OxfordTVG-HIC: Can Machine Make Humorous Captions from Images?

Runjia Li[1*] Shuyang Sun[1*] Mohamed Elhoseiny[2] Philip Torr[1]
[1]Torr Vision Group, University of Oxford [2]KAUST
https://torrvision.com/tvghic/

Abstract

This paper presents OxfordTVG-HIC (Humorous Image Captions), a large-scale dataset for humour generation and understanding. Humour is an abstract, subjective, and context-dependent cognitive construct involving several cognitive factors, making it a challenging task to generate and interpret. Hence, humour generation and understanding can serve as a new task for evaluating the ability of deep-learning methods to process abstract and subjective information. Due to the scarcity of data, humour-related generation tasks such as captioning remain under-explored. To address this gap, OxfordTVG-HIC offers approximately 2.9M image-text pairs with humour scores to train a generalizable humour captioning model. Contrary to existing captioning datasets, OxfordTVG-HIC features a wide range of emotional and semantic diversity resulting in out-of-context examples that are particularly conducive to generating humour. Moreover, OxfordTVG-HIC is curated devoid of offensive content. We also show how OxfordTVG-HIC can be leveraged for evaluating the humour of a generated text. Through explainability analysis of the trained models, we identify the visual and linguistic cues influential for evoking humour prediction (and generation). We observe qualitatively that these cues are aligned with the benign violation theory of humour in cognitive psychology.

1. Introduction

Humour has been recorded as a universal and high-level cognitive perception since Sumerians wrote down the first joke and remains a complicated concept due to its dependence on culture, visual and linguistic stimuli, as well as fundamental affective factors. Generating humorous content poses a significant challenge. As the arousal-reduction mechanism [31] suggests, the optimal level of novelty for humour should be neither too low nor too high, so that most audience can comprehend and appreciate it. Achieving this balance can be difficult. Another study from neuroscience [14] states: "successful jokes involve a cognitive juxtaposition of mental sets, followed by an affective feeling of amusement". Linguistically, semantic and phonological violations of mental sets interact to generate humour [32]. When multiple modalities such as vision and language are involved, the complexity of juxtaposition will significantly increase. Thus, Humour captioning, as a multimodal humour generation task, can be a useful task to investigate the upper limit of deep learning to handle high-level abstraction and creativity.

The definition of humour captioning is not clear and the task is under-explored. We argue that humour captioning should be distinguished from conventional image captioning [9, 45, 51], which mainly describes the objects or scenes in the image. The main goal of humour captioning should be to elicit a cognitive perception of humour

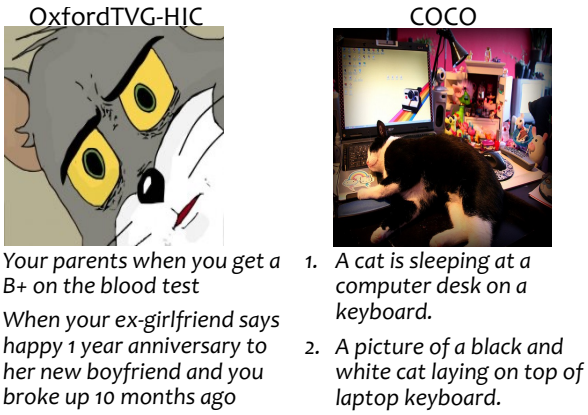


Figure 1: **Image-text samples from OxfordTVG-HIC and COCO** [9]. In OxfordTVG-HIC , the captions for a cat image do not describe the physical features of the cat, but rather the situations that could elicit the cat's facial expression. These situations create a humorous effect with the expression, as they are not offensive and violate the audience's everyday-life expectations (Benign violation theory [32]). On the other hand, the captions for a similar cat image in COCO [9] explicitly describe the facts in the image.

*These authors contributed equally to this work. Correspondence to Shuyang Sun (kevinsun@robots.ox.ac.uk).

Dataset	#Images	#Captions	#Captions /image	Dataset type
OxfordTVG-HIC	54k	2885k	**53.7**	Humour
COCO [9]	123k	617k	5	Object
CC3M [45]	3334k	3334k	1	Object
Flicker30k [51]	32k	159k	5	Object
ArtEmis [2]	80k	455k	5.7	Emotion
ArtELingo [35]	80k	1224k	15.3	Emotion

Table 1: **Image captioning dataset statistics comparison.** The average captions per image of OxfordTVG-HIC is significantly higher than other popular captioning datasets, and the total number of captions is larger than most of the datasets.

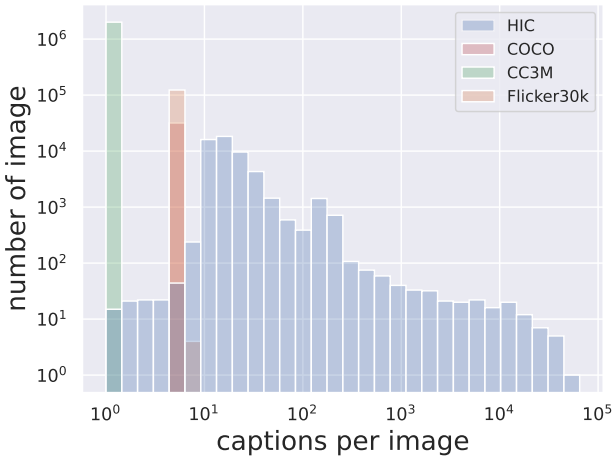


Figure 2: OxfordTVG-HIC are much greater than the other traditional image captioning datasets [9, 45, 51] in terms of the mean and variance for **the number of captions per image**.

rather than to provide factual information. This implies that a humorous sentence may not be relevant to the image content and still be a valid example. Hence, standard metrics [39, 12, 46, 25] that assess the quality of captions and conventional image captioning training framework may not be appropriate for humour captioning.

