0.28	0.24	0.42	0.3	0.28	0.41	0.51	0.22	0.19	0.11	0.33	0.434	0.489	0.437	0.48
Audio [1]	0.47	0.51	0.40	0.10	0.61	0.38	0.58	0.55	0.51	0.37	0.11	0.34	0.39	0.30	0.35	0.55	0.16	0.15	0.13	0.12	0.454	0.449	0.400	0.537
Motion [6]	0.5	0.59	0.74	0	0.62	0.33	0.63	0.56	0.55	0.36	0.2	0.38	0.45	0.24	0.37	0.57	0.23	0.14	0.10	0.13	0.463	0.487	0.448	0.494
Image [12]	0.48	0.63	0.79	0.12	0.65	0.41	0.60	0.59	0.55	0.42	0.25	0.47	0.42	0.29	0.50	0.54	0.34	0.19	0.12	0.31	0.516	0.554	0.493	0.572
Image + Audio	0.52	0.63	0.78	**0.15**	0.65	0.42	0.68	0.6	0.63	0.46	0.25	0.50	0.51	0.34	0.49	0.59	0.38	**0.28**	0.12	0.42	0.544	0.558	0.476	0.65
Image + Motion	0.59	0.64	0.78	0	0.59	0.39	0.66	0.6	0.6	0.5	0.29	0.54	0.53	0.25	**0.52**	0.57	0.4	0.2	0.24	0.12	0.535	0.553	0.511	0.583
Image + Scene	0.52	0.61	0.80	0.12	0.61	0.37	0.65	**0.62**	0.58	0.49	0.15	0.51	0.49	0.37	0.48	0.56	**0.43**	0.26	0.12	0.46	0.531	0.539	0.490	0.600
Naive Concat	0.56	0.61	0.64	0.09	0.64	0.35	0.69	0.60	0.58	0.39	0.19	0.49	0.45	0.21	0.48	0.6	0.39	0.28	0.27	0.41	0.525	0.497	0.522	0.551
MMX-Trailer-20	**0.62**	**0.69**	**0.71**	0.11	**0.71**	**0.53**	**0.73**	**0.62**	**0.64**	**0.51**	**0.34**	**0.56**	**0.60**	**0.45**	0.50	**0.64**	0.30	0.11	**0.13**	**0.55**	**0.597**	**0.583**	**0.554**	**0.697**

Table 3: Comparison of our proposed approach with existing methods for genre classification.

Method	no genres	no labels	$\overline{AU(PRC)}$	$AU(\overline{PRC})$	$AU(\overline{PRC})_w$
Random 9 Class	9	3	0.206	0.204	0.294
Random 20 Class	20	6	0.134	0.130	0.208
VLLF [24]	9	3	0.278	0.476	0.386
AV [14]	9	3	0.455	0.599	0.567
CTT-MMC [28]	9	3	0.646	0.742	0.724
Moviescope [7]	13	3	0.703	0.615	-
Proposed MMX-Trailer-20	20	6	0.456	0.589	0.583

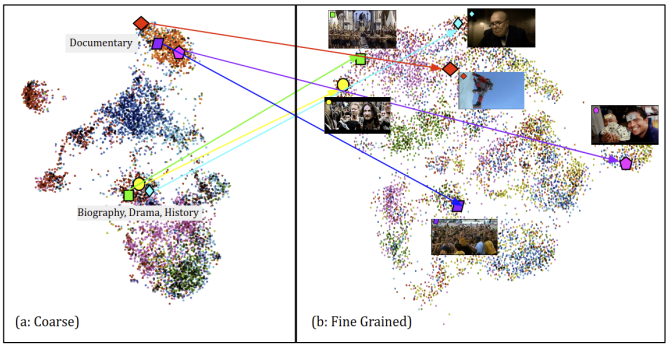


Fig. 3: Self-supervised Genre clustering via collaborative experts. (a) A T-SNE plot is showing the output of the coarse genre encoder network. (b) The fine-grained genre model encourages embeddings to cluster according to their multi-modal content.

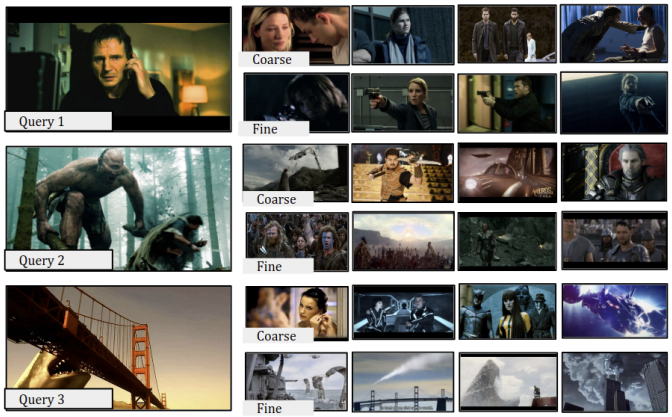


Fig. 4: Retrieval results obtained from the bottleneck embedding layer after training for coarse genre classification and after fine-tuning with the fine-grained self-supervised network. We show that the latter is much more effective at retrieving trailers that share multi-modal information yet have the same genre label.

towards a cluster featuring historical thrillers such as 'The Imitation Game'. We can also show illustrative retrieval results. For example, in Fig. 4 we offer how the fine-grained network finds movies with greater contextual similarity than the coarse encoder. For example, given the movie trailer, "Giant Shark vs Mega Octopus", the fine-grained network generates clusters of movies that feature sea monsters. You can view further examples in the supplementary materials.

