International Journal of Innovative Research in Engineering
Volume 6, Issue 2(March-Aril 2025), PP: 27-33.
www.theijire.com
ISSN No: 2582-8746

Deep Meme: Automated Image-Text Meme Production Via Convolutional Neural Networks

Moni Chaurasiya[1], Sanchi Deshmukh[2], Kaikashan Siddavatam[3], Sarita Kori[4], Nikki Shukla[5]
[1,2,4,5] *University of Mumbai, Maharashtra, India.*
[3] *Professor, University of Mumbai, Maharashtra, India*

How to cite this paper:
Moni Chaurasiya[1], Sanchi Deshmukh[2], Kaikashan Siddavatam[3], Sarita Kori[4], Nikki Shukla[5]," Deep Meme: Automated Image-Text Meme Production Via Convolutional Neural Networks", IJIRE-V6I02-27-33.

Abstract: *In this paper, we present DeepMeme, a system that uses convolutional neural networks (CNNs) and sophisticated natural language processing algorithms to automatically generate image-text memes. In contrast to conventional methods, DeepMeme incorporates a multi-modal attention mechanism that enables the text generation process to concentrate on an image's most prominent visual elements. To extract reliable picture embeddings, we specifically use a ResNet-based architecture that has been pre-trained on a sizable image dataset. After being refined on a carefully selected dataset of meme captions, these embeddings are subsequently input into a transformer-based language model, which allows for the creation of contextually appropriate and maybe amusing text. The method uses a new loss function that promotes stylistic consistency with well-known memes as well as semantic coherence between the image and text.*

I.INTRODUCTION

In the digital age, memes have emerged as a prevalent form of communication, encapsulating humor, cultural nuances, and shared experiences in easily digestible formats. Defined by Richard Dawkins as an idea, behavior, or style that spreads within a culture, memes thrive on their ability to convey complex emotions succinctly through language and imagery. Their ubiquitous presence across social media platforms underscores their significance as a modern-day lingua franca, enabling users to express thoughts, feelings, and opinions rapidly and effectively.

The dynamic nature of memes presents both opportunities and challenges for automated generation. Creating a meme that resonates requires not only an understanding of visual elements but also the ability to generate contextually relevant and humorous text. Traditional methods of meme creation are largely manual, relying on human creativity and cultural insight to blend images with captions that elicit the desired reaction. However, as artificial intelligence and deep learning technologies advance, there is growing interest in automating this creative process to meet the high demand for fresh and engaging content.

Existing approaches to automated meme generation often employ encoder-decoder architectures, utilizing convolutional neural networks (CNNs) for image feature extraction and recurrent neural networks (RNNs) or long short-term memory networks (LSTMs) for text synthesis. While these methods have demonstrated preliminary success, they frequently struggle with maintaining the subtle balance between semantic relevance and comedic effect. Additionally, ensuring that the generated captions align stylistically with popular meme formats remains a persistent challenge.

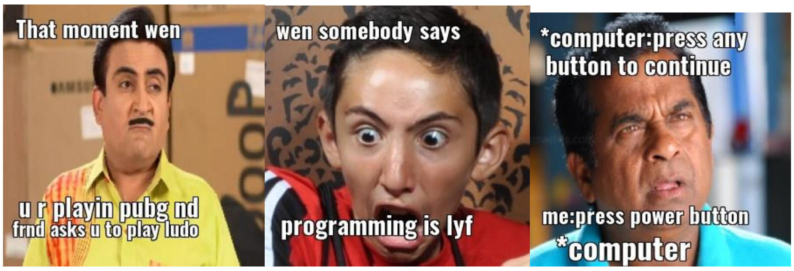


Figure I: Images Generated by DeepMeme

To address these limitations, we introduce **DeepMeme,** an advanced system designed to automatically generate image-text memes with enhanced coherence and humor. DeepMeme leverages a ResNet-based architecture pre-trained on extensive image datasets to extract robust visual embeddings. These embeddings are then refined through a multi-modal attention mechanism, allowing the system to focus on the most salient features of the image that contribute to effective captioning. The refined image representations are subsequently fed into a transformer-based language model, renowned for its ability to generate contextually appropriate and fluent text. A novel aspect of DeepMeme is the implementation of a specialized loss function that not only promotes stylistic consistency with established meme formats but also ensures semantic coherence between the image and the generated text. This dual-focus approach facilitates the creation of memes that are both visually and contextually aligned, enhancing their potential to engage and amuse audiences.

