Supplementary Material for
Punny Captions: Witty Wordplay in Image Descriptions

Arjun Chandrasekaran[1] **Devi Parikh**[1,2] **Mohit Bansal**[3]
[1]Georgia Institute of Technology [2]Facebook AI Research [3]UNC Chapel Hill
{carjun, parikh}@gatech.edu mbansal@cs.unc.edu

1 Additional Experiments

1.1 Relevance of witty caption to image

We compared the relative relevance of the top witty caption from our generation approach against a machine generated boring caption (either for the same image or for a different, randomly chosen image) in a pairwise comparison. We showed Turkers an image and a pair of captions, and asked them to choose the more relevant caption for the image. We see that on average, the generated witty caption is considered more relevant than a machine generated boring caption for the same image 37.5% of the time. People found the generated witty caption to be more relevant than a random caption 97.2% of the time. This shows that in an effort to generate witty content, our approach produces descriptions that are a little less relevant compared to a boring description for the image. But our witty caption is clearly still relevant to the image (almost always more relevant than an unrelated caption).

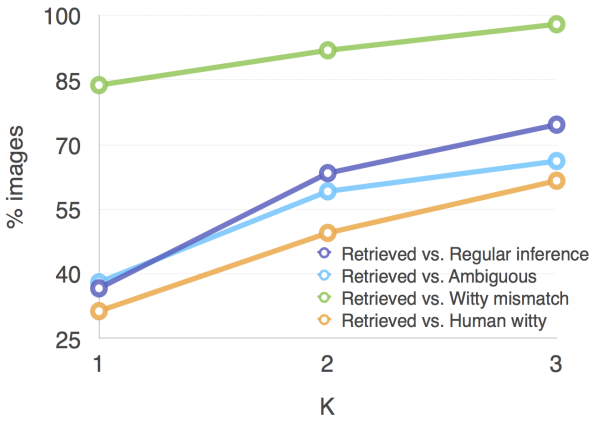


Figure 1: Comparison of wittiness of the top 3 captions from our retrieval approach vs. other approaches. The y-axis measures the % images for which at least one of K captions from our approach is rated wittier than other approaches. As we increase the number of retrieved captions (K), recall steadily increases.

1.2 Retrieved captions vs. baselines

Humans evaluate the wittiness of each of the 3 top-ranked retrieved captions against baseline approaches and a human witty caption. As we see in Fig. 1, at $K = 1$, the top retrieved description is found to be wittier than only a human-written witty caption that is mismatched with the given image (witty mismatch) 83.8% of the time. The top retrieved caption is found *less* witty than even a typical caption (regular inference) about 63.4% of the time. Similarly, the retrieved caption is also found to be less witty than a naive method that produces punny captions (ambiguous) about 62% of the time. We observe the trend that as K increases, recall also increases. On average, at least one of the top 3 retrieved captions is wittier than the (constrained) human witty caption about 61.6% of the time, compared to generated captions which are wittier 84.0% of the time.

Poor performance of retrieved captions could be due to the fact that they are often not perfectly apt for the given image since they are retrieved from story-based corpora. Please see Sec. 4 for examples, and a more detailed discussion. As we will see in the next section, these issues do not extend to the generation approach which exhibits strong performance against baseline approaches, human-written witty captions and the retrieval approach. While these captions might evoke a sense of incongruity, it is likely hard for the viewer to resolve the alternate interpretation of the retrieved caption as being applicable to the image.

2 Design choices

In this section, we describe how our architecture design and parameter choices in the architectures influence witty descriptions. During the design of our model, we made choices of parameters based on observations from qualitative results. For instance, we experimented with different beam sizes to generate a set of high precision captions with few false positives. We found that a beam size of 6 resulted in a sufficient number of candidate sentences which were reasonably accurate. We extract image tags from the top-K predictions of an image classifier. We experimented with different values of K, where $K \in \{1, 5, 10\}$. We also tried using a score threshold, where classes predicted with a score above the threshold were considered valid image tags. We found that $K = 5$ results in reasonable predictions. Determining a reasonable threshold on the other hand was difficult because for most images, class prediction scores are extremely peaky. We also experimented with the different positions that a pun counterpart can be forced to appear in. Based on qualitative examples, we found that the model generated witty descriptions that were somewhat sensible when a pun word appeared at any of the first or last 5 positions of a sentence. We also experimented with a number of different methods to re-rank the candidate of witty captions, e.g., language model score (Jozefowicz et al., 2016), image-sentence similarity score (Kiros et al., 2014), semantic similarity (using Word2Vec (Mikolov et al., 2013)) of the pun counterpart to the sentence, a priori probability of the pun counterpart in a large corpus of English sentences to avoid rare / unfamiliar words, likelihood of the tag (under the image captioning model or the classifier as applicable). etc. We qualitatively found that re-ranking using log. prob. score of the image captioning model, while being the simplest, resulted in the best set of candidate witty captions.

3 Pun List

Recall that we construct a list of puns by mining the web and based on automatic methods that measure the similarity of pronunciation of words. Upon inspecting our list of puns, we observe that it contains puns of many frequently used words and some pun words that are rarely used in everyday language, e.g., 'wight' (which is the counterpart of 'white'). Since a rare pun word can be distracting to a perceiver, the corresponding caption might be harder to resolve, making it less likely to be perceived as witty. Thus, we see limited benefit in increasing the size of our pun list further to include words that are used even less frequently.

