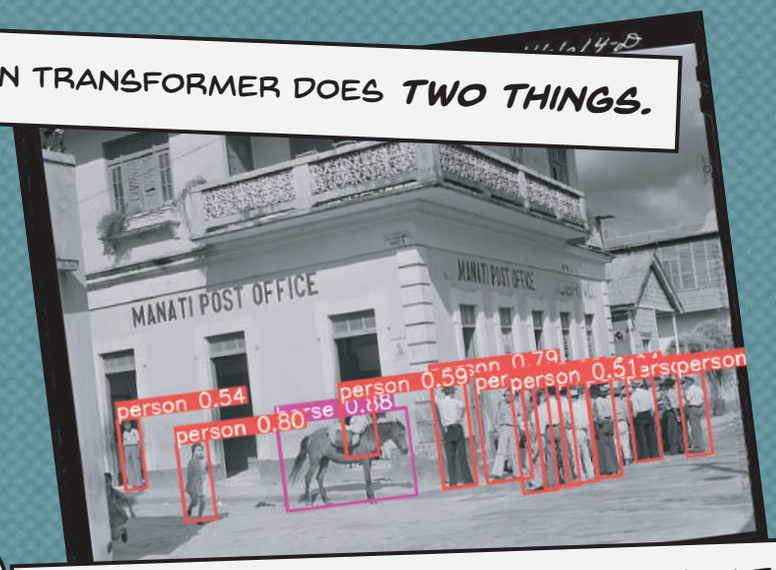
LET'S GIVE IT A TRY WITH THIS IMAGE.

WE'LL USE A MACHINE LEARNING MODEL CALLED ***THE VISION TRANSFORMER*** (VIT).

THE VISION TRANSFORMER DOES ***TWO THINGS.***



1. IT TRIES TO DETECT KNOWN ELEMENTS IN THE IMAGE (THINGS LIKE PEOPLE OR CARS OR FLOWERS), AND IT LEARNS THE SIZE & POSITION OF THOSE THINGS.

2. IT USES WHAT IT HAS LEARNED FROM THE LABELLED IMAGES IT WAS TRAINED ON TO GENERATE A NEW CAPTION.

```
{
  "generated_text":
      "a man standing next to
       a building with a horse "
}
```

THEY'RE OFTEN WORSE THAN HAVING NO CAPTION AT ALL.

THESE MACHINE-GENERATED CAPTIONS COME NOWHERE NEAR THE LIBRARY'S EXACTING STANDARDS.

```
{
  "generated_text":
      "a woman is standing in
       a parking lot "
}
```

... AND PLACES THAT AREN'T IN THEIR WORLD VIEW, LIKE LOW-INCOME SETTLEMENTS.

```
{
  "generated_text":
      "a man in a field with
       a bunch of dead trees "
}
```

VERY OFTEN THOUGH, THE MODELS STRUGGLE WITH OBJECTS THAT AREN'T FOUND A LOT IN THEIR TRAINING SETS (LIKE SUGAR CANE)

```
{
  "generated_text":
      "people standing around a building "
}
```

SOMETIMES THE IMAGE DESCRIPTIONS WE GET FROM JACK'S IMAGES OF PUERTO RICO ARE ACCURATE... IF A LITTLE BIT UNIMAGINATIVE.