II. RELATED WORK

The field of automated meme generation is a rapidly developing field that combines computer vision and natural language processing (NLP) techniques. Peirson et al. [13] investigated meme creation using encoder-decoder architectures with Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks. These models approached meme generation as an image captioning task, generating textual content based on input images, but they frequently struggled to produce humor and contextually relevant captions, limiting their effectiveness in real-world meme creation.

With advances in deep learning, newer approaches have shifted towards transformer-based architectures and multi-modal learning, which improve text coherence and humor detection, making meme generation more engaging. Despite these advancements, there are still difficulties in aligning visual elements with text-based humor and making sure that they are consistent with The Synthetic Minority Over-sampling Technique (SMOTE) is a more effective method because it creates synthetic data points for minority classes, which improves model generalization and ensures a diverse representation. Attention-based techniques have been proposed in recent studies to refine image-text alignment, leading to more contextually appropriate meme captions. Class imbalance is a significant challenge in meme dataset creation, where specific meme templates or styles dominate, leading to biased model outputs. To address this issue, data balancing techniques like undersampling, which increases the presence of underrepresented classes but frequently involves duplicating existing samples, which may not contribute to model generalization.

In order to train reliable meme generation models, image augmentation is essential. While maintaining key visual properties, methods including rotation, flipping, cropping, scaling, color jittering, and Gaussian noise addition aid in adding unpredictability to training data. Augmentation approaches aid in the expansion of datasets, the reduction of overfitting, and the reinforcement of models against variations in meme formats in the context of meme production. Furthermore, by guaranteeing improved generalization in vision-based tasks, adaptive augmentation strategies—where changes are applied dynamically based on dataset characteristics—further improve performance.

Convolutional Neural Networks (CNNs) have been crucial in computer vision, allowing effective feature extraction and classification from images. While traditional CNN architectures like LeNet, AlexNet, and VGGNet established the groundwork for deep learning in image processing, more recent models like ResNet introduced residual learning to solve the vanishing gradient problem, allowing for deeper networks with improved performance. CNNs are also essential in meme generation, as they extract meaningful visual features from meme templates. For instance, pretrained ResNet models have been used to encode image embedding's that capture the essence of a meme, which are then integrated into NLP models allowing multi-modal learning that bridges the gap between image-based humor and textual content.

By eliminating the need for manual involvement, Automated Machine Learning (AutoML) has become a potent tool for streamlining machine learning procedures. By automating processes including feature selection, model selection, neural architecture search (NAS), and hyper parameter tuning, autoML improves the efficiency of machine learning operations. By boosting inference performance, increasing training efficiency, and fine-tuning hyper parameters, AutoML improves CNN and transformer-based models in meme production. Furthermore, AutoML-powered ensemble approaches can combine several models to enhance the quality of meme captions, guaranteeing that produced memes have stylistic coherence and contextual relevance.

All things considered, automated meme production has greatly improved with the use of sophisticated deep learning techniques, data balancing tactics, picture augmentation, CNN-based feature extraction, and AutoML-driven optimization. Even though there are still issues with text-image alignment, comedy recognition, and stylistic consistency, research is still being done to improve these models and open the door to more complex and interesting meme generating systems.

III.DEEPMEME GENERATOR APPROACH

3.1 Overview of the System

DeepMeme is a cutting-edge web program that uses user-provided photos to automatically create memes. By enabling users to upload a clear image, the application streamlines the process of creating memes by analyzing it to identify any facial expressions. DeepMeme uses this analysis to create humorous and contextually appropriate captions that blend in perfectly with the original image. The technology makes it simple for users to download the generated memes, which speeds up distribution on several social media platforms. This automated method uses cutting-edge machine learning techniques to reproduce the creativity that has historically depended on human intelligence, thereby meeting the growing demand for new and captivating meme content.

3.2 Methods of Data Balancing

3.2.1 The Value of Equitable Information

Training successful machine learning models requires balanced datasets to ensure fair representation of all classes. In cases of data imbalance, where some classes have significantly more samples than others, models tend to develop biases toward majority classes. For instance, if Class A contains 1,000 samples while Class B has only 300, the model may overfit to Class A and underperform on Class B. This imbalance weakens the model's generalization ability, particularly in complex tasks such as facial expression recognition in meme generation. To mitigate this issue, DeepMeme relies on effective data balancing techniques.

