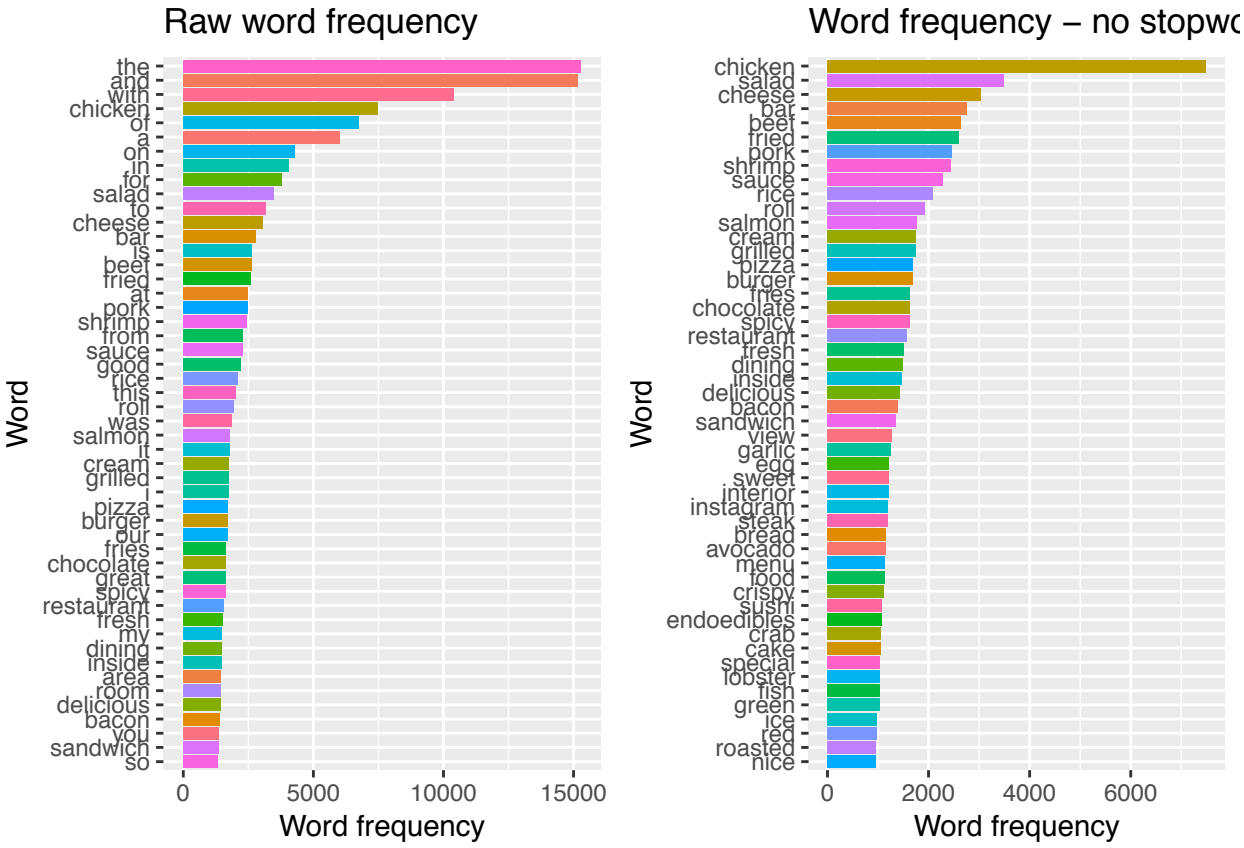
Chicken and Waffles :D

In this notebook I'll try and be more clear about what I was explaining earlier about what I think is causing the "chicken and waffles" phenomenon.

In short - my argument was that these three words are common in the training data, and so if my model cannot derive any information from an input image (image is too confusing, different from training data, etc), it will resort to the training captions and naively pick a sequence that is common there. Here I'll try and back that argument up.

Word Frequency

First what words are common in the dataset?