Humour captioning has received little attention. An initial attempt [41] used a relatively simple model [47] and a small-scale dataset and followed the conventional image captioning framework. Another recent image-text dataset [19] that aims to evaluate how well deep-learning models can comprehend humour has too few images (1.6k) with limited diversity to train a robust model.

To overcome the limitation faced by previous research, we introduce OxfordTVG-HIC (Humorous Image Captions), a large-scale image-text dataset for humour captioning. To our knowledge, it is the first dataset of this kind that contains 2.9 million image-text pairs with a humour score to measure their funniness. In OxfordTVG-HIC, each image has 53.7 captions on average, while the conventional image caption datasets (e.g., [9, 45, 51]) have less than 5 captions per image. Based on OxfordTVG-HIC, we develop humour generators that can automatically produce humorous captions given any images. the humour generators are trained on our proposed position-conditioned loss to address the issues of low diversity and humour caused by using cross-entropy loss on OxfordTVG-HIC. Specifically, since OxfordTVG-HIC is a dataset with high grammatical and semantic diversity, the cross-entropy loss tends to generate captions that are close to all ground truths given an input image, which is detrimental for training. The reason is that a generated sentence that is close to a set of distinct sentences is also far from those sentences. Therefore, the humour level and diversity are reduced with cross-entropy loss. Our position-conditioned loss solves this problem with promising results.

The qualitative nature of human thoughts makes it hard to evaluate and understand humour quantitatively. However, there are still some theories [32, 6, 37] proposed by psychologists to explain and measure humour. Drawing insights from one of the theories: Benign violation [32], we propose the benign violation humour score based on a humour classifier trained on OxfordTVG-HIC and a benign level classifier. Together with other linguistic metrics such as diversity and fluency, we benchmark the performance of our model on OxfordTVG-HIC.

Moreover, the learned ability to perceive humour as digital signals sheds light on the quantitative understanding and explainability of humour. From our observation and analysis, the model regards the abnormal and emotionally intense parts of images and sentences as the most important factors for creating humour. We hypothesise our observation is consistent with the Benign violation theory [32].

In summary, **our main contributions** are as follows:

1. We introduce *OxfordTVG-HIC*, a large-scale humour-oriented image-text dataset that addresses the lack of data in image-text-based humour generation and detection.

2. We show the diversity and richness of *OxfordTVG-HIC* and design the position-conditioned loss to better train a humour caption generator on the diverse data of *OxfordTVG-HIC*.

3. By explaining the learned humour models, we analyse the visual and linguistic stimuli that evoke humour and provide insights regarding humorous cues within the data. We further observe that the insights are aligned with the Benign violation theory [32].

2. Background and related work

Humour in deep learning. Humour is a complex cognitive phenomenon that requires a high level of abstraction to comprehend. Even humans may struggle to articulate the mechanisms and sources of humour when we encounter it. Having a sense of humour is so hard and valuable for humans that we have professional comedians to make jokes. Nowadays deep learning models have achieved remarkable performance in object-centric multi-modal tasks, but their ability to generalise to tasks involving abstract and high-level concepts like humour remains uncertain. Previous research has attempted to extract and analyse abstract concepts from various modalities, especially humour detection in texts [34, 7, 8, 50, 49, 3] and videos [27, 21, 40, 16], as well as text-based joke generation [28, 30, 17, 48].

Image captioning data. The goal of image captioning is to generate natural language descriptions for a given image. Most existing datasets [9, 45, 51] focus on the physical aspects of the scenes in an image and evaluate the descriptive skills of the models. Some recent datasets [2, 35] introduce the challenge of capturing the emotional effects of the visual stimulus and expressing them in natural language. However, to the best of our knowledge, no public large-scale dataset exists that specifically targets humour generation in captioning and examines how visual and linguistic cues trigger humour as a high-level cognitive phenomenon.

Humour-oriented caption generation. Despite its potential applications and benefits, humour-oriented captioning remains a relatively under-explored area. An initial attempt [41] adapted a simple model [47] to a small self-collected dataset and another dataset [19] was proposed to benchmark the machine's ability to comprehend humour. However, these datasets are limited in size and diversity and do not support the development of a robust model. Thus, a new large-scale humour captioning dataset is in need. Moreover, existing approaches have not tailored their image captioning modules to the specific features of humour-oriented captioning and have relied on conventional metrics that may not capture the humorous aspects of the captions. Though humour is complex to measure, psychologists have proposed several criteria such as Benign violation theory [32], Relief theory [6], and Superiority theory [37] to measure humour. Rather than using standard metrics [39, 46, 25, 12], redesigning evaluation metrics for humour captioning based on the psychological theories of humour should be more appropriate.

3. Humour Dataset: OxfordTVG-HIC

Current public datasets [41, 19] for humour-oriented image captioning contain too few image samples for the trained model to generalise well on unseen images. The lack of a large-scale available dataset hinders the development and generalisation of humour-centric multi-modal tasks such as humour captioning. Therefore, we present a large-scale humour captioning dataset: OxfordTVG-HIC, containing **2,885,326** image-text pairs in English and **53,728** unique images acquired from popular online websites and communities including Bokete Ogiri (Japanese) [52], ImgFlip (English) [20], and online image generators (English) [33]. Each image-text pair is annotated with a funny score in the form of votes from the website users, which could be crucial to quantitatively evaluate how much humour the visual and linguistic stimuli create jointly.