3.2.2 Excessive and Insufficient Sampling

Oversampling is DeepMeme's main tactic for addressing class imbalance. Using methods such as the Synthetic Minority Oversampling Technique (SMOTE), which creates synthetic instances by interpolating between preexisting data points, this strategy expands the number of samples in the minority class. Oversampling maintains the integrity of majority class data while improving minority class representation, ensuring the model generalizes well across all classes.

To enhance meme-based facial expression detection, DeepMeme used SMOTE to balance publically accessible datasets like AffectNet and FER-2013. By ensuring that the model learns from representative and diverse data without favoring the majority class, this technique increases the model's accuracy and robustness in identifying facial expressions in a variety of settings.

3.3 Augmentation of Images

One essential method used in DeepMeme to artificially increase the amount and variety of the training dataset is image augmentation. This procedure entails transforming pre-existing photos using a variety of techniques, including cropping, scaling, brightness modifications, rotations, and flips both horizontally and vertically. Image augmentation improves the model's resilience by adding these variations, which helps it deal with real-world situations where images could appear in various viewpoints, lighting conditions, or orientations. In contrast to oversampling, which targets class imbalance explicitly, image augmentation enhances the dataset's general quality and variety. As a result of the model being exposed to a wider variety of visual inputs during training, DeepMeme is able to provide captions that are more accurate and flexible.

3.4ArchitectureoftheModel

3.4.1 CNNs, or convolutional neural networks-

The Convolutional Neural Network (CNN), a specific kind of artificial neural network made to process and evaluate visual data, is the brains behind DeepMeme's image processing powers. CNNs are perfect for jobs like facial expression detection, which is crucial for meme creation, because they can automatically identify patterns, forms, edges, and textures in images. To extract reliable visual embeddings from user-uploaded photographs, DeepMeme uses a CNN architecture based on ResNet that has been pre-trained on large image datasets. The vanishing gradient issue is successfully mitigated by ResNet's deep design, which is defined by residual connections and enables the network to learn complex characteristics with great precision. In order to create feature maps that capture different local and global patterns, the convolutional layers apply a sequence of learnable filters (kernels) throughout the image. DeepMeme can reliably categorize facial expressions because to this hierarchical feature extraction, which is essential for producing funny and contextually relevant captions.

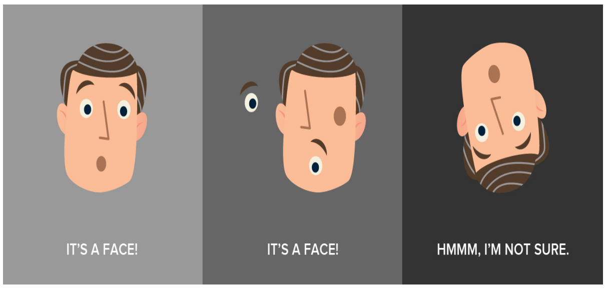


Figure II: Challenges in Face Detection Using CNNs

3.4.2. Machine Learning Automation (AutoML)-

To enhance DeepMeme's facial expression recognition capabilities, AutoML plays a crucial role. The approach involves training multiple models on diverse datasets to maximize accuracy and reliability in detecting expressions.

DeepMeme utilizes three key datasets for this purpose:

AffectNet: A large-scale dataset containing facial images labeled with various emotions.

FER-2013: A widely used dataset for facial expression recognition, featuring images classified into basic emotions.

CelebA: A dataset containing celebrity facial images with multiple attribute annotations, including expressions.

Weighted Ensemble Learning Strategy

The accuracy of models trained on these datasets varies depending on the detected expression. For instance:

A model trained on AffectNet achieves 75% accuracy for "Happy" expressions.

A model trained on FER-2013 achieves 39% accuracy for "Sad" expressions.

A model trained on CelebA achieves 60% accuracy for "Neutral" expressions.

To improve prediction accuracy, DeepMeme employs a weighted ensemble strategy, where models are assigned different weights based on their performance for specific expressions. For example, when predicting a "Sad" expression:

AffectNet (Weight: 0.8) contributes the most.