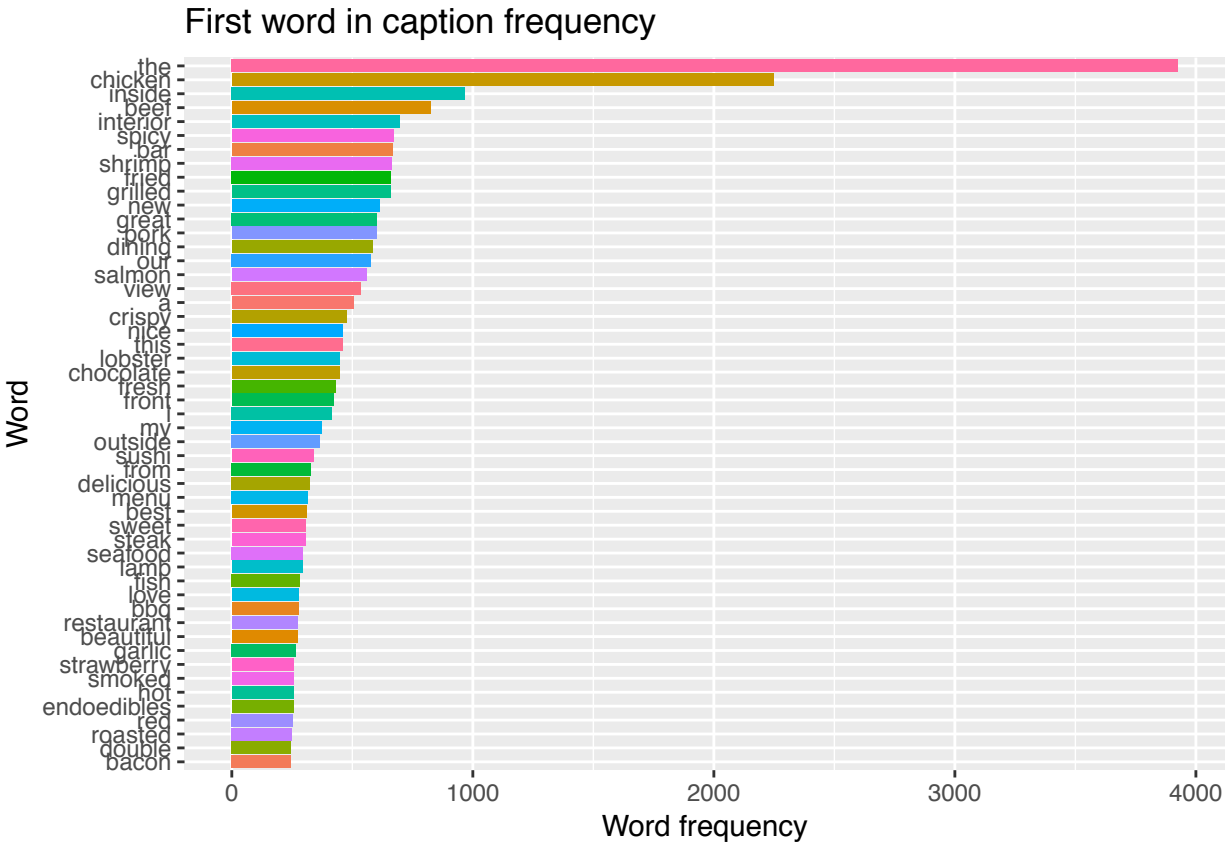


The top couple of words are stopwords. If I remove stopwords, chicken is the most frequent word in the captions.

As in the global frequency, the two most commonly occuring words to appear **first** in a caption are *the* and *chicken*:

Beginning word frequency

But, this model learned to add the first to a predicted sequence by looking at the first word of the training captions. So more interesting than the global word frequency is the frequency of words that start the training captions.