All raw captions in OxfordTVG-HIC are translated into English with DeepL translator [11] and cleaned with MangaOCR [5] and TextBlob [29] to ensure no linguistic hints appear in the images and captions with non-English words caused by homophonic translation are deleted so that culturally-exclusive contents and words will be removed.

Dataset	Anger	Digust	Fear	Joy	Neutral	Sad	Surprise
OxfordTVG-HIC	**0.136**	**0.138**	0.058	0.086	0.363	**0.078**	**0.145**
COCO[9]	0.036	0.076	**0.149**	0.048	**0.609**	0.016	0.065
CC3M[45]	0.061	0.074	0.074	**0.163**	0.533	0.059	0.035
Flicker30k[51]	0.049	0.136	0.129	0.074	0.549	0.023	0.037

Table 2: **Emotion analysis for captions in OxfordTVG-HIC and object-centric datasets.** The emotion richness of OxfordTVG-HIC is at most 24.6% greater than object-centric datasets.

Dataset	#Combination of Parts Of Speech		
	noun/verb	noun/verb/adj.	noun/verb/adj./conj.
OxfordTVG-HIC	**294,794**	**455,280**	**545,013**
COCO [9]	6,012	21,020	33,997
CC3M [45]	59,914	168,982	225,933
Flicker30k [51]	11,796	34,034	43,601
ArtEmis [2]	11,327	99,101	153,324
ArtELingo [35]	37,253	329,852	452,531

Table 3: **Number of grammar patterns of image captioning datasets.** The grammar pattern is defined by the combination of the relative position of different POS (Parts Of Speech) in sentences. OxfordTVG-HIC has a much larger number of grammar patterns than other popular captioning datasets.

3.1. Data statistics

Rather than attending to the physical content of an image, humour captioning focuses on a perceived aspect of the image and further develops a thought that relates to the objects, the objects' point of view, or the audience's comments on the image. The diverse form of humour indicates that the distribution of appropriate humour captions for an image can be much wider than object-centric captions. More

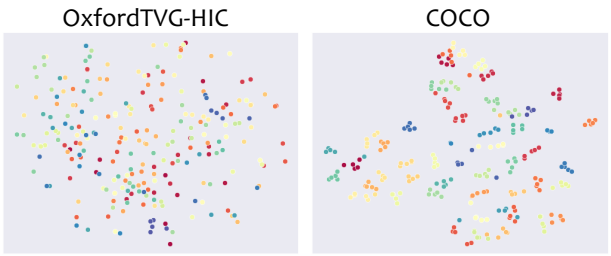


Figure 3: **t-SNE distribution of sampled captions.** Points with the same colour represent captions of the same image. Captions belonging to the same image are spread **distinct** in OxfordTVG-HIC, but they are clustered together in COCO [9].

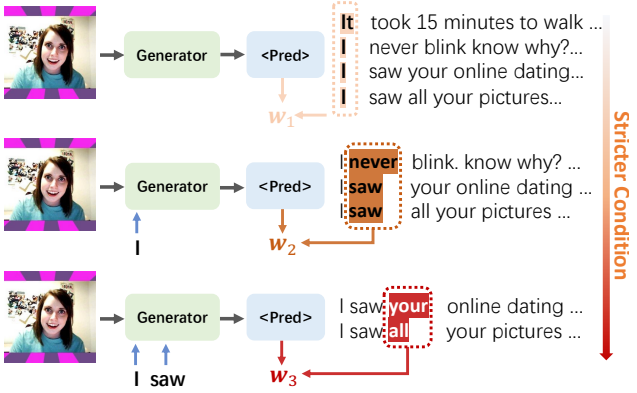


Figure 4: **Intuition of the position-conditioned loss: stricter previous-word conditions constrain the next-word distribution.** For the first few tokens, there are diverse ground truths that exclude one another (more diverse than the demo here). In this case, we don't want to penalize false positive predictions too much because they are potentially ground truths. As more previous tokens are determined, the diversity of plausible ground truths of the next token decreases. Now the false positive predictions become less likely to be ground truths. Thus we increase the false positive loss weight w_j to penalize those predictions.

specifically, the OxfordTVG-HIC stands its way out due to the fact that humour can take a variety of forms, or in other words, an infinite number of semantically distinct captions can stimulate humour in combination with the image. Another significant difference is that OxfordTVG-HIC has more emotional richness than object-oriented image captioning datasets, which requires more ability to extract abstract and psychologically high-level concepts from the input image. For this reason, as shown in Figure 2, OxfordTVG-HIC has 53.7 captions per image on average and most of the images from the OxfordTVG-HIC have 10 to 1,000 captions, which is significantly higher than that of popular image captioning datasets [9, 45, 51, 2, 35] as demonstrated in Table 1 and Figure 2. The high number of captions per image poses a greater challenge for captioning models to tackle than conventional image captioning.

3.2. Richness and diversity

Unlike object-centric image captioning datasets [9, 45, 51] that emphasise the physical relationship between objects in the image, OxfordTVG-HIC captures high-level mental perception in its captions. The subjectivity of OxfordTVG-HIC results in a rich and diverse set of emotions, semantics, and grammar.

Emotional richness. Though not an emotion itself, humour is built on emotions [14]. Quantitatively reflected in the dataset statistics, 63.7% captions in the OxfordTVG-HIC are detected to be emotional by DistilRoBERTa [15], compared to only 39.1% captions in COCO [9] are considered emotional as shown in Table 2. In more detail, when analysed apart from the image, the captions in OxfordTVG-HIC have more emotion in surprise, disgust, and anger than the object-centric dataset. Thus, although humour is generally considered a positive feeling, its emotional components could be non-positive.

