Florence2 weaknesses

Abstracts: Given that takes input in the following format: "<TASK_ID>your_input_prompt", Florence2 has 4 weaknesses:

1. Given Caption-To-Phrase-Grounding and Open Vocabulary Detection tasks where input prompt is natural English text and model outputs detected object-category from input prompt and bounding boxes per object-category (sometimes polygons for Open Vocabulary Detection): if your_input_prompt has English plurals or having multiple object-categories, then Florence2 will hallucinate extra bounding boxes. From 2 consistent working cases **where 1 surprisingly worked despite my expectation** (Test 1 and Test 3a) and all other failed examples, I find when Florence2 hallucination is triggered by plurals or multiple object-categories in input prompt, **Florene2 takes every subword that can no longer be split by Bart tokenizer at face value and some weirder behaviour** as shown by comparing Test 2a vs Test 2b; Test 3a vs Test 3b; Test 4a vs **Test 4b**.
2. Given Referring-Expression task where input prompt is natural English text and model outputs polygon. Input prompt can contain as many object-categories as you like and use plurals but Florence2 only cares about whichever object-category stands at the beginning of input prompt (kinda like first come first serve). See test 5.
3. Given Region-to-Description task where the input prompt is bounding box and model captions that bounding box, the model can only input 1 bounding box per input prompt.
4. Given OCR detection where there's no input prompt and the model outputs 1 quadrilateral bounding box per word, Florence2 insists that the label for each box has to be unique… So much so that it intentionally misspelt the same recurring word in the image. See test 6

Conclusions

From the contradicting weakness (1) and weakness (2), and cross-examining with the GeoChat results reported by Alexey where this issue didn't occur despite GeoChat model being also a coordinate-format focus VLM just like Florence2, **this weakness is likely a symptom of underfitting where not enough training input prompt with plurals and multiple object-categories in Florence2 pretraining dataset**. **So when encountering scenarios like those, Florence2 text-decoder defaults to its most fundamental pretrained knowledge back when it was just a text-decoder for LLM not VLM, which explains it closely followed Bart tokenizer except for test 4b**. Despite not having access to Florence2 pretraining dataset, **I think finetuning Florence2 on GeoChat and see the results can prove whether this was a training data problem or model architecture problem.**

Finetune only on grounding makes model performance worse!!!

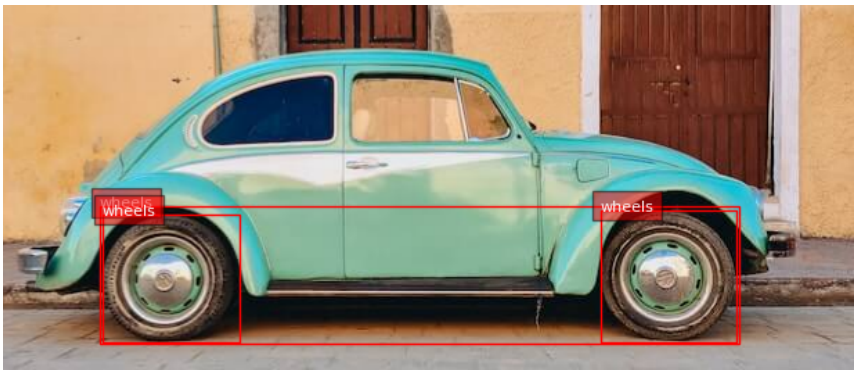
Finetune on all tasks still led to overfit on grounding-caption where all different test images result exact same description, that is: "In the satellite image, there are two gray Boeing 747 airplanes<loc_700><loc_340><loc_860><loc_480><angle_83><loc_700><loc_340><loc_900><loc_480><angle_83> positioned close to each other at the right side of the image." For bounding box visualisation <DENSE_REGION_CAPTION>, it's nowhere better: there's no prediction at all.

Test 1: 1 object-category in input prompt, only singular

Input prompt = “wheels, windows”. Here, even if Bart Tokenizer splits the prompt into subword, there’s still only 2 object- also gave exact same result

Test 2: 1 object-category in input prompt, plurals

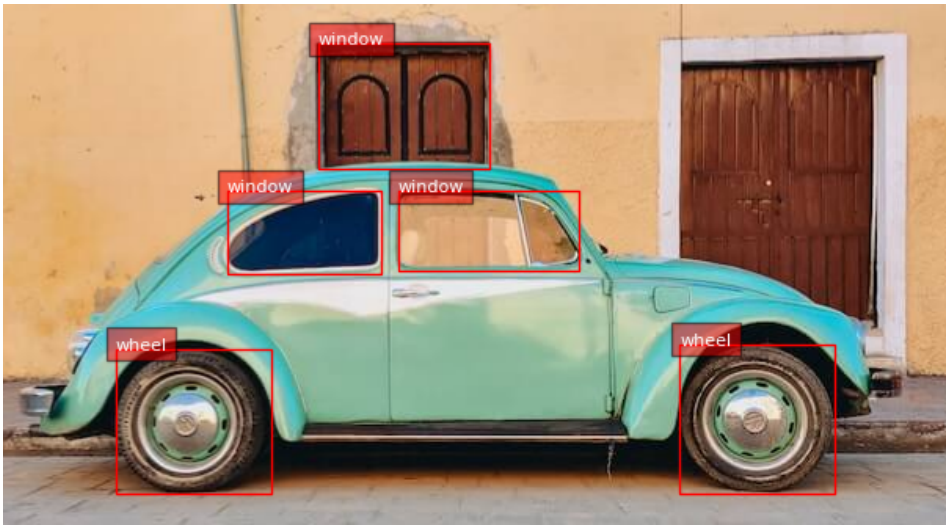
(a)Input prompt = “wheels”. Notice how the unnecessary bounding box surrounds both wheels, this is what I meant by “duplicated bounding box of the same object-category”



