Title:
Do Androids Laugh at Electric Sheep? Humor "Understanding" Benchmarks from The New Yorker Caption Contest.

Link: https://aclanthology.org/2023.acl-long.41.pdf

Authors: Number of Citations from Google Scholar:

- Jack Hessel from The Allen Institute for AI: 29 citations
- Ana Marasović from the University of Utah: 20 citations
- Jena D. Hwang from The Allen Institute for AI: 38 citations
- Lillian Lee from Cornell University: 69 citations
- Jeff Da from the University of Washington: 12 citations
- Rowan Zellers from OpenAI: 30 citations
- Robert Mankoff from Air Mail and Cartoon Collections: 8 citations
- Yejin Choi from the Allen Institute for AI and the University of Washington: 351 citations
    - The author with the most citations

Problem Statement:

How well do large neural networks "understand" humor? The authors of this paper developed three tasks to challenge how well these AI/ML models understand humor and are of increasing complexity. These tasks are matching a joke to the New Yorker Caption Contest, identifying the winning caption for that cartoon, and explaining why that winning caption and image is funny.

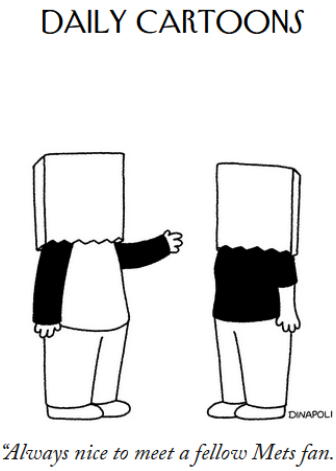

Figure 1: An example of a daily cartoon with the winning caption from the New Yorker
Source: https://www.newyorker.com/humor

Prior Works:

The New Yorker hosts a contest, where they provide a cartoon image with no caption. The audience (denoted as the "crowd by the authors of this research paper) then submits their funniest captions for that cartoon. The editors choose three finalists and then the audience votes on the three finalists and the finalist with the most votes in the chosen caption.

There are three great families of humor. Superiority theory is humor that involves the perception of superiority over something or someone. An example of this is a political cartoon that satirizes a politician, Relief theory is humor that serves as a release to some stress or tension. An example of this is a nervous laugh during a tense scene in a thriller movie. The Incongruity Theory is humor is the subversion of expectations, or when something unexpected happens. Puns are an example of this type of humor. Most of the humor from the New Yorker cartoons falls under the Incongruity Humor category.

There already exists developed methods for identifying features that determine the funniest caption of The New Yorker cartoons. These include "perplexity, match to image setting and uncanniness description, readability, proper nouns, overlap with WordNet's, …, "person" and "relative" synsets, lexical centrality among submissions and sentiment.

Ervin Tanczos, Robert Nowak, and Bob Mankoff created a ranking algorithm for a caption contest. However, it does not directly use the data from The New Yorker. Lalit Jain, Kevin Jamieson, Robert Mankoff, Robert Nowak, and Scott Sievert created a dataset for The New Yorker cartoons and captions that the authors of this paper utilized in their research.

There was past research that explored various related aspects of computational humor. There exists a paper that explored humor recognition in images, research that explores prediction when laughter will occur, studies of political cartoons, whether a sentence is humorous or not, and ways to automatically generate humor from specific contexts.

The researchers of this paper also utilized past research on qualitatively explaining humor. They specifically used a "qualitative exploration of (non-visual) joke explanations and mechanisms by which one can understand the humor from the text.

Unique Contributions:

The authors added two distinct features to the methods of identifying features for the funniest caption. The first feature is using new data to rank whether or not a caption matches with a cartoon. The second feature is evaluating on two different audience preferences, The New Yorker editors and the "crowd" (the readers of the New Yorker).

The paper states that the three tasks are covered in two methods, or what they call two settings. The first setting is called the "from pixels" setting. At test time, the models are given the cartoon images and must perform computer vision on those images. The only available information from the contest is the image of the cartoon. The second setting is called the "from description" setting. In this setting, the model is provided with human written annotations about the cartoon, simulating "access to a human-level computer-vision system".

The authors state that most of their contribution is within this "from description" setting. Interestingly, these annotations are mainly created by crowdworkers from Amazon Mechanical Turk. These human-written annotations contain the setting, the description, an explanation of the humorous irregularities, and several relevant Wikipedia links to each cartoon. These annotations are utilized in the "from description" or the training section of the "from pixels" setting.

The authors created several models to evaluate the three tasks. To evaluate the models, they first "split all 704 cartoons into 5 cross-validation splits" to ensure that the entire contents are split between the training splits and testing splits.