Grammar pattern diversity. As shown in Table 3, if we consider the relative position of the nouns, verbs, adjectives, and conjunctions as grammar patterns, the captions in OxfordTVG-HIC have significantly more grammar types than other popular image captioning datasets [9, 45, 51, 2, 35]. The grammatical diversity results from the fact that humour can take an infinite number of forms due to its subjectivity.

Semantic diversity. The semantic diversity of humour captions is another characteristic of OxfordTVG-HIC. The captions for the same image in OxfordTVG-HIC may have different meanings when considered in isolation. To illustrate this, we used the CLIP [42] text encoder to obtain sentence-level semantic embeddings as shown in Figure 3. In COCO [9], the semantic embeddings of captions for the same image tend to form a cluster, indicating that they have similar meanings. However, in OxfordTVG-HIC , the semantic embeddings of captions for the same image are scattered across the embedding space. Moreover, we observed that some semantic embeddings of captions from different images are close to each other in the embedding space. This suggests that semantically similar captions can generate humour for different images.

3.3. Ethics

The raw dataset contained some offensive content such as profanity, sexual references, and hate speech based on race, gender, or belief. To ensure the quality and safety of OxfordTVG-HIC, we applied several filters to remove

the harmful content at different levels: image-level: NSFW offensive content detection [23], text-level: filtering captions with words from English profane word dictionary, and image-text-level: Villo [38] hateful memes detection. Finally, we discarded 10,352 images and 133,341 captions from the raw dataset.

4. Humour Captioning Baseline

Humour captioning is different from conventional image captioning [9, 45, 51], which aims to describe the objects or scenes in the image. The primary objective of humour captioning is to evoke the perception of humour rather than to convey factual information. This means that a humorous sentence can be valid even if it is not relevant to the physical image content, which introduces high grammatical and semantic diversity into the ground truths in OxfordTVG-HIC as shown in Figure 1. In this paper, the diversity in grammar and semantics is taken into account by our baseline humour generator.

4.1. Humour Generator

We experiment with two popular backbone architectures when designing humour generators trained on OxfordTVG-HIC: ClipCap [36], which combines CLIP [42] image encoder with a GPT-2 [43] text decoder connected by a mapping network; and the recent captioning model BLIP [24], which utilises the noisy web data by bootstrapping the captions. Those two backbones are trained with a position-conditioned loss we designed specifically for OxfordTVG-HIC.

4.2. Position-conditioned Loss.

Token-wise cross-entropy is a commonly used supervision loss in natural language processing tasks, and prominent language models such as BERT [13], GPT2 [43], and RoBERTa [26] have been trained using this method. While cross-entropy loss has demonstrated effectiveness in object-centric image captioning tasks, it may introduce confusion to models during training when presented with diverse captions for the same image that differ significantly in semantics and grammar patterns. This confusion is particularly troublesome when predicting the first few tokens of a sentence, as the initial label tokens for the same image may vary significantly in meaning as demonstrated in Figure 4. Training the model using cross-entropy loss under these circumstances leads to the model predicting the most similar token to all different label tokens to minimise the loss. However, if the predicted token is comparable to all label tokens that are distinct from one another, it will also be dissimilar to all of those label tokens, resulting in generated tokens that lack humour and are "neither fish nor fowl" as illustrated in Figure 5.

To tackle the problem, we propose a token-position-conditioned loss based on the assumption that the closer to the start of the caption, the less strict the condition is for the predicted token. For example, the condition of the first token on the image prompt should be weak because the same image prompt could lead to a variety of distinct first tokens. In more detail, we do not wish to penalise false negative prediction too much at the beginning of the sentence because the wrong prediction for one label could potentially be correct for another. For implementation, we designed a false positive loss weight function w_j based on token position (j is the j-th token position in the caption). Mathematically, the position-conditioned loss $\mathcal{L}_j$ can be formulated as the following:

$$\mathcal{L}_j = -\frac{1}{B}\sum_{i=1}^{B}\sum_{z=1}^{K}\begin{cases}\log(c_{ij}^z) & y_{ij}^z = 1\\ w_j \log(1-c_{ij}^z) & y_{ij}^z = 0\end{cases} \quad (1)$$

where $\mathbf{y}_{ij} = [y_{ij}^1, ..., y_{ij}^K] \in \{0,1\}^K$ is a one-hot vector for the label token and c_{ij}^z represents the z-th word probability of the predicted token $\mathbf{c}_{ij} \in [0,1]^K$, and K is the number of word tokens in the label space. B represents the data batch.

The false positive loss weight function w_j could take various forms. One general principle for choosing w_j is that the function needs to be monotonically increasing as the position index increases because we assume the condition gets stricter at the end of the caption. Next, we discuss choices for the false negative loss weight function w_j.

Linear. The most straightforward monotonically increasing function is the linear function. In this paper, we consider:

$$w_j = \gamma\frac{j}{M}, w_j \in \mathbb{R}^+ \quad (2)$$

where M denotes the number of tokens in the caption and γ is a hyperparameter that controls how drastically the false positive penalty should increase as the token position moves along the sentence.

