PAPER READING PRESENTATION

Analysis of ‘Learning Transferable Visual Models From Natural Language Supervision’

By Krishak Aneja

Core Idea

- The paper introduces **CLIP** (Contrastive Language-Image Pre-training), a model trained on 400 million image-text pairs to learn visual concepts through natural language supervision. Unlike traditional models that rely on fixed object labels like ‘dog’ or ‘cat’, CLIP uses contrastive learning to associate images with free-form text, enabling **zero-shot transfer** to diverse tasks without task-specific training.
- **Key Innovations**:

1. *Natural Language Supervision*: CLIP leverages unstructured text paired with images, bypassing the need for labeled datasets. This allows the model to generalize across 30+ tasks, from OCR to action recognition.
2. *Contrastive Learning:* Instead of predicting exact text captions, CLIP learns a shared embedding space where matching image-text pairs are aligned. This approach improves training efficiency and scalability.

Motivation & Mechanism

- Traditional models (e.g., ResNet) use narrow supervision i.e. hand picked labels that limits their ability to generalize.
- CLIP uses **natural language** as supervision which is richer and more expressive. For example, a crane might be labeled ‘construction equipment’ or ‘bird’ depending on context. CLIP trains on text scraped from the web, text that captures diverse visual concepts. This also makes it scalable: No manual labeling; leverages vast internet data.
- How? Images and text are mapped to a shared space.
  - **Dual Encoders**: Image encoder: ResNet or Vision Transformer(ViT) + Text encoder :Transformer→ joint embedding space.
  - **Contrastive Loss**: During training, maximize cosine similarity for matching (image,text) pairs; minimize for mismatched. This avoids predicting exact text and scales better than generative methods like VirTex (4x faster).
  - **Prediction**: Compare image embedding with *all possible text embeddings* (e.g., class names).

Generalization & Prompt Engineering

- Task flexibility: OCR, geo-localization, action recognition via text prompts.
- Replace rigid labels with **descriptive prompts** (e.g., "A photo of a {label}, *a type of pet*" where label could be 'dog', 'cat' etc).
- Zero-shot power: Outperforms supervised ResNet-50 on 16/27 datasets *without training on them.* For example, CLIP achieves 76.2% accuracy on ImageNet without ever seeing its labels—matching a fully supervised ResNet-50.
- Ensembling multiple prompts (e.g., 'big' or 'small' versions of labels) adds another 3-5% accuracy, all without fine-tuning.

Robustness & Bias

Robustness:

Resists adversarial examples (sightly altered versions of sample images that get classified wrongly) since there is no overfitting to fixed labels.

Handles distribution shifts better: 75% smaller accuracy drop on natural distribution shifts vs. ImageNet models.

Bias Concerns:

Inherits biases from web data (e.g., racial/gender stereotypes

- FairFace: Black faces misclassified as “animal” 14% vs. 8% for other races.
- Crime labels: 16.5% of male images vs. 9.8% female.

the model’s reliance on unfiltered internet data remains a critical ethical concern

Limitations & Ethical Impacts

- **Limitations**:
  - Struggles with highly specialized tasks (e.g., satellite imagery, medical diagnostics) and abstract reasoning (e.g., counting objects in CLEVR).
  - Compute-heavy (64x ResNet-50, 18 days on 592 GPUs).
- **Ethical Risks**:
  - Surveillance misuse (e.g., facial recognition).
  - Environmental cost of large-scale training.
- **Comparison to Human Learning**:
  - Humans adapt faster (1-shot improves pet classification accuracy by 22%).
  - CLIP's few-shot learning lags, but zero-shot bridges gaps.
  - Humans use context and reasoning; CLIP relies on static text prompts. General trend with Machine Learning is that progress is driven by mimicking human capabilities. In this case, to work towards actually extract semantics (meaning) from text and images.

Conclusion and Impact

What is CLIP?
At its core lies the creation (pre-training) of the shared image-text embedding space–intending to capture semantic relationships, not just pixel patterns.

The good
Democratizes AI: No labeled data needed for new tasks.
Multimodal Foundation: bridges vision + language, redefining computer vision.

The bad
Social Bias: Tendency to inherit the biases present in the dataset.
Data Dependency: Uses massive compute and fails to address the poor data efficiency of deep learning.

Power vs. Power trade-off:
CLIP is a *power*ful, robust and general (task and dataset-agnostic) approach at the cost of equivalent resource and *power* (energy) consumption.