Introduction

Computer vision model deals with visual information such as identifying objects in the image, recognizing shapes and understanding the details like color and texture. So, for computer vision models to be effective across different tasks model need to understand both global concepts and the final details. Global concepts mean the overall scene understanding of the image and final details means pixel level details like recognizing individual objects and their exact shapes and positions.

So, this dual understanding is important for creating versatile computer vision model that can accurately interpret, analyze visual information in various applications. Spatial hierarchy refers to the understanding of visual information at different scales or levels of detail within an image. It involves recognizing objects and their spatial relationship from global context down to the specific details.

For example, in a scene where person is riding a bike on the street. Spatial Hierarchy would involve recognizing the street bicycle and the person as distinct objects and understanding their positions relative to each other and also knowing finer details like the texture of the road or the color of the bicycle. Semantic Granularity refers to how much detail we can understand from visual information. It ranges from getting a general idea of what's happening like describing the scene with few words to noticing small specific details like identifying the color or specific model.

So for example in our street scene it's not just about recognizing a person in a bicycle but also noticing details such as what color clothes the women is wearing or what model the bicycle is. This level of detail helps in accurately describing and understanding the visual content.

Section 2.1

For multitask learning, what is multitask learning? Multitask learning means teaching a model to do multiple related tasks at the same time to improve its overall understanding and performance and to create a strong foundation model for computer vision. Researchers employ multitask learning, this approach involves setting up multiple learning objectives that are specifically designed to enhance the different aspects of visual comprehension.

Let's discuss few key objectives,

So first one is **Image Level Understanding**. So, image level understanding means it recognizes what is in the image and describe it. So, this involves 2 tasks identifying objects in the image like a car or a tree right and then generates a caption that describes the whole scene, like the red car parked under tree.

We have **Region / Pixel Level Recognition** this objective is to accurately identify and localize objects within the image. It will not only identify objects, but it will also precisely outline where each object is in the image and task in this category include object detection means putting bounding boxes on the objects and then segmentation where we put our mask on the images. Then the third one is **Fine Grained Visual Semantic Alignment Task** now these tasks require detailed understanding of both textual understanding and the visual elements so their task involve pointing which part of an image correspond to specific text phrases for example given a description that says a person riding a bicycle so the model should accurately pinpoint the exact area in that image where the person and bicycle are located.

So, in multitask learning the model learned to handle all these three tasks simultaneously. So by training on variety of tasks like these, the model become more capable and it can do things like describing objects in the image, describing scenes accurately and then understanding how the objects relate to each other. So, when the model learned all these tasks together it does not just get good at one thing like spotting objects like detecting objects or writing captions, it gains a vital and deeper understanding of images. This method is crucial for making smarter computer vision systems that can be used in lots of ways like in self driving cars or in medical technology right so this is what this multitask learning is.

Model:

What is **Universal Representation Learning**, it means training a single model that can understand different type of information without needing a task specific adjustment or retraining.

So Florence 2 is designed to handle multiple vision tasks using a single set of weights and a unified architecture. This means that instead of creating a separate model for tasks like image classification, object detection or image captioning, Florence 2 can perform all these tasks with the same underlying model. So Florence 2 use **Sequence To Sequence Learning** approach. So this approach commonly used in natural language processing task. Let's try to understand this image so Florence 2 work by taking image and the task instructions which are prompts as input and it uses a vision encoder to turn images into a vision visual embeddings and which are then combined with the text embeddings and these combined embeddings go through a transformer based encoder decoder to produce a final response and depending upon the task and the prompt, the response can either be text or region. So for example if you are providing a simple text, it will directly process them in the sequence to sequence format but if you are providing image and text, region specific tasks, so for region specific task, you provide image and text to model, for task like object detection and segmentation.

So for those tasks Florence 2 use location tokens. So location tokens are special tokens that are added to the models vocabulary to handle the specific region within the image. So for example if you'll provide text input so you will get an output in the format of sequence to sequence format but if you are providing a region specific detail, for these kinds of tasks, model use location tokens which are specifically used to handle the location of the objects in the images.

So the first one is **Box Representation,** box representation is used for tasks like object detection, where tokens represents the coordinate of the bounding box around the object.

Then we have **Quad Box Representation,** now what is it. So they are used for text detection whenever we want to detect text in the images quad box representation is used for that.

Then we have a **Polygon Representation,** so this is used for tasks like segmentation where tokens represent the vertices of a Polygon outline angle segmented area. So by adding these location tokens into the vocabulary, Florence 2 can process region specific information uniformly across different tasks without needing a task specific modifications.

Florence 2 model is trained using a standard language modeling techniques like cross entropy loss, this ensures that the model can generate accurate and contextually relevant output based on the input image and the task blocks.

Let's talk about this data so Florence 2 is trained on FLD 5B dataset.so they have done three main type of annotations one is **Text Annotation**, second is **Region Text Pair** annotation and the third is **Text Phrase Region Triplets.**

So the first one is Text Annotation, takes annotation categorize images based on different levels of detail from brief descriptions to more detailed narratives. So text annotations are categorized into 3 levels of detail **brief text, detailed text and more detailed text.**

Second one is region text pair, region text pairs provide textual description for specific region within the image. So we are defining our text and in that text we are defining the region.

Third one is text phrase region triplets. Text phrase region triplets combined descriptive text, noun phrases related to the image objects and annotations for these objects.