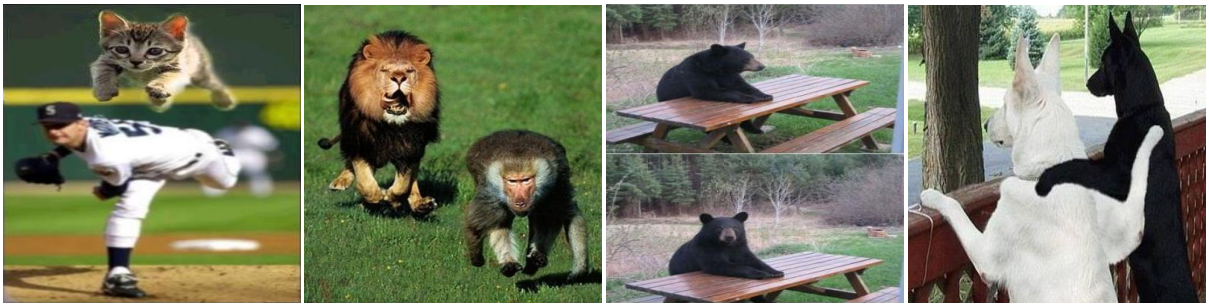
Gaussian. Gaussian distribution in the right half plane has properties of smooth change and monotonically decreasing at a fixed range. If we consider the transformed Gaussian function described as the following:

$$w_j = 1 - e^{-(\frac{j}{\beta})^2}, w_j \in \mathbb{R}^+ \quad (3)$$

where β controls the standard deviation, then the false negative loss weight function will be monotonically increasing in the right half plane and ranges from $[0, 1]$.

Sigmoid. The sigmoid function is another choice of monotonically increasing function with nice properties to smoothly converge at a constant value as the position index increases. We consider the sigmoid weight function as:

$$w_j = \frac{2}{1+e^{-\frac{j}{\alpha}}} - 1, w_j \in \mathbb{R}^+ \quad (4)$$



I'm going to go to the bathroom

Go Pikachu!

I'm going to go to the convenience store

Pool over! Give me your driver's license.

I'm going to give you a couple of examples.

Next interview, please

Hey, hey, what's wrong with you?

The cats all over the country have been deported as promised...

Figure 5: **Captions generated on unseen images by models that are trained on different losses**. Captions generated by the position-conditioned loss are rendered in blue, from which we can find that the position-conditioned loss solves the problem of limited diversity in cross-entropy. More examples will be shown in the appendix.

where α is a hyper-parameter that controls the scaling of the sigmoid function, or in other words, how fast the false negative loss weight increases. In our ablation experiment, we find $\alpha = 6$ gives the highest humour intensity.

5. Evaluation

Object-centric image captioning task uses linguistic evaluation metrics such as BLEU [39], ROUGE [25], METEOR [12], and CIDEr [46] to measure how close the predicted caption is to the ground truth objectively. And all previous work [41] on humorous caption generation used the objective metrics above to evaluate their humour generation model. However, we believe those linguistic metrics are inappropriate for the humour-centric image captioning task where ground truths per image could be distinct from one another in both semantics and grammar patterns.

To evaluate humour, we borrow insights from psychology and propose a new evaluation metric: Humour score and Benign score, which we believe serve better as a proxy for humour intensity evaluation to our task than conventional linguistic metrics mentioned above.

5.1. Humour metrics

Humour has been studied for decades, some famous explanations include the Benign violation theory [32], Superiority theory [37], and Relief theory [6]. The Benign violation theory [32] states that humour happens when a statement violates how the audience thought the world ought to be and the violation sounds benign to the audience. The subject of violation includes social norms, linguistic norms, moral norms, and self-dignity. An example of benign violations is shown in Figure 1. We believe designing an evaluation metric based on Benign violation theory [32] could allow us to measure how well a model can generate humour.

Humour Score. In this paper, the approach for evaluating the humour intensity in image-text pairs is based on a classification model trained on OxfordTVG-HIC. Positive samples are drawn from OxfordTVG-HIC, while negative samples of an image are comprised of semantically distinct captions from other images, randomly generated text from GPT-3 [4], and generated captions from captioning models trained on COCO [9]. The model is trained to distinguish between these positive and negative samples and ultimately to output the humour confidence given an image-text pair, which is positively correlated with the violation level. The classifier is composed of 3 components: ResNet50 [18] image encoder, GPT-2 [43] text encoder, and linear classification head to conduct a binary classification task. The classifier is trained on binary cross entropy loss and achieved an in-domain accuracy of 89% and out-domain accuracy of 77% (in-domain means seen images; out-domain means unseen images). The relative non-perfect performance of the evaluator implies the difficulty to comprehend and judge humour.

Benign Score. According to the benign violation theory, to generate humour, the benign level should be high and no offensive content to the audience should be included. We measure the benign value by Villo [38], a classification model trained on the Hateful Memes Challenge dataset [22]. Given an image-text pair, the model can tell whether the information the image-text pair conveys is hateful or offensive. The benign score of an image-text pair is the output probability of the Villo [38] model.

The Benign score and Humour scores are positively correlated with the humour intensity of the generated captions.

Model	Dataset	*w/* position-condition?	Humour score↑	Benign score [38]↑	Fluency (Parrot [10])↑
BLIP [24]	COCO [9]	×	0.416	**0.999**	**0.956**
	Ours	×	0.749	0.990	0.897
	Ours	✓	**0.828**	0.991	0.899
ClipCap [36]	COCO [9]	×	0.270	**1.000**	0.899
	Ours	×	0.650	0.935	**0.901**
	Ours	✓	**0.832**	0.984	0.831

Table 4: **Evaluation results of humour generators**. The humour scores of models trained on OxfordTVG-HIC and position-conditioned loss are the highest as expected while the benign and fluency scores of the models are comparable, which indicates models trained on OxfordTVG-HIC with position-conditioned loss generate the highest humour intensity.

5.2. Linguistic metrics

Besides the humour evaluation, fluency and diversity should also be evaluated to measure how well the generated sentences make sense and diversify linguistically.