They explored two "from pixels" models: CLIP and OFA -> LM. The researchers also explored several "from description" models: T5-Large, T5-11B, GPT-3, GPT-3.5, and GPT-4. Additionally, two baselines were developed, a Random-guess baseline and a fine-tuned T5-11B baseline that was only given

the caption of the cartoon (but not the cartoon itself). The paper goes into more detail on each implementation of each model.

Evaluation:

For evaluation, the models are passed through three separate tests for each of the three tasks that the authors of this paper developed.

To test how well a caption matches a given cartoon, the model is given the cartoon and a multiple-choice question with five captions. One of these captions is the correct match for the given cartoon, while the incorrect captions are correct for cartoons from different contests. The model then must choose the option to be the correct caption. A correct instance would be if the model chooses the correct caption, and the incorrect instance otherwise

How well a model can identify the higher-rated, or funnier, caption? To evaluate this question, the model is given a finalist (higher rated) caption and a caption not selected as a finalist (lower rated). The model has to then choose which one of these captions is the better one for this cartoon. If the finalist caption is chosen, then it is a correct instance, otherwise, it is an incorrect instance.

		Matching	Quality Ranking	
		Accuracy (↑)	CrowdAcc (↑)	NYAcc (↑)
	Random	20.0	50.0	50.0
	Caption Only (T5-11B)	19.4	59.4	64.5
FP	CLIP ViT-L/14@336px (finetuned)	62.3	57.0	66.9
FP	↳ Zero-shot	↳ 56.6	↳ 55.8	↳ 56.8
FP	OFA-Huge → T5-Large	45.2	59.1	64.3
FP	OFA-Huge → T5-11B	51.8	60.3	65.0
FD	T5-Large	59.6	61.8	64.8
FD	T5-11B	70.8	62.3	65.6
FD	GPT3-175B (finetuned)	75.1	64.8	**69.8**
FD	↳ 5-shot	↳ 57.2	↳ 55.1	↳ 54.8
FD	↳ Zero-shot	↳ 51.6	↳ 56.2	↳ 55.6
FD	GPT 3.5 (5-shot)	63.8	55.6	55.2
FD	↳ Zero-shot+CoT	↳ 50.4	↳ 52.8	↳ 55.4
FD	GPT-4 (5-shot)	**84.5**	**73.3**	68.2
FD	↳ Zero-shot+CoT	↳ 81.9	↳ 66.2	↳ 64.3
	Human Estimate From Pixels (FP)	94.0	83.7	64.6

Figure 2: Prediction Results from the matching and quality ranking tasks.
(Source: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 688–714
July 9-14, 2023 ©2023 Association for Computational LinguisticsDo Androids Laugh at Electric Sheep? Humor "Understanding"
Benchmarks from The New Yorker Caption Contest)

The last task is how well can a model explain why a caption and cartoon image are funny. To evaluate this very subjective task, a human panel of judges is utilized. The panel of three crowdworkers from Amazon Mechanical Turk are given an unlabeled pair of descriptions of why the cartoon and caption are funny. One of the pair's descriptions is the model's description and the other is the human annotated

description. A correct instance is if the human panel has a majority vote for the model's description. If the panel chooses the human-annotated description, then it is an incorrect instance.

	A	B	% A wins	# ratings	G-γ
Q1	T5-11B	Caption only	84.7%	393	64.4
Q2	T5-11B	OFA $\rightarrow$ T5-11B	74.6%	393	41.6
Q3	T5-11B	T5-Large	68.5%	390	45.9
Q4	FT-GPT-3	In context GPT-3	50.0%	396	23.2
Q5	5-shot GPT-4	Zero-shot GPT-4	64.3%	396	19.7
Q6	5-shot GPT-4	5-shot GPT-3	93.0%	384	86.4
Q7	Human	5-shot GPT-4	67.7%	390	20.9

Table 3: Pairwise human evaluations for explanation, with per-instance agreement according to Gwet's (2014) γ. Q1-Q7 notations refer to the corresponding paragraphs in §3.2.

Figure 3: Prediction Results from the description.
Source: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 688–714
July 9-14, 2023 ©2023 Association for Computational LinguisticsDo Androids Laugh at Electric Sheep? Humor "Understanding"
Benchmarks from The New Yorker Caption Contest

Conclusion:

Humor is a very subjective topic. Something one might find funny, another might find it unfunny. However, computers are, excuse my humor, very computative and objective. The research shows that, as of right now, large neural network models perform worse than humans at identifying humor in all three tasks. However, the authors' research on creating models to objectively identify and classify humor using machine learning and natural language processes could be a step toward further understanding humor for both machines and humans.