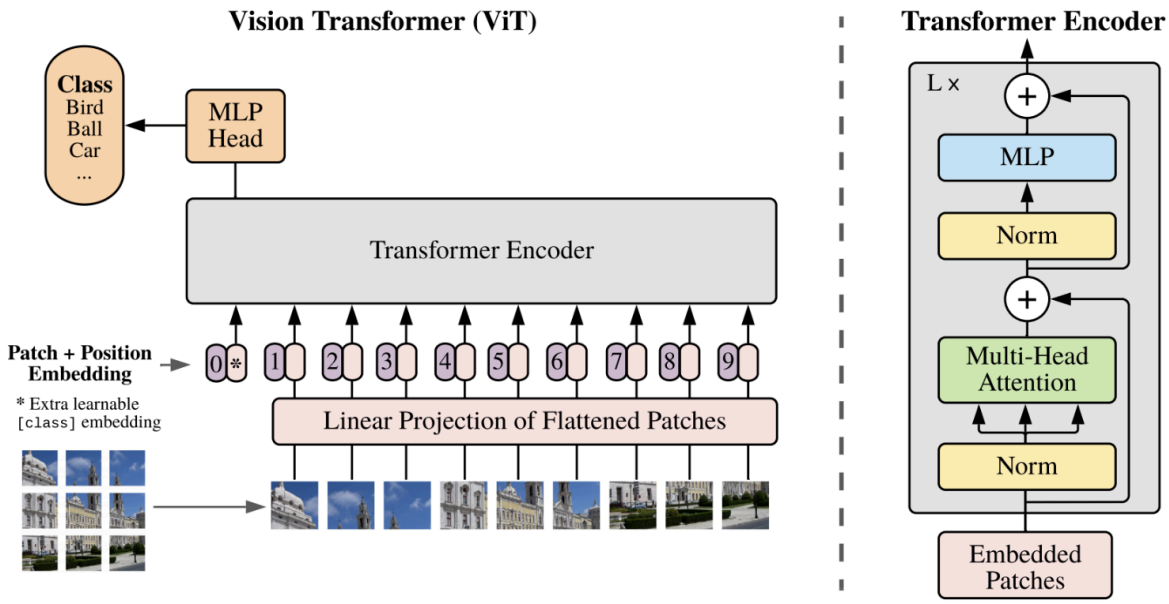
Bangla OCR

What is Bangla OCR:

আমাদের মূল লক্ষ্য হলো এমন একটি **Bangla OCR (Optical Character Recognition)** সিস্টেম তৈরি করা, যা হ্যান্ডরিটেন বাংলা টেক্সটের ছবি ইনপুট হিসেবে পেলে সেগুলোকে সঠিকভাবে রিড করতে পারবে।

এই কাজের জন্য আমরা ব্যবহার করছি **ViT (Vision Transformer)**-ভিত্তিক মডেল।



What is ViT (Vision Transformer)?

ViT (Vision Transformer) হচ্ছে একটি Transformer-based Vision Model, যা মূলত ছবির উপর বিভিন্ন টাস্ক (যেমন: Classification, OCR, Detection) খুব কার্যকরভাবে করতে পারে — যদি এটি হাই কোয়ালিটি ডেটাসেটে ট্রেইন করা হয়।

ViT-এর একটি সীমাবদ্ধতা:

> Transformer ভিত্তিক মডেলগুলো খুব "data-hungry" — এদের ভালোভাবে কাজ করাতে হলে অনেক বড় ও পরিষ্কার ডেটাসেট লাগে। ছোট বা নোইজি ডেটাতে এদের পারফরম্যান্স খুব কমে যায়।

কিন্তু **ViT-**এর সবচেয়ে বড় সুবিধা:

> আগের প্রচলিত মডেল যেমন CNN বা RNN-এর তুলনায় ViT অনেক বেটার long-range dependency বোঝে এবং কম্প্লেক্স ফিচার রিলেশনশিপ ধরতে পারে।

ViT কীভাবে কাজ করে (Data Flow)?

ViT মডেল মূলত দুইটি অংশে বিভক্ত থাকে:

1. Patch Embedding Block (Vision-specific part)
2. Transformer Encoder Block (Generic Transformer layers)

🔍 ViT-এর কাজের তিনটি ধাপ (OCR-এর জন্য):

Patch Extraction

- আমাদের ইনপুট ইমেজ (যেমন Bangla লেখা) কে প্রথমে ছোট ছোট ফিক্সড-সাইজ প্যাচে ভাগ করতে হয়, যেমন 16x16 বা 22x22।
- পুরো ইমেজকে ধরে N সংখ্যক ছোট টুকরা (patch) বানানো হয়।

Linear Projection + Positional Encoding

- প্রতিটা প্যাচকে একটির ভেক্টরে কনভার্ট করা হয় — একে বলে embedding।
- তারপর প্রতিটি embedding-এর সাথে তার অবস্থান বোঝাতে Positional Encoding যোগ করা হয়।

Transformer Encoder

- সব প্যাচ embeddings কে Transformer Encoder ব্লকে পাঠানো হয়।
- সেখানে Self-Attention Mechanism ব্যবহার করে ViT পুরো ইমেজের বিভিন্ন অংশের মধ্যে সম্পর্ক বোঝে।
- ফাইনাল আউটপুট হয়: ইমেজের একটি ধারাবাহিক ভেক্টর রেপ্রেজেন্টেশন।

👉 এই ভেক্টর সিকোয়েন্সকেই আমরা পরবর্তীতে CTC অথবা Transformer Decoder দিয়ে ডিকোড করে মূল টেক্সট (বাংলা লেখা) বের করি।