Storify: Music Suggestion Model for Social Media Stories

Rahul Ajith, Sanjana IIITD, Hardi Parikh, Vishnu Mothukuri, Shlok Mehroliya, Om Mehroliya

PROBLEM FORMULATION

In the digital storytelling realm, integrating music into social media stories significantly boosts the emotional and aesthetic appeal of user-generated content across platforms such as Instagram, Facebook, and WhatsApp. Despite its crucial role in enhancing narratives, users often encounter difficulties in selecting music that harmoniously complements their visual stories, directly impacting viewer engagement and satisfaction.

Addressing this, the Storify model emerges as a solution aimed at automating the selection of background music by meticulously analyzing the visual content. The challenge primarily lies in the users' struggle to find music that aligns with the mood and theme of their images, stemming from the intricate process of interpreting visual themes and matching them with appropriate musical genres.

To overcome this, Storify leverages advanced deep learning and natural language processing (NLP) techniques to scrutinize visual content and its accompanying captions, extracting pivotal themes and emotions. This thorough analysis guides the classification of suitable music genres and the selection of congruent music tracks, thereby simplifying the task of enriching social media stories with fitting background music and enhancing the overall storytelling experience.

LITERATURE REVIEW

In the realm of digital storytelling, particularly within the context of social media platforms such as Instagram, Facebook, and WhatsApp, the integration of music into visual narratives has been recognized as a pivotal enhancement to the emotional and aesthetic appeal of user-generated content. This literature review delves into the existing body of research pertinent to the Storify model—a sophisticated system designed to automate the selection of background music that aligns with the mood and theme of visual content, leveraging advanced deep learning (DL) and natural language processing (NLP) techniques for content analysis and music selection.

Markus Schedl's exploration in "**Deep Learning in Music Recommendation Systems**" sheds light on the utilization of DL techniques in music recommendation systems (MRS), with a particular emphasis on extracting latent factors from music items and learning sequential patterns from music playlists or listening sessions. This foundational work underscores the relevance of applying similar DL methodologies for the analysis of visual content and its accompanying captions, facilitating the automated classification of suitable music genres and the selection of congruent tracks within the Storify model. The emphasis on content-based filtering (CBF) and sequence-aware music recommendation illuminates the potential of **Convolutional Neural Networks (CNNs)** for extracting features that resonate with the themes and emotions depicted in visual narratives.

Further expanding on the intricacies of music recommendation, Yading Song, Simon Dixon, and Marcus Pearce's comprehensive survey, **"A Survey of Music Recommendation Systems and Future Perspectives**" examines the multifaceted approaches encompassing collaborative filtering (CF), content-based models (CBM), and user-centric strategies, including emotion-based and motivation-based models. Their analysis offers invaluable insights into tailoring music recommendations **to enhance viewer engagement and satisfaction**, pivotal to the objectives of the Storify model. The integration of CF and CBM to refine the quality of recommendations, coupled with the proposition of leveraging emotion-based and motivation-based models, presents a robust framework for Storify to ensure that the music selection profoundly resonates with the visual and emotional context of digital stories.

Markus Schedl's subsequent work, "**Integrating Music Content, Music Context, and User Context for Improved Music Retrieval and Recommendation**" proposes a holistic approach to music recommendation by intertwining music content, context, and user context. This perspective aligns with Storify's mission by advocating for a recommendation system that transcends mere audio features to include contextual information about the music, visual content, and user context. Schedl's emphasis on user-centric models and the exploration of adaptive playlist generation based on user context underscores **the potential of creating a more personalized and context-aware music recommendation system for digital storytelling.**

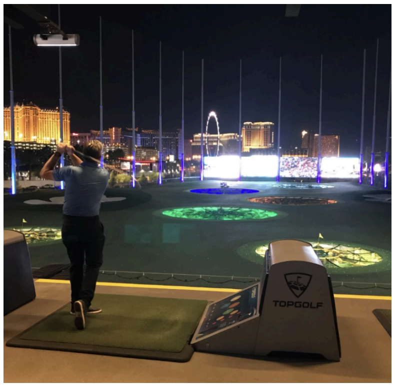
The synthesis of these scholarly contributions highlights the complexity and potential of employing deep learning and user-centric methodologies in developing automated systems for music recommendation in the context of digital storytelling. The Storify model, through its innovative approach to matching visual content with congruent musical tracks, stands at the forefront of enhancing user engagement and satisfaction in social media storytelling. These referenced works collectively provide a theoretical and practical foundation for advancing the Storify model's capabilities, emphasizing the importance of nuanced music preference understanding, deep learning's potential for content analysis, and the significance of considering both music and user context in personalized music recommendations.

RESULTS

Dataset

Our project utilizes the Flickr dataset, containing 5,000 images with rich metadata, including multiple captions per image. These captions provide insights into the scene, mood, and elements present, which are crucial for our analysis. We pre-processed this data by lowercasing the text, removing punctuation, and applying stemming to focus on the core semantics of the captions. This dataset was split into an 80% (4000 images) training set and a 20% (1000 images) validation set for the purposes of developing our model.