(b)Input prompt = “Car tires”. In this test, there are 2 subwords where the subwords will be covered when it passes through Bart tokenizer. So “Car tires” => “Car”, “tires” after passing through Bart tokenizer

Test 3: Many object-category in input prompt, only singular

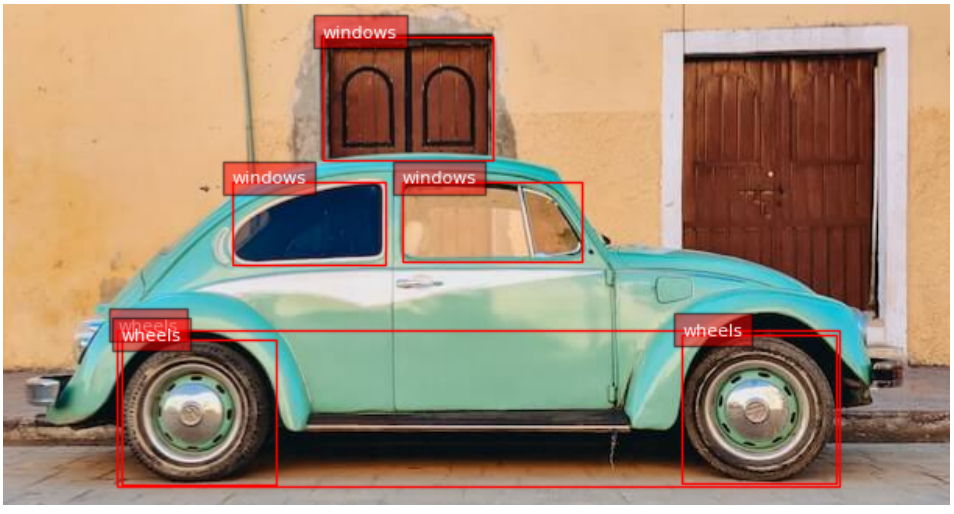
(a)Input prompt = “wheel, window”, “wheel and window” also gives the same correct result.



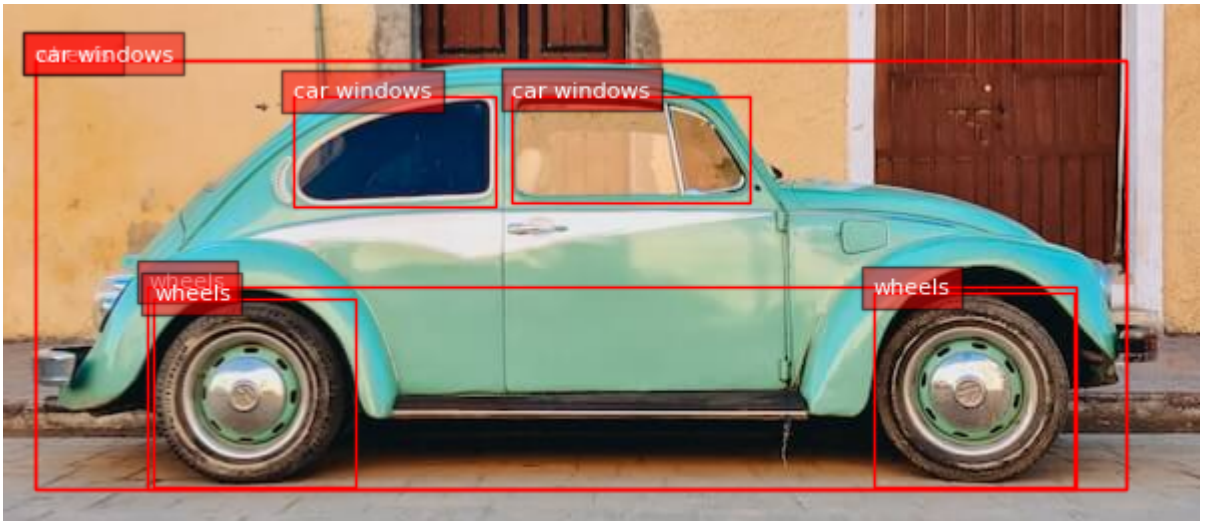
(b) Input prompt = “wheel, car window”. Again, Bart Tokenizer split “car window” into 2 subwords: “car”, “window”. Therefore, the fabricated bounding box is the car itself.

Test 4: Many object-categories in input prompt, plurals

(a)Input prompt = “wheels, windows”. Even with plurals, “wheels” got duplicated but not “windows”.



(b) Input prompt = “wheels, car windows”. Surprisingly, there is 1 fabricated bounding box labelled “wheels” that is overlaid by another bounding box labelled “car window”. It’s surprising because using Bart tokenizer logic, there are only “car”, “windows”, “wheels” , which are 3 subwords that can no longer be split. In this sense, “wheels” shouldn’t cause Florence2 fabricated bounding boxes like how “car windows” did in this input prompt. But here it happens



Test 5: Florence2 behaviour in segmentation tasks

Input prompt = “car tires”

Input prompt = “car windows”.

Input prompt = “car windows and car tires”

Test 6: Florence2 behaviour in OCR Detection

In the example below, it intentionally mispellt “REPEAT”, “REPEATE”, “REPEAY”, “REPEAAT”, “REPE AT” to avoid having more than 1 quadrilateral bounding box per detected phrase (where detected phrase in this case is “REPEAT”).

In the 2nd test, I cropped the image to only have the lower half portion to see if Florence2 could detect the “REPEAT” word anywhere. My prediction was correct, but it was stuck to its misguided pretraining rule so it had to intentionally misspelt the 2nd “REPEAT” word -> “REPEATE”


REPEAT
REPEAT