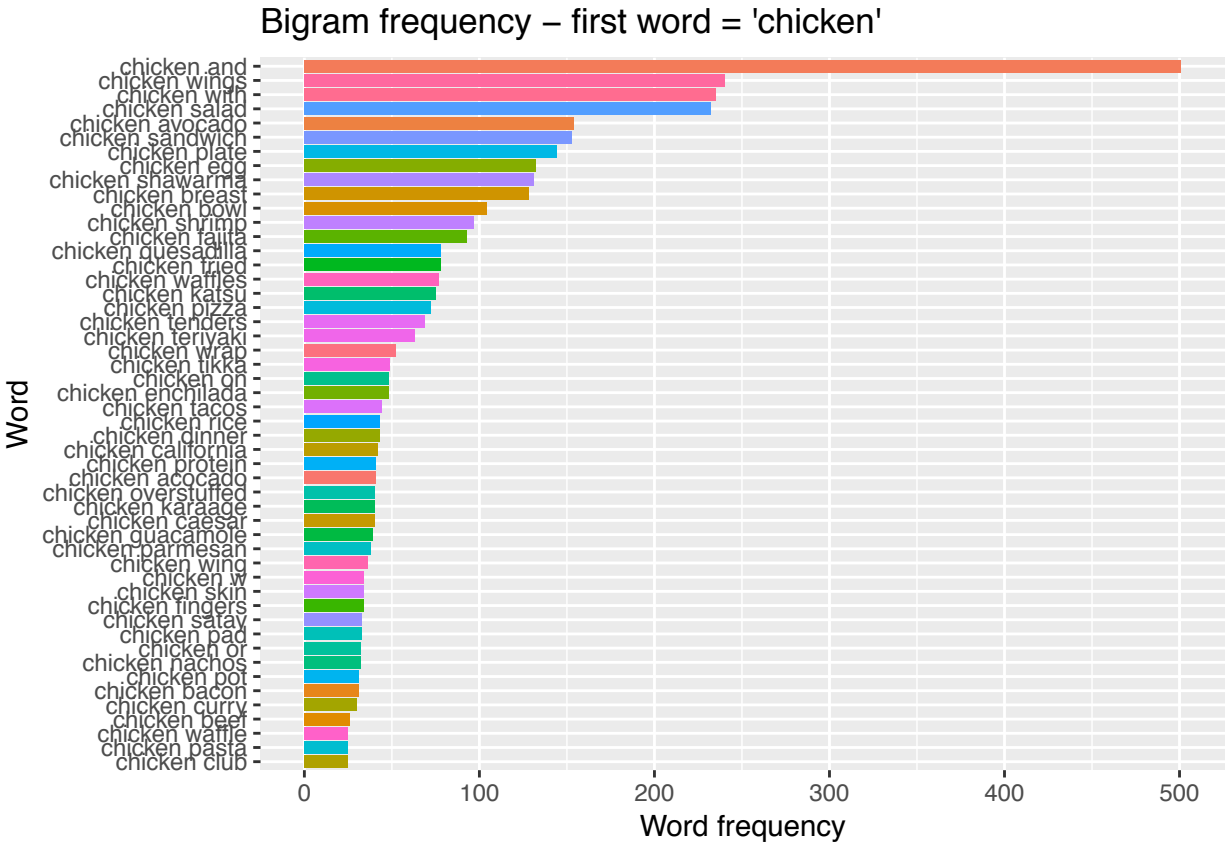
Again the most common words are *the* and *chicken*.

A completely uninformed agent would predict that the first word of every sequence is *the* - i.e completely ignore the image information and go off of the training corpus word frequency.

BUT, assume that my model sees something in the input photo that makes it believe that the first word should be *chicken* instead of *the*. It does so, and so the running sequence is `[chicken]`, which is fed back to the model to predict the next word.

In the training corpus, what words commonly come after *chicken*?

```
## Warning: Too many values at 12709 locations: 3, 4, 46, 64, 65, 134, 277,
## 456, 461, 462, 468, 469, 502, 506, 623, 624, 645, 646, 723, 724, ...
```