FER-2013 (Weight: 0.3) provides additional input.

CelebA (Weight: 0.1) contributes minimally.

3.5 Execution

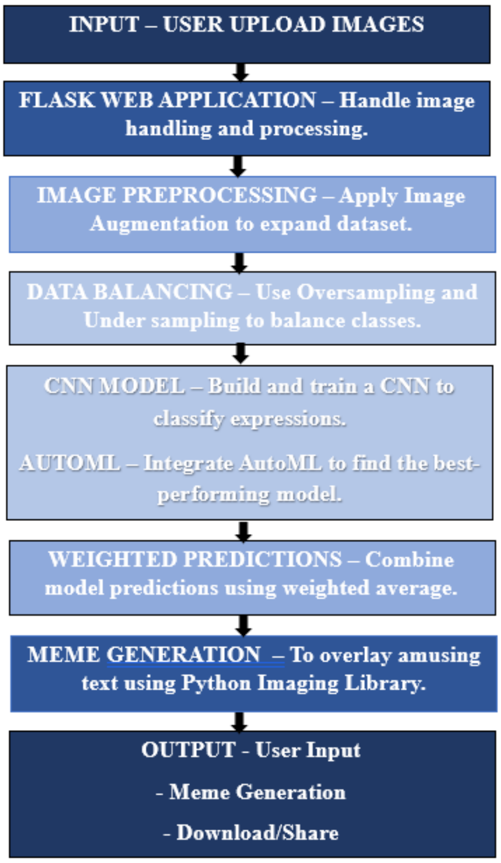
3.5.1 Framework for the Web

Flask, a lightweight and adaptable Python web framework, is used in the development of DeepMeme. Because Flask offers necessary features without the overhead of more complicated frameworks, it makes it easier to create web apps. This decision allows for the quick creation and simple integration of multiple components, including model inference, meme production, and image upload management. The simplicity of Flask guarantees that DeepMeme will continue to be scalable and maintained, enabling future feature additions and improvements with little difficulty.

IV.PROPOSED METHOD

4.1 Overview of the Approach

The suggested method, DeepMeme, is an online platform that automatically creates memes from user-uploaded photos by applying machine learning algorithms. Convolutional neural networks (CNNs), ensemble learning, data balance, and picture preparation are some of the parts of the system.


INPUT – USER UPLOAD IMAGES
FLASK WEB APPLICATION – Handle image handling and processing.
IMAGE PREPROCESSING – Apply Image Augmentation to expand dataset.
DATA BALANCING – Use Oversampling and Under sampling to balance classes.
CNN MODEL – Build and train a CNN to classify expressions.
AUTOML – Integrate AutoML to find the best-performing model.
WEIGHTED PREDICTIONS – Combine model predictions using weighted average.
MEME GENERATION – To overlay amusing text using Python Imaging Library.
OUTPUT - User Input
- Meme Generation
- Download/Share

Fiowchart I: DeepMeme FlowChart

Step 1: Preparing the image: Image preprocessing, the initial stage of the DeepMeme methodology, entails using a variety of methods to improve the input photos' diversity and quality. This stage consists of:
Image Augmentation: To expand the amount and diversity of the dataset, the system randomly alters the input photos through rotation, flipping, scaling, and cropping.
Data Balancing: To balance the dataset and avoid class imbalance, the system employs strategies like oversampling.
Step 2: Convolutional Neural Networks (CNNs): A CNN architecture, which is intended to extract features and patterns from the preprocessed images, is then fed the images. Multiple convolutional and pooling layers make up the CNN, which is followed by fully linked layers. A collection of feature maps that depict the input image are the CNN's output.
Step 3: Group Education: After that, an ensemble learning module processes the CNN-generated feature maps, combining the predictions of several models that were trained on various datasets. The system's accuracy and resilience are enhanced using the ensemble learning technique.
Step 4: Creation of Memes: Meme production, the last phase in the DeepMeme technique, is superimposing text on the input image to produce a hilarious and contextually appropriate description. The system creates a single output by combining the predictions of several models using a weighted ensemble approach.