4. CONCLUSION

Previous works have shown the effectiveness of convolutional neural networks and deep learning for genre classification. However, these methods do not address the unique semantic and contextual differences within these discreet labels. Using a collaboratively gated multi-modal network, we show that genre labels can be subdivided and extended using only visual and audio features, with applications in video recommendation, retrieval, and archiving.

5. REFERENCES

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. YouTube-8M: A Large-Scale Video Classification Benchmark. 2016.

[2] Rick Altman. 3 . A Semantic / Syntactic Approach to film genre. *Cinema Journal*, 23(3):6–18, 1984.

[3] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1437–1451, 2018.

[4] Aida Austin, Elliot Moore, Udit Gupta, and Parag Chordia. Characterization of movie genre based on music score. In *2010 IEEE International Conference on*

Acoustics, Speech and Signal Processing, pages 421–424. IEEE, 2010.

[5] Olfa Ben-Ahmed and Huet Benoit. Deep multimodal features for movie genre and interestingness prediction. In *International Conference on Content-Based Multimedia Indexing (CBMI)m*, 2018.

[6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[7] Paola Cascante-Bonilla, Kalpathy Sitaraman, Mengjia Luo, and Vicente Ordonez. Moviescope: Large-scale analysis of movies using multiple modalities. *arXiv preprint arXiv:1908.03180*, 2019.

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[9] Jianfeng Dong, Xirong Li, and Cees GM Snoek. Word2visualvec: Image and video to sentence matching by visual feature prediction. *arXiv preprint arXiv:1604.06838*, 2016.

[10] David Eigen, Marc'Aurelio Ranzato, and Ilya Sutskever. Learning factored representations in a deep mixture of experts. *arXiv preprint arXiv:1312.4314*, 2013.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR-16*, pages 770–778, 2016.

[12] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR-18*, pages 7132–7141, 2018.

[13] Hui-Yu Huang, Weir-Sheng Shih, and Wen-Hsing Hsu. A film classifier based on low-level visual features. In *2007 IEEE 9th Workshop on Multimedia Signal Processing*, pages 465–468. IEEE, 2007.

[14] Yin-Fu Huang and Shih-Hao Wang. Movie genre classification using svm with audio and video features. In *International Conference on Active Media Technology*, pages 1–10. Springer, 2012.

[15] Robert A Jacobs. Mi jordan, sj nowlan, and ge hinlon," adaptive mixtures of local experts. *Neural Computation*, 3(1):7It87, 1991.

[16] Sanjay K Jain and RS Jadon. Movies genres classifier using neural network. In *2009 24th International Symposium on Computer and Information Sciences*, pages 575–580. IEEE, 2009.

[17] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. *BMVC-19*, 2019.

[18] Gregory Luklow and Steven Ricci. The "audience" goes "public": Inter-textuality, genre, and the responsibilities of film literacy. (12):29, 1984.

[19] Antoine Miech, Jean-Baptiste Alayrac, Piotr Bojanowski, Ivan Laptev, and Josef Sivic. Learning from video and text via large-scale discriminative clustering. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5267–5276, 2017.

[20] Antoine Miech, Ivan Laptev, and Josef Sivic. Learnable pooling with Context Gating for video classification, jun 2017.

[21] Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516*, 2018.

[22] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 19–27, 2018.

[23] Steve Neale. Questions of Genre. *Film Genre Reader IV*, (July):178 – 202, 2012.

[24] Zeeshan Rasheed, Yaser Sheikh, and Mubarak Shah. On the use of computable features for film classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(1):52–64, 2005.

[25] Prashant Giridhar Shambharkar, MN Doja, Dhruv Chandel, Kartik Bansal, and Kunal Taneja. Multimodal kdk classifier for automatic classification of movie trailers. *IJRTE*, 2019.

[26] Prashant Giridhar Shambharkar, Pratyush Thakur, Shaikh Imadoddin, Shantanu Chauhan, and MN Doja. Genre classification of movie trailers using 3d convolutional neural networks. *ICICCS 2020*, 2020.

[27] Xin Wang, Fisher Yu, Lisa Dunlap, Yi-An Ma, Ruth Wang, Azalia Mirhoseini, Trevor Darrell, and Joseph E Gonzalez. Deep mixture of experts via shallow embedding. In *Uncertainty in Artificial Intelligence*, pages 552–562. PMLR, 2020.

[28] Jônatas Wehrmann and Rodrigo C Barros. Movie genre classification: A multi-label approach based on convolutions through time. *Applied Soft Computing*, 61:973–982, 2017.

[29] Jonatas Wehrmann, Rodrigo C Barros, Gabriel S Simoes, Thomas S Paula, and Duncan D Ruiz. (deep) learning from frames. In *Intelligent Systems (BRACIS), 2016 5th Brazilian Conference on*, pages 1–6. IEEE, 2016.

[30] Rui Wei Zhao, Jianguo Li, Yurong Chen, Jia Ming Liu, Yu Gang Jiang, and Xiangyang Xue. Regional Gating Neural Networks for Multi-label Image Classification. *British Machine Vision Conference 2016, BMVC 2016*, 2016-Septe:72.1–72.12, 2016.

[31] Howard Zhou, Tucker Hermans, Asmita V Karandikar, and James M Rehg. Movie genre classification via scene categorization. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 747–750, 2010.