Fluency score. We use Parrot [10], a T5-based language model, to measure the fluency score of the generated sentences. Parrot was originally designed to paraphrase a given sentence and also includes an evaluation model to measure the fluency of its paraphrased sentence. We argue that Parrot is suitable as a fluency evaluation model for the image captioning tasks because Parrot's original purpose aligns with our goal to evaluate generated sentences.

Diversity score. CLIP [42] is a powerful pre-trained model to extract semantic information from images and texts. We use CLIP [42] as a semantic diversity evaluator based on the cosine similarity of the extracted semantic features from generated captions. Mathematically, let p represent the caption semantic feature extracted by CLIP [42]. Then the diversity score can be described as :

$$diversity = \sqrt{1 - (\frac{1}{N}\sum_{m=1}^{N}\max_{n,n\neq m}\frac{p_m^T p_n}{|p_m||p_n|})^2} \quad (5)$$

6. Experiments

6.1. Humour caption generation

Implementation details. We train all models on OxfordTVG-HIC with COCO [9] pre-trained weights. The training process consists of 20 epochs with 500 warm-up steps for both cross-entropy and position-conditioned loss. We use the cosine learning rate decaying schedule with the initial learning rate of 10^{-5}.

Experimental results. As shown in Table 4, humour generators trained on OxfordTVG-HIC obtain at most 56.2% more in humour score than those trained on COCO [9], and humour generators trained with position-conditioned loss improve humour score by 18.2% compared to those trained with cross-entropy. The position-conditioned loss improves the diversity by 13% (as shown in Table 6) with comparable fluency.

Kernel function	Humour score↑	Benign score [38]↑	Fluency (Parrot [10])↑	Diversity (CLIP [42]))↑
sigmoid	**0.832**	**0.984**	0.831	**0.373**
linear	0.792	0.911	0.795	0.348
gaussian	0.826	0.906	**0.873**	0.362

Table 5: **Influence of kernel functions of position-conditioned loss on the performance of ClipCap** [36]. Experimentally, the sigmoid kernel function conveys the highest benign and humour scores, or equivalently, the highest humour intensity. Besides, models trained with the sigmoid kernel loss generate the most diverse captions.

α	Humour score↑	Benign score [38]↑	Fluency (Parrot [10])↑	Diversity (CLIP [42]))↑
/	0.650	0.935	**0.901**	0.254
2	0.818	0.909	0.881	0.358
4	0.823	0.911	0.889	0.362
6	**0.832**	**0.984**	0.831	0.373
8	0.822	0.905	0.795	**0.384**

Table 6: **Influence of the hyper-parameter α of the sigmoid kernel function defined in Eq. (4) on diversity and humour level of ClipCap** [36]. As α increases, the false positive penalty increases slowlier and more diversity is introduced because the supervision is less strict. Also, models trained on position-conditioned loss generally tend to have more diversity than cross-entropy (denoted by $\alpha =$"/").

6.2. Ablation study

Influence of different kernel functions. The sigmoid function of the position-conditioned loss conveys the highest benign and humour score as well as diversity as demonstrated in Table 5.

Influence of kernel function hyperparameters. The hyper-parameter α for the sigmoid kernel function of the position-conditioned loss controls how fast the false positive penalty increases with positions. The motivation to introduce the position-conditioned loss is to overcome the problem of limited diversity caused by cross-entropy and α plays an important role in diversity control. As demonstrated in Table 6, when α increases, the diversity of generated captions increases as well while the humour level seems without correlation.

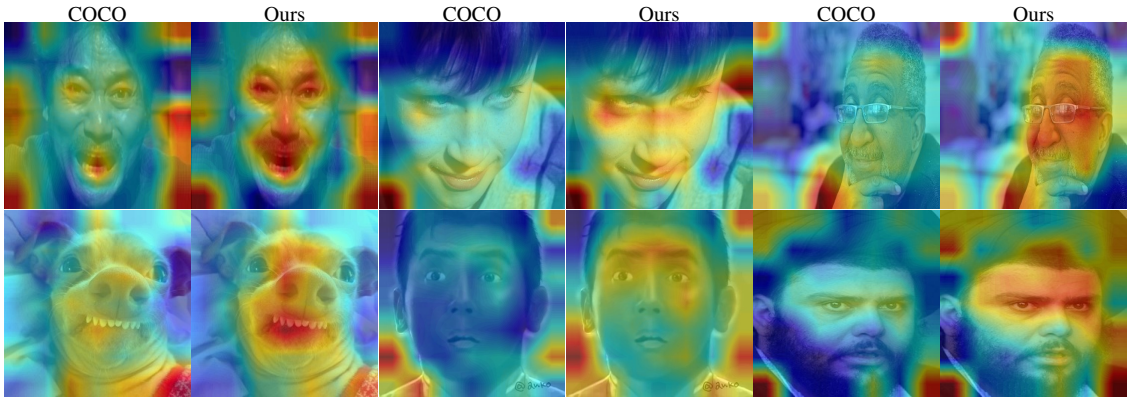


Figure 6: **Attention heatmap of the humour generator image encoder.** The generator tends to focus more on facial expressions, especially the "dramatic" parts which involve more emotions.

