(10-1-2024)

Does Ai have sense of humor. It tests AI models on three tasks: matching jokes to pictures, picking the best caption, and explaining why it's funny. Turns out, even the best AIs still struggle. They're good at wordplay and basic references, but miss subtle human quirks. The paper shows there's a long way to go before robots can truly laugh with us, but it helps us build better AIs with a sense of humor!

Paper Overview

Key Research Question/Problems Addressed:

The question delves into 2 main problems:

1. **Understanding basic humor beyond linguistic comprehension(Understanding Language):** The paper goes beyond simply recognizing puns or wordplay and focuses on whether LLMs can grasp the nuances of humor such as cultural references, unexpected twists, and interplay between visuals and text.
2. **Benchmarking LLM humor capabilities against human performance:** The New Yorker Cartoon Contest serves as a challenging and representative testing ground to evaluate LLM humor understanding and compare it to human judgment.

overall this paper talks about “Can large neural networks(LLMs) generate and understand humor.

Methodology:

This paper uses two types of AI models (multimodal and language-only) to tackle "New Yorker" cartoon humor. Both are tested on matching captions to pictures, choosing the funniest caption, and explaining why it's funny. Human annotations and diverse data fuel their training, and human judgment ultimately evaluates their success.

Data:

1. Cartoon and Caption Acquisition:

- 475 unique cartoons from the acclaimed New Yorker Cartoon Caption Contest were selected. Each cartoon was paired with five diverse, human-rated captions that varied in humor level and style. And to ensure a comprehensive dataset, captions ranged from simple wordplay to more complex cultural references and unexpected twists.

2. Human Annotation and Enrichment:

- Annotators provided detailed descriptions of each cartoon's visual elements, including objects, characters, and actions. They identified any unusual or unexpected elements in the scene that might contribute to the humor. For each caption, annotators offered an explanation of why and how the caption was funny, providing valuable insights into the mechanics of humor.

Model Architecture:

The study does not delve into specific details of the model architectures due to potential implementation biases. The choice of architecture within each category (multimodal and language-only) might have variations, but the paper mainly focuses on the high-level comparison between the two approaches.

1. Multimodal Models:

- These models process both the cartoon image and the caption text, similar to athletes performing in unison.
- **Architecture:** They often use transformer-based models with specialized image embeddings to handle visual information alongside textual input.

2. Language-Only Models:

- These models rely solely on textual descriptions of the scene, like poets weaving humor from a whispered description of a painting.
- **Architecture:** They typically use standard transformer models trained on textual data.

Evaluation:

1. Accuracy-based tasks:

- **Matching:** Can the model correctly pair a caption with the cartoon it best fits?
- **Ranking:** Given five captions, can the model identify the funniest one, similar picking the perfect punchline?
- **Explanation:** Can the model break down the mechanics of a joke and explains what makes it funny?

2. Human preference:

- Human judges compare explanations generated by the LLM models with human-written explanations for the same joke.
- They then choose the funnier explanation, regardless of model type.

This subjective evaluation adds a layer of nuance, reflecting on humans' perception of humor rather than just relying on objective accuracy metrics.

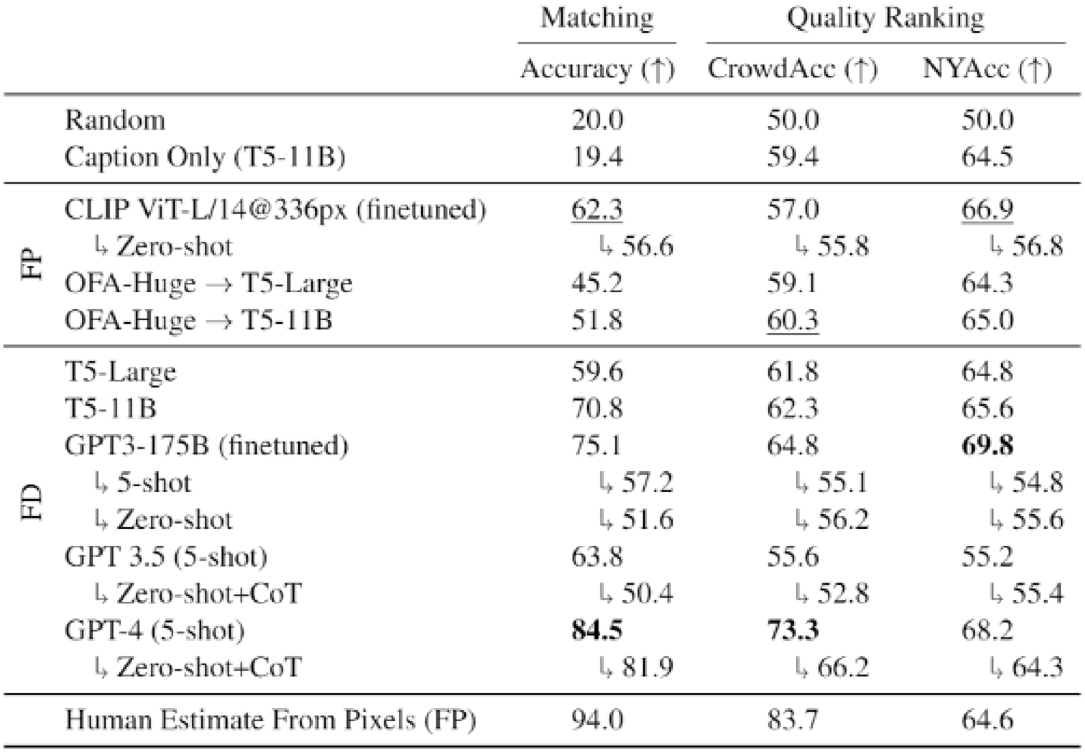
3. Analysis of strengths and weaknesses:

- The study analyzes the models' performance on different types of jokes, identifying where they excel and where they struggle.
- This includes evaluating their ability to handle subtle humor, cultural references, and unexpected twists.

This analysis goes beyond simple accuracy figures and provides deeper insights into the limitations and potential of different LLM techniques for understanding humor.

Overall, the evaluation provides a comprehensive picture of the LLM models' capabilities in the realm of humor, encompassing both objective and subjective assessments, along with a detailed analysis of their strengths and limitations.

Analysis



		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
	↳ Zero-shot	↳ 56.6	↳ 55.8	↳ 56.8
	OFA-Huge → T5-Large	45.2	59.1	64.3
	OFA-Huge → T5-11B	51.8	60.3	65.0
FD	T5-Large	59.6	61.8	64.8
	T5-11B	70.8	62.3	65.6
	GPT3-175B (finetuned)	75.1	64.8	**69.8**
	↳ 5-shot	↳ 57.2	↳ 55.1	↳ 54.8
	↳ Zero-shot	↳ 51.6	↳ 56.2	↳ 55.6
	GPT 3.5 (5-shot)	63.8	55.6	55.2
	↳ Zero-shot+CoT	↳ 50.4	↳ 52.8	↳ 55.4
	GPT-4 (5-shot)	**84.5**	**73.3**	68.2
	↳ Zero-shot+CoT	↳ 81.9	↳ 66.2	↳ 64.3
	Human Estimate From Pixels (FP)	94.0	83.7	64.6

The results showed that:

- GPT-4 generally performed best, especially at matching captions and predicting which captions editors would choose.
- Both "from pixels" and "from description" models did better than the baseline, suggesting they use interactions between image and text for humor understanding.
- Fine-tuning CLIP worked best for matching captions, while OFA+T5-11B was good for ranking and generating captions.
- Zero-shot models performed worse than models with training data, but GPT-4's performance dropped less than others.

Overall, the study shows that AI models are making progress in understanding humor, but there is still room for improvement.

Personal Reflection

What did i learn:

- **AI models are becoming better at understanding humor:** GPT-4 and other models performed well on the tasks.
- **Multiple approaches work for humor understanding:** Both "from pixels" and "from description" models were successful, suggesting multiple strategies can be effective.
- **There is still room for improvement:** Even the best models did not reach human levels of performance, especially when it comes to understanding subtler humor or predicting audience preferences.
- **Chain-of-thought reasoning may be helpful:** GPT-4's less severe performance drop in zero-shot settings suggests its chain-of-thought approach might be a promising avenue for future research.

How does it connect to my Interests/future goals:

1. **Pushing the boundaries of NLP:** This research delves beyond basic language understanding by exploring how LLMs can grasp complex concepts like humor.
2. Humor understanding requires AI to interpret cultural references, unexpected twists, and social cues, delving into its creative and reasoning abilities. This research paves the way for AI with deeper creative potential, something I am interested in contributing to.
3. **Building more engaging AI systems:** Understanding humor is crucial for creating AI systems that can interact with humans on a deeper level, engaging in conversations and responding to our emotional cues. This research avenue is something i would love to explore.
4. This research introduces a new humor-annotated dataset and opens up exciting avenues for further investigation.

Open Questions/ areas for further investigation:

1. How can we create diverse humor datasets(for future research) beyond the New Yorker style and address subjectivity in evaluations, while incorporating cultural avenues into models?
2. Can we use chain-of-thought reasoning across architectures, further develop multimodal fusion(text and pixel), and equip models with better explainability?
3. How can we explore new avenues of research where the LLM have to explore the social cues and understand cultural references?