Below is a sample image from our Flickr dataset, representing a typical social media post, along with the actual captions.



Captions:

1. "Hitting the perfect drive into the Vegas night."
2. "Swinging into the weekend under the city lights."
3. "Topgolf views, top-notch vibes."
4. "Where the grass is green and the lights shine brighter."

Preprocessed Captions for LSTM Analysis (Stop words removed):

1. "Perfect drive Vegas night."
2. "Swinging weekend city lights."
3. "Topgolf views, top-notch vibes."
4. "Grass green, lights shine brighter."

LSTM Processed Result:
The LSTM model, after processing the preprocessed captions, determines the mood as lively and upbeat, considering the references to weekend activities and vibrant city life.

ResNet Processed Visual Features:
The ResNet model detects the golfing activity, night setting, and urban background, reinforcing the mood suggested by the LSTM analysis.

Music Recommendation Result:
Taking cues from both the LSTM-processed text and ResNet-analyzed visuals, the system, using Spotify API, recommends music that mirrors the energetic yet leisurely experience. A possible suggestion could be a playlist named

"Electro-Swing in the City," featuring upbeat electronic tracks with chill overtones that match the dual nature of activity and relaxation represented in the post.

User Experience Enhancement:
This integrated solution, combining both textual and visual analysis with music recommendations from Spotify, provides social media users with a synchronized and dynamic addition to their posts. By enhancing the visual narrative with a suitable musical backdrop, the platform significantly enriches the overall user experience.

Model Description:

Our baseline model comprises two main components: a Convolutional Neural Network (CNN) for visual analysis and a Long Short-Term Memory (LSTM) network for textual analysis.

Visual Component: We employed a pre-trained ResNet-50 CNN model to extract visual features from the images. The final layer's output serves as a compact representation of the visual content, capturing essential elements like color, texture, and objects within the image.

Textual Component: Parallel to the visual analysis, our LSTM network is tasked with deciphering the preprocessed captions. Leveraging advanced word embeddings, the network translates the textual data into a numerically encoded format, enabling the extraction of thematic and mood-related elements.

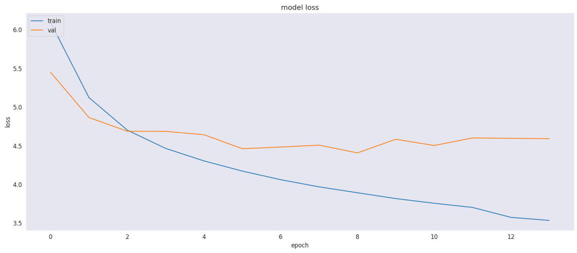
The fusion of the visual and textual interpretations is accomplished through a specialized layer that merges the outputs from both the CNN and the LSTM. This merged output is then processed through a series of dense neural layers, leading to a softmax classifier. This classifier assigns each pair of images and caption to specific mood categories, like reflective, joyful, or vibrant. The classified mood then interfaces with the Spotify API, which facilitates the selection of music tracks that complement the mood and setting of the social media content.

Model evaluation:

The model combines feature vectors from DenseNet201, processed through an LSTM for caption generation.

The model is making progress with both training accuracy and validation loss showing improvement over the epochs. An accuracy of around 27.47% by the 20th epoch, given the complexity of image captioning tasks, indicates that the model is learning from the training data. Only 1500 images were used in the training set and 100 images were used in testing due to complexity constraints.

The repetitive nature of generated captions ("startseq two people are playing in the air" for different images) suggests the model might be overfitting to the most common patterns seen during training or not adequately learning the diversity of the dataset (as indicated by a BLUE score of 0.1680). The model's validation loss is improving, which is a good sign.



Model Architecture Improvements:

Attention Mechanism: Incorporating an attention mechanism can significantly improve performance by allowing the model to focus on relevant parts of the image at each step of the caption generation. This could help address the issue of generic captions by making the model's output more contextually relevant to the image content.

Experimenting with Different Architectures: While DenseNet201 for feature extraction is a solid choice, experimenting with other architectures for the LSTM decoder part or even trying transformer-based models could yield improvements.

Expanding the training dataset with more images and captions can help improve the model's ability to generate diverse and accurate captions.

Following this, the genre will be generated from the ‘mood’ of the image by the model. The model would utilize the image features and captions (if available) to obtain an appropriate music genre for an image. Based on this genre, the top tracks in each will be searched using the Spotify API.

CONTRIBUTION

All Group Members have contributed equally to the research, literature review as well as in the development of the Model to help analyze images to give contextually relevant songs in the background for uses dependent on user use case

REFERENCES

Markus Schedl. "Deep Learning in Music Recommendation Systems." Frontiers in Applied Mathematics and Statistics.

Yading Song, Simon Dixon, and Marcus Pearce. "A Survey of Music Recommendation Systems and Future Perspectives." International Symposium on Computer Music Modelling and Retrieval (CMMR) 2012.

Markus Schedl. "Integrating Music Content, Music Context, and User Context for Improved Music Retrieval and Recommendation." Proceedings of the MoMM2013 Conference.