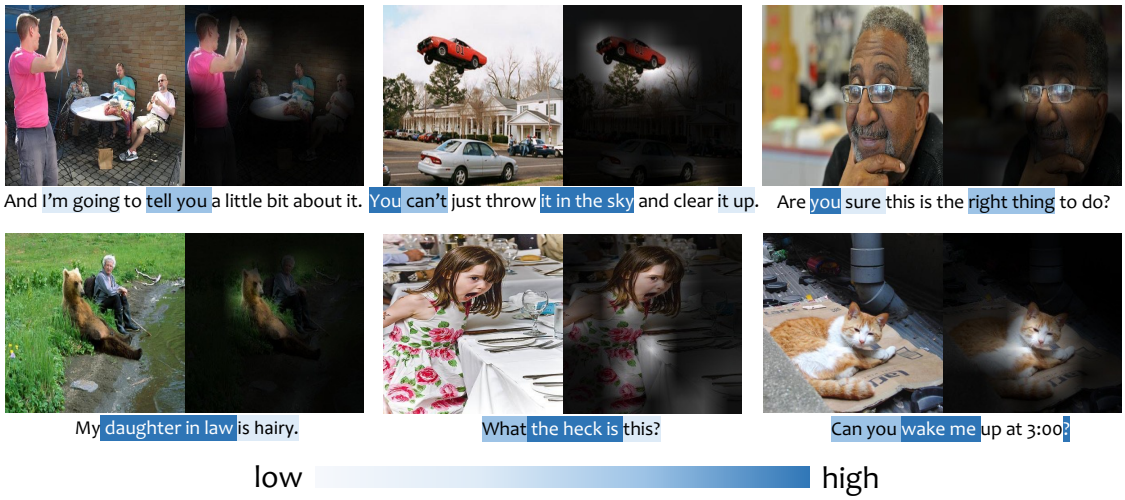


Figure 7: **Gradient visualization of image-text pairs from humour classifier.** The evaluator tends to focus on abnormal parts and facial expressions of the images which involve more emotion; More attention is paid to the pronouns in the text which induces engagement from the audience. Deeper colour on the texts means a larger contribution to humour.

6.3. Understanding humour

The humour generator and classifier can both provide insight into the quantitative understanding of humour, which is one of the major goals of this paper. In general, we aim to quantify what visual and linguistic clues the models use to generate and evaluate humour.

Generator attention visualisation. We visualise [1] the attention map of the image encoder and examine how it attends to different parts of the image to produce humorous captions. Figure 6 shows the comparison of the attention heatmaps between the image encoder trained on COCO [9] and OxfordTVG-HIC. The results indicate that the OxfordTVG-HIC-trained encoder pays more attention to affective features in the image. For example, it focuses more on the facial expressions of people than the COCO [9]-trained encoder, as they often convey rich emotions. Moreover, it highlights some unusual or exaggerated aspects of facial expressions, such as a dog's odd smile or a child's evil eyes. These features suggest that humour arises from some form of incongruity or violation of mental sets. This observation is consistent with the benign violation theory [32] of humour.

Evaluator gradient visualisation. To visualise the most contributive parts of the image and text to humour generation judged by the humour classifier, we apply Grad-CAM [44] on images and gradient magnitude analysis on text embedding space. Figure 7 demonstrates that the violation evaluator also focuses on facial expressions and incongruous elements in the image (*e.g.* "a cat's perplexed face" and "a car in the air"), similar to the image encoder of the generator. For the captions, pronouns seem to play a significant role as they involve the audience in the situa-

tion portrayed in the picture. This involvement and empathy may be the key to eliciting humour in the audience's mind.

7. Discussion and Future Works

Besides humour-oriented image captioning, OxfordTVG-HIC as an image-text dataset could be applied in any humour-oriented task involving visual and linguistic modalities. Potential directions include humorous image generation guided by texts and simultaneous humour-oriented image-text generation. Since OxfordTVG-HIC is collected from different cultural websites and communities, the explainability and understanding of the cultural and linguistic differences in humour could be explored with OxfordTVG-HIC. If similar video-based datasets are created, the humour generation and evaluation tasks can be extended to a more complicated and challenging level.

8. Conclusion

Humour is a unique and iconic characteristic of the human cognitive process. It reflects out-of-context and unexpected interpretations of information. In this work, we pave the way for achieving humour understanding and generation of artificially intelligent systems by (1) the release of the OxfordTVG-HIC dataset that for use in humour-oriented vision language generation and evaluation; and (2) a demonstration of humour generators that can create jokes given any images. The capacity to computationally deal with humorous attributes of text and images opens an exciting new direction in human-computer communication and interaction and unlocks a novel hallmark of cognition.

Acknowledgement. We would like to thank Ashkan Khakzar, Yangchen Pan, and Jindong Gu for their helpful suggestions and feedback on the paper. This work is supported by the UKRI grant: Turing AI Fellowship EP/W002981/1 and EPSRC/MURI grant: EP/N019474/1. We would also like to thank the Royal Academy of Engineering and FiveAI.

8.1. Limitations

As the first large-scale study of humour captioning, we acknowledge the following limitations of our work. (1) Our humour classifier is trained on positive samples from OxfordTVG-HIC, which may bias the classifier towards generators trained on the same dataset. (2) The humour classifier only correlates to the benign violation theory, not directly linked. A stronger connection to psychological theory is desirable. We leave this for future work. (3) Because the funny scores in OxfordTVG-HICare collected based on user votes from different humour-oriented communities, the funny scores are not entirely proportional to the true humour intensity, but only positively correlated. (4) Due to the fact that certain humorous captions are originally composed in Japanese and subsequently translated into English via DeepL [11], it is possible that the intended humour of the original Japanese captions may be diminished as a result of inaccuracies in translation

References

[1] Samira Abnar and Willem H. Zuidema. Quantifying attention flow in transformers. *CoRR*, abs/2005.00928, 2020.