V.RESULT

5.1 Datasets

We start by describing the structure of our dataset utilized in this study in order to give a thorough grasp of the basis of our model. There are over 700 image-text pairings in the collection, each of which represents a distinct combination of textual and visual information. A facial expression label and an accompanying story are two essential elements that are carefully applied to each data instance in the dataset. The seven main categories of face emotions—happy, sad, angry, afraid, shocked, disgusted, and neutral—are represented by these names. Interestingly, there are about 80 to 100 examples in each of these groups, indicating a very balanced sample distribution. By guaranteeing that the model is exposed to a wide variety of emotional expressions throughout training, this balanced representation improves the model's capacity for effective generalization.

There are two fundamental components to every data instance in the dataset:

Label for Facial Expression: This denotes the emotional state depicted in the related picture. Through supervised learning, the model is able to understand the subtleties of various emotions by using the label as the ground truth for its predictions.

Caption Text: In line with the recognized facial expression, this element offers a written explanation. In addition to providing further verbal clues that enhance the visual information, the caption depicts the emotional context. These two components work together to produce a rich multimodal dataset that makes it easier to build models that can comprehend emotional cues in both text and images.

The dataset is publicly accessible via the following link: https://www.kaggle.com/datasets/monichaurasiya/deepmeme-datasets/data

Table I: Sample datasets

Caption_ID	Facial_Expression	Caption_Text
1	Happy	The moment we know | our roommate bought| a bike
2	Happy	The moment wen u write |an error free code
3	Fear	The moment wen u | get caught while | cheating in an exam
4	Fear	When you hear the phrase | "we need to talk"
5	Angry	?When ur sibling | steals ur food
6	Angry	That moment wen ur | frnd gets credit | for ur hard work
7	Disgust	The moment ur | frnd uploads | life-changing quotations
8	Disgust	When your friend | won?t stop talking | about their diet
9	Sad	When you find out your | favorite show is canceled
10	Sad	Wen u prepare all night | and exam is | called off due to rain
11	Sad	Wen u have |1% charge left on mobile!
12	Surprise	When your college decides|to conduct online exams
13	Surprise	The moment your frnd | uploads life-changing|quotations
14	Surprise	Wen u see ur ever|late frnd come on time.
15	Happyy	When you see | your favorite | childhood movie

5.2 Accuracy of DeepMeme

Our system's accuracy evaluation, which is based on the creation of emotions from textual input, shows encouraging outcomes. The system's total accuracy across the four main emotion categories—Happy, Sad, Angry, and Fear—was 65.71%, as indicated in Table 1. With 15 out of 20 test instances created correctly, the Happy emotion category had the highest accuracy of 75% among them. With the lowest accuracy of 50%, the angry category suggested that it could be difficult to capture its complex facial expressions. At 60% accuracy each, the system's performance for the emotions of fear and sadness stayed modest. These findings both demonstrate how well the system produces acceptable emotional displays and point out areas for development, especially in recognizing minute differences in negative emotions. To improve overall accuracy, future improvements might entail fine-tuning the model's training with a wider variety of datasets and advancing facial expression synthesis methods.

Table II: Accuracy of our Deep Meme Generator

Emotion Category	Total Test Cases	Correctly Generated	Incorrectly Generated	Accuracy (%)
Happy	20	15	05	75%
Sad	15	09	06	60%
Angry	15	10	05	50%
Fear	20	12	08	60%
Overall	70	46	24	65.71%