4 Qualitative analysis of retrieved descriptions

The retrieved witty descriptions are retrieved from story-based corpora. They often contain sentences that describe a very specific situation or instance. Although these sentences are grounded in objects that are also present in the image, the entire sentence often contains a few words that are irrelevant for a given image, as we see in Fig. 2b, Fig. 2d and Fig. 2e. This is a likely reason for why a retrieved sentence containing a pun is perceived as less witty when compared with witty descriptions generated for the image.

(a) **Generated**: a bear that is bare (bear) in the water.
Retrieved: water glistened off her bare (bear) breast.
Human: you won't hear a creak (creek) when the bear is feasting.

(b) **Generated**: a bored (board) bench sits in front of a window.
Retrieved: Wedge sits on the bench opposite Berry, bored (board).
Human: could you please make your pleas (please)!

(c) **Generated**: a woman sell (cell) her cell phone in a city.
Retrieved: Wright (right) slammed down the phone.
Human: a woman sighed (side) as she regretted the sell.

(d) **Generated**: a loop (loupe) of flowers in a glass vase.
Retrieved: the flour (flower) inside teemed with worms.
Human: piece required for peace (piece).

(e) **Generated**: a female tennis player caught (court) in mid swing.
Retrieved: my shirt caught (court) fire.
Human: the woman's hand caught (court) in the center.

(f) **Generated**: broccoli and meet (meat) on a plate with a fork.
Retrieved: "you mean white folk (fork)".
Human: the folk (fork) enjoyed the food with a fork.

Figure 2: Sample images and witty descriptions from our generation model, retrieval model and a human. The puns (counterparts) that are used in captions (a) to (f) are bare/creak, bored, sell/Wright/sighed, loop/flour/peace, caught and meet/folk respectively. The word in the parenthesis following each counterpart is the pun associated with the image. It is provided as a reference to the source of the unexpected pun which is used in the caption.

5 Interface for 'Be Witty!'

We ask people on Amazon Mechanical Turk (AMT) to create witty descriptions for the given image. We also ask them to utilize one of the given pun words associated with the image. We show them a few good and bad examples to illustrate the task better. Fig. 3 shows the interface that we used to collect these human-written witty descriptions for an image.

6 Interface for 'Which is wittier?'

We showed people on AMT two descriptions for a given image and asked them to click on the description that was wittier for the image. The web interface that we used to collect this data is shown in Fig. 4.

7 Sample images and witty descriptions

We provide qualitative examples for about 30 of the 100 images in our evaluation set as part of the supplementary material. The full set of 100 examples can be found on the author's webpage. Each image is accompanied by 4 witty descriptions from our generative and retrieval models – 3 top-ranked descriptions, and 1 low-ranked 'bad' description. We also provide the descriptions produced by the 3 baseline approaches – Regular inference, Witty mismatch and Ambiguous, which are described in Sec. 4 of the main paper. Please see the accompanying pdf titled, 'Sample witty descriptions - all methods. pdf'.

References

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410* .

Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539* .

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*.

Hi, my name is **(F)Punky**. I am an **Artificial Intelligence (AI)**.
I am learning how to be witty. **Please write a caption about this image that contains a pun.**

Keyboard shortcuts

Previous	Left arrow
Next	Right arrow

Task: Write a witty sentence about the image containing one of the puns listed beside the image.

Please see a few good examples (green font) and bad examples (red font) below.

HIDE EXAMPLES

Good examples

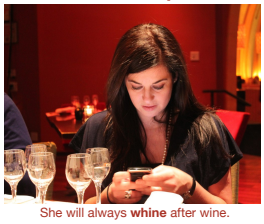
Witty sentence with a pun:
Emotional wedding where the cake is in **tiers**.

Witty sentence with a pun:
A woman at a dine and **whine**.

Witty sentence with a pun:
A cat is pressing **pause** on the phone.

Bad examples

A bridesmaid is in **tiers** at a wedding.
[Pun word should make sense! This caption makes sense for the *original* word but not for the pun.]



She will always **whine** after wine.
[No personal viewpoints. No first person accounts!]

Sleepy cat said, "Dance to the music without pause".
[Shouldn't be what a character in the picture might say!]

If you don't follow these instructions, your work will be rejected.

PREVIOUS NEXT

Task 1/5

List of puns: waul (wall), wight (white), stile (style), poll (pole), sine (sign)

Write a caption about this image using waul, wight, stile, poll, or sine.
Remember: The caption should be relevant to the image, and the sentence should make sense for: waul, wight, stile, poll, or sine.

Witty sentence here ...

Figure 3: AMT web interface for the 'Be Witty!' task.

Hi, my name is **(F)Punky**. I am an **Artificial Intelligence (AI)**.
I'm trying to learn to be **witty by using puns** while describing images. I'm not very good yet, and I'd like to learn so I can slowly get better.

Please tell me which of the following two captions are wittier for this image. To give you a sense for what pun I was going for -- I'll also show you in parenthesis what I saw in the image which I then made a pun around.

Even if both captions seem not all that witty, please indicate the one that seems (ever so slightly) wittier.

I will benefit from this positive feedback! Thanks :)

PREVIOUS **Task 5/15** NEXT

Which of the two captions for the image is wittier?

CAUGHT (COURT) A TENNIS PLAYER HITTING THE BALL .

THE TEDDY BEAR WERE BARE (BEAR)

Keyboard shortcuts	
Top caption	Ctrl+j
Bottom caption	Ctrl+k
Previous	Ctrl+d
Next	Ctrl+h

Figure 4: AMT web interface for the 'Which is wittier?' task.