[2] Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas Guibas. Artemis: Affective language for visual art. *CoRR*, abs/2101.07396, 2021.

[3] Issa Annamoradnejad. Colbert: Using BERT sentence embedding for humor detection. *CoRR*, abs/2004.12765, 2020.

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020.

[5] Maciej Budyś. Manga OCR. original-date: 2022-01-15T17:18:34Z.

[6] Moniek Buijzen and Patti M. Valkenburg. Developing a typology of humor in audiovisual media. *Media Psychology*, 6(2):147–167, 2004.

[7] Lei Chen and Chong Min Lee. Predicting audience's laughter during presentations using convolutional neural network. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 86–90, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics.

[8] Peng-Yu Chen and Von-Wun Soo. Humor recognition using deep learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 113–117, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[9] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server, 2015.

[10] Prithiviraj Damodaran. Parrot: Paraphrase generation for nlu., 2021.

[11] DeepLcom/deepl-python: Official python library for the DeepL language translation API.

[12] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[14] Vinod Goel and Raymond J. Dolan. The functional anatomy of humor: segregating cognitive and affective components. *Nat Neurosci*, 4(3):237–238, 2001.

[15] Jochen Hartmann. Emotion english distilroberta-base. https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/, 2022.

[16] Md Kamrul Hasan, Sangwu Lee, Wasifur Rahman, Amir Zadeh, Rada Mihalcea, Louis-Philippe Morency, and Ehsan Hoque. Humor knowledge enriched transformer for understanding multimodal humor. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12972–12980, May 2021.

[17] He He, Nanyun Peng, and Percy Liang. Pun generation with surprise. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1734–1744, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[19] Jack Hessel, Ana Marasović, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest, 2022.

[20] Imgflip - create and share awesome images.

[21] Yuta Kayatani, Zekun Yang, Mayu Otani, Noa Garcia, Chenhui Chu, Yuta Nakashima, and Haruo Takemura. The laughing machine: Predicting humor in video. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2072–2081, 2021.

[22] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *CoRR*, abs/2005.04790, 2020.

[23] Gant Laborde. Deep nn for nsfw detection.

[24] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[25] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[26] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[27] Zhisong Liu, Robin Courant, and Vicky Kalogeiton. Funnynet: Audiovisual learning of funny moments in videos. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 3308–3325, December 2022.

[28] Nikita Login, Alexander Baranov, and Pavel Braslavski. Jokingbird: Funny headline generation for news. In Evgeny Burnaev, Dmitry I. Ignatov, Sergei Ivanov, Michael Khachay, Olessia Koltsova, Andrei Kutuzov, Sergei O. Kuznetsov, Natalia Loukachevitch, Amedeo Napoli, Alexander Panchenko, Panos M. Pardalos, Jari Saramäki, Andrey V. Savchenko, Evgenii Tsymbalov, and Elena Tutubalina, editors, *Analysis of Images, Social Networks and Texts*, pages 97–109, Cham, 2022. Springer International Publishing.

[29] Steven Loria. TextBlob: Simplified text processing. original-date: 2013-06-30T18:29:18Z.

[30] Fuli Luo, Shunyao Li, Pengcheng Yang, Lei Li, Baobao Chang, Zhifang Sui, and Xu Sun. Pun-GAN: Generative adversarial network for pun generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3388–3393, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

[31] Colin Martindale. *Romantic progression: The psychology of literary history*. Harpercollins, 1975.

[32] A Peter McGraw and Caleb Warren. Benign violations: Making immoral behavior funny. *Psychological science*, 21(8):1141–1149, 2010.

[33] Meme generator.

[34] Rada Mihalcea and Carlo Strapparava. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 531–538, Vancouver, British Columbia, Canada, Oct. 2005. Association for Computational Linguistics.

[35] Youssef Mohamed, Mohamed Abdelfattah, Shyma Alhuwaider, Feifan Li, Xiangliang Zhang, Kenneth Ward Church, and Mohamed Elhoseiny. Artelingo: A million emotion annotations of wikiart with emphasis on diversity over language and culture. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

[36] Ron Mokady, Amir Hertz, and Amit H. Bermano. Clipcap: CLIP prefix for image captioning. *CoRR*, abs/2111.09734, 2021.

[37] John Morreall. *Taking Laughter Seriously*. State University of New York Press, 1983.

[38] Niklas Muennighoff. Vilio: State-of-the-art visio-linguistic models applied to hateful memes. *CoRR*, abs/2012.07788, 2020.

[39] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA, 2002. Association for Computational Linguistics.

[40] Badri N. Patro, Mayank Lunayach, Deepankar Srivastava, Sarvesh Sarvesh, Hunar Singh, and Vinay P. Namboodiri. Multimodal humor dataset: Predicting laughter tracks for sitcoms. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 576–585, 2021.

[41] Abel L. Peirson V and E. Meltem Tolunay. Dank learning: Generating memes using deep neural networks.

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021.

[43] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[44] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391, 2016.

[45] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[46] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. *CoRR*, abs/1411.5726, 2014.

[47] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555, 2014.

[48] Orion Weller, Nancy Fulda, and Kevin Seppi. Can humor prediction datasets be used for humor generation? humorous headline generation via style transfer. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 186–191, Online, July 2020. Association for Computational Linguistics.

[49] Orion Weller and Kevin Seppi. The rJokes dataset: a large scale humor collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6136–6141, Marseille, France, May 2020. European Language Resources Association.

[50] Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. Humor recognition and humor anchor extraction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2367–2376, Lisbon, Portugal, Sept. 2015. Association for Computational Linguistics.

[51] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[52] Bokete ogiri.