Understanding Dataset Collection: (Problem-3)

এখন আসা যাক মেইন প্রবলেমে— ViT (Vision Transformer) মডেল কিন্তু region detect করতে পারে না। মানে, সে ইমেজের মধ্যে কোন অংশে লেখা আছে সেটা বুঝতে পারে না।

তাকে যদি আমরা OCR টাস্কের জন্য ট্রেইন করি, তাহলে সে শুধুমাত্র ওয়ার্ড-লেভেল ইমেজ নিতে পারবে এবং সেই ইমেজের টেক্সটটা এক্সট্রাক্ট করতে পারবে।

কিন্তু আমরা যে ডেটা কালেক্ট করছি, সেটা তো মূলত সেন্টেন্স বা প্যারাগ্রাফ লেভেলের—
একটা ইমেজে হয়তো অনেকগুলো ওয়ার্ড বা একাধিক লাইন লেখা থাকবে। এই ক্ষেত্রে ViT মডেল তো বুঝতেই পারবে না—"ইমেজের কোন অংশে কোন ওয়ার্ড আছে?"

তাহলে সমাধান কী?

আমাদের হাতে দুইটা অপশন আছে:

Option 1: Manual Word-Level Annotation + Cropping

1. আমরা ডেটাগুলো ম্যানুয়ালি ওয়ার্ড-লেভেল এনোটেশন করবো (bounding box সহ)।
2. এরপর সেই এনোটেশনের region ধরে ওয়ার্ডগুলো আলাদা করে ক্রপ করবো।
3. প্রতিটা ক্রপ করা ওয়ার্ড ইমেজকে fix size (যেমন 224x224) এ স্কেল করবো।
4. তারপর সেই ইমেজ ViT-এর মতো মডেলে ফিড করবো।

📌 এতে কাজ হবে ঠিকই—but মডেল ওয়ার্ড লেভেল পর্যন্তই পারবে।
⇒ কোন বড় সেন্টেন্স বা প্যারাগ্রাফ দেওয়া হলে, সেটা ViT হ্যান্ডেল করতে পারবে না।
⇒ এটা মূলত “single word recognition” system হবে, full OCR নয়।

Option 2: Full-Fledged OCR System (Object Detection + Recognition)

যদি আমরা একটা পুরোপুরি end-to-end OCR বানাতে চাই— যেটা ইমেজে থেকে পুরো সেন্টেন্স বা লাইনগুলো detect করে, এবং তারপর সেগুলো এক্সট্রাক্ট করবে, তাহলে আমাদের করতে হবে:

1. প্রথমে Object Detection Module (যেমন YOLO বা অন্য কোনও lightweight CNN detector) এটা detect করবে ইমেজে কোথায় কোথায় লেখা আছে (line/word bounding box)।

2. এরপর Recognize করার জন্য ViT বা অন্য OCR model ইউজ করবো— যেটা প্রতিটা detected রিজিওন থেকে গ্রাফিম লেভেলে টেক্সট বের করবে।

📌 এই এপ্রোচ অনেক বেশি powerful, স্কেলেবল, এবং future-proof।
→ বড় সেন্টেন্স হোক বা প্যারাগ্রাফ—মডেল পারবে সবকিছু detect + extract করতে।