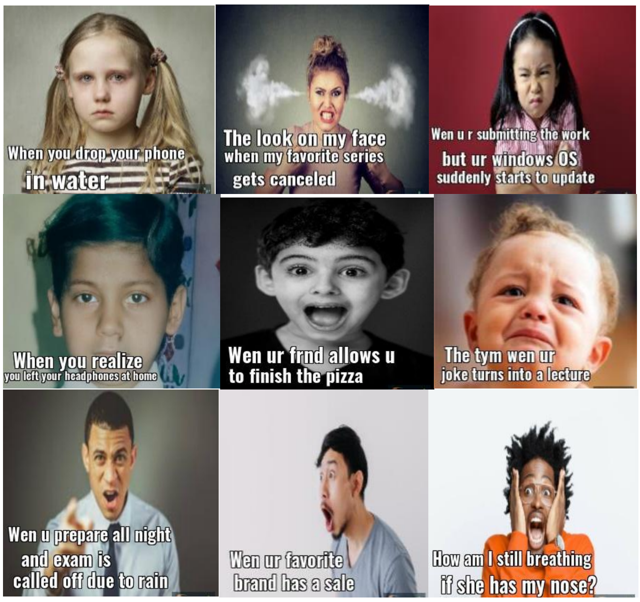


Figure 3: Memes generated from DeepMeme

VI.CONCLUSION

DeepMeme is revolutionizing the creation of memes by using machine learning to automatically create captions and identify facial emotions. Fundamentally, DeepMeme uses a weighted ensemble approach in conjunction with a CNN-based ResNet model to guarantee high expression prediction accuracy. It can identify a broad spectrum of emotions and produce contextually relevant meme captions because it has been trained on a variety of datasets, including AffectNet, FER-2013, and CelebA. Techniques like picture augmentation and the Synthetic Minority Over-sampling Technique (SMOTE) are used to further improve the model's robustness while guaranteeing a diverse and balanced dataset.

A scalable and effective environment for user interactions is provided by the Flask-built backend architecture, which makes it possible to create memes with little manual input. DeepMeme streamlines content creation and increases interaction on social media platforms by automating the whole meme generation process. In addition to streamlining the procedure, the combination of AI-powered facial recognition and captioning enables more expressive and customized memes.

Future developments for DeepMeme could include adding real-time user customization options, improving the model continuously for increased accuracy and adaptability, and extending its facial expression recognition capabilities to capture more complex emotions. In order to enable the system to produce even more contextually sensitive captions, future advancements might also concentrate on using deep learning approaches for text generation. DeepMeme is at the vanguard of the ongoing evolution of AI-powered content creation, revolutionizing the development and dissemination of memes in the digital era.

References

1. *PavelBerkhin.2006. ASurveyofClusteringDataMiningTechniques. InGrouping Multidimensional Data.*
2. *Florian Colombo, Alexander Seeholzer, and Wulfram Gerstner. 2017. Deep artificial composer: A creative neural network model for automated melody generation. In International Conference on Evolutionary and Biologically Inspired Music and Art. 81–96.*
3. *Shifman L. Memes in a digital world: Reconciling with a conceptual troublemaker. Journal of computer mediated communication. 2013;18(3):362-377. doi:10.1 111/jcc4.12013*
4. *Gal N, Shifman L, Kampf Z. "it gets better": Internet memes and the construction of collective identity. New media & society. 2016;18(8):1698-1714,.*

5. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, Show and tell: A neural image caption generator, 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
6. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet classification with deep con volutional neural networks. Proceedings of the 25th International Conference on Neural Information Pro cessing Systems- Volume 1 (NIPS'12)
7. Xu, K. et al. Show, attend and tell: Neural image caption generation with visual attention. Proc. Interna tional Conference on Learning Representations (2015).
8. Mun, J., Cho, M., Han, B.: Text-guided attention model for image captioning. AAAI (2016).
9. Ferrara E, JafariAsbagh M, Varol O, Qazvinian V, Menczer F, Flammini A. Clustering memes in social media. In: 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASO.AM 2013). IEEE; 2013:548-555. doi:10.1145/249 2517.2492530
10. Merriam-webster.com. (2018). Definition of MEME. [online] Available at: https://www.merriam webster.com/dictionary/meme [Accessed 21 Mar. 2018].
11. GitHub. (2018). tensorflow/models. [online] Available at: https://github.com/tensorflow/models/tree /master/research/im2txt#model-overview [Accessed 21 Mar. 2018].
12. Tensorflow authors, Official 'Show and Tell: A neural image caption generator model' implementation, https://github.com/tensorflow/models
13. Peirson V and Tolunay2018. Abel L Peirson V and E Meltem Tolunay. 2018. Dank learning: Generating memes using deep neural networks. arXiv preprint arXiv:1806.04510
14. Karpathy and Fei-Fei2015. Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic align ments for generating image descriptions. In Proceedings of the IEEE conference on com puter vision and pattern recognition, pages 3128–3137
15. Wang and Wen2015. William Yang Wang and Miaomiao Wen. 2015. I can has cheezburger? a nonparanormal approach to combining textual and visual information for predicting and gener ating popular meme descriptions. In Proceed ings of the 2015 Conference of the North Amer ican Chapter of the Association for Computa tional Linguistics: Human Language Technolo gies, pages 355–365.
16. QuickMeme. 2016. Quick Meme Website. http://quickmeme.com/.