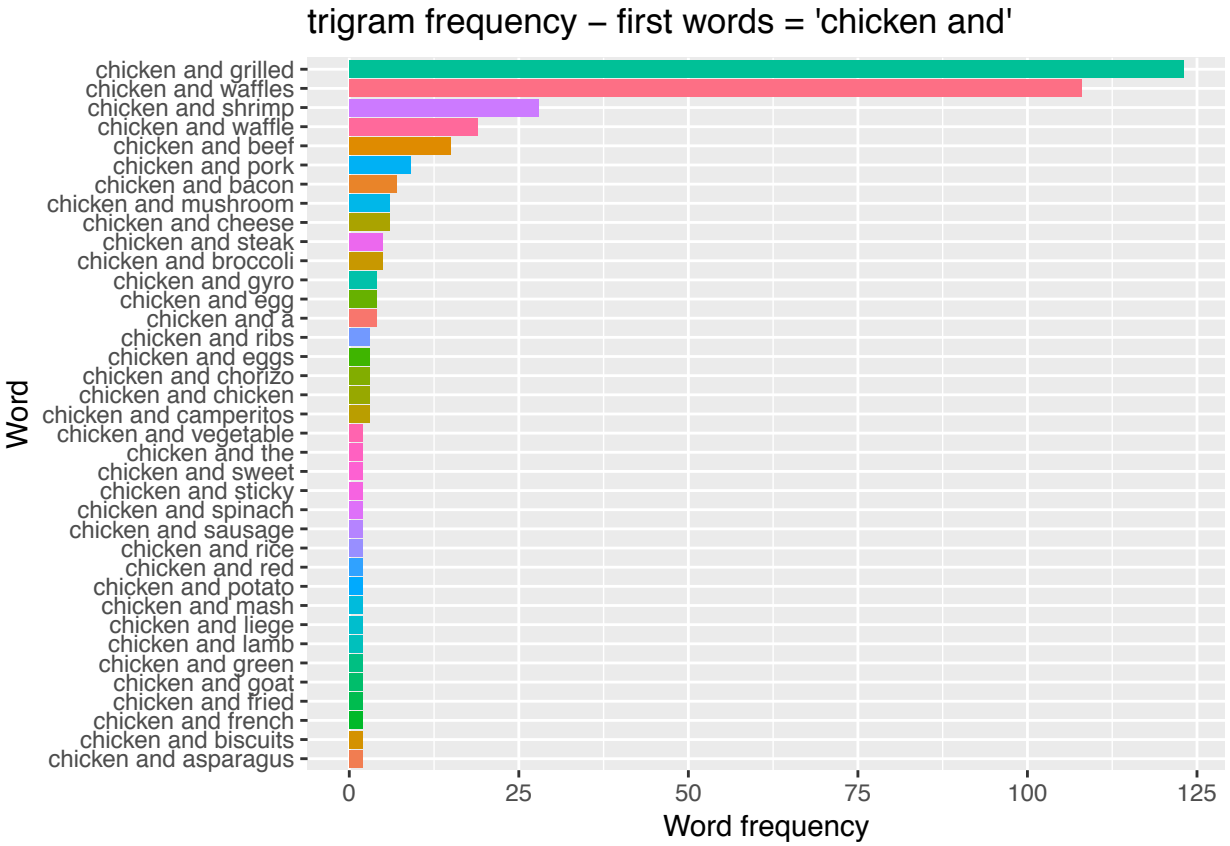
Ok, so by far the most frequent word to come after "chicken" is "and" in the training set.

So lets assume our model has no idea what a picture looks like, but it doesn't feel like putting *the* at the beginning of the sentence, so it puts *chicken*. After it has chicken it naively picks the most frequently occuring word after chicken in the training set - *and*.

So our running sequence is `[chicken, and]`, which is fed back into the model to predict he next word.

. . . What word commonly comes after "chicken and. . . "?

```
## Warning: Too many values at 81 locations: 22, 184, 185, 186, 187, 189, 190,
## 191, 248, 294, 295, 296, 330, 361, 362, 363, 372, 373, 374, 511, ...
```



The second most common word to come after "chiken and" is "waffles".

This high Frequency of captions that start with "chicken", and then the high frequency of the word "and" after "chicken", and high frequency of "waffles" after "chicken and" is what I suspect is fooling my model into the "chicken and waffles" phenomenon.

You might say: "hey - the trigram *chicken and grilled* is more frequent than *chicken and waffles* - why would it predict the latter?"

The reason (I suspect) is that this process of predicting, appending and feeding back into the model doesn't stop until a special token `<endseq>` is predicted - which marks the end of a caption. Thus, the model may have learned that it is less likely for a sentence to end after *"chicken and grilled"* than *"chicken and waffles"*, as the former is ungrammatical.