CLIP: Contrastive Language-Image Pre-Training

Reproduction, Analysis, and Trilingual Extension

Marine Vieillard & Mohamed Amine Grini

M2 Maths & AI

December 12, 2025

Presentation Outline

Why this Paper? A Paradigm Shift

Context:

- Traditional models (ResNet, etc.) are limited by predefined classes (e.g., ImageNet 1k).
- Problem: Impossible to classify an object outside the training vocabulary without retraining.

CLIP's Proposal (Radford et al., 2021):

- Learning **visual concepts** from natural language.
- Bridging the semantic gap between pixels and words.

Mechanism: Contrastive Learning

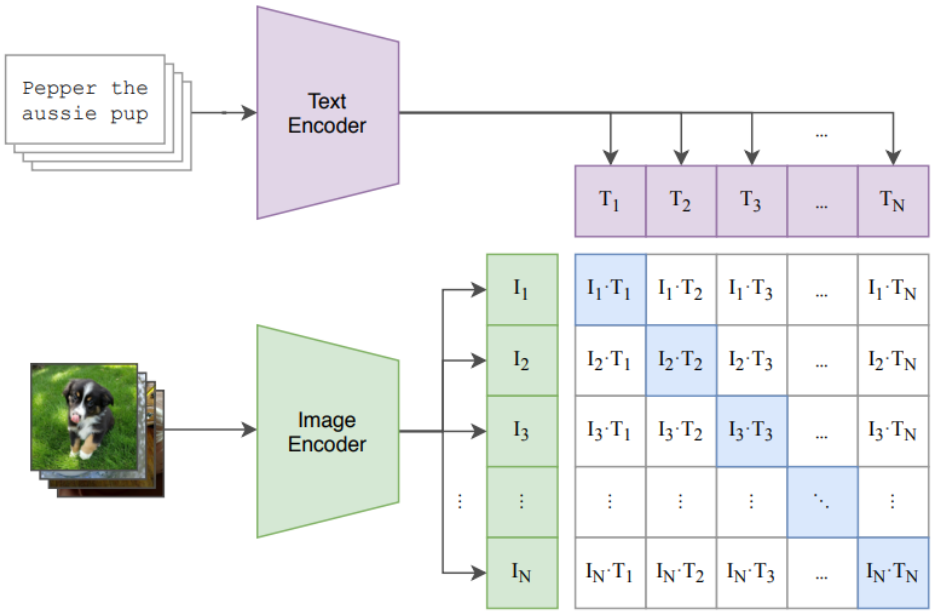


Figure: Contrastive pre-training

Training Objective:

- Maximize cosine similarity for correct pairs (diagonal).
- Minimize similarity for incorrect pairs.
- Result: A shared multi-modal embedding space.

Major Contribution: Zero-Shot Transfer

The Concept:

- Classification becomes a **retrieval** task.
- No specific classifier training needed for CIFAR-100.

Process:

1. Feed the image to the model.
2. Provide a list of prompts: *"A photo of a {label}"*.
3. CLIP predicts which description best matches the image.

→ High robustness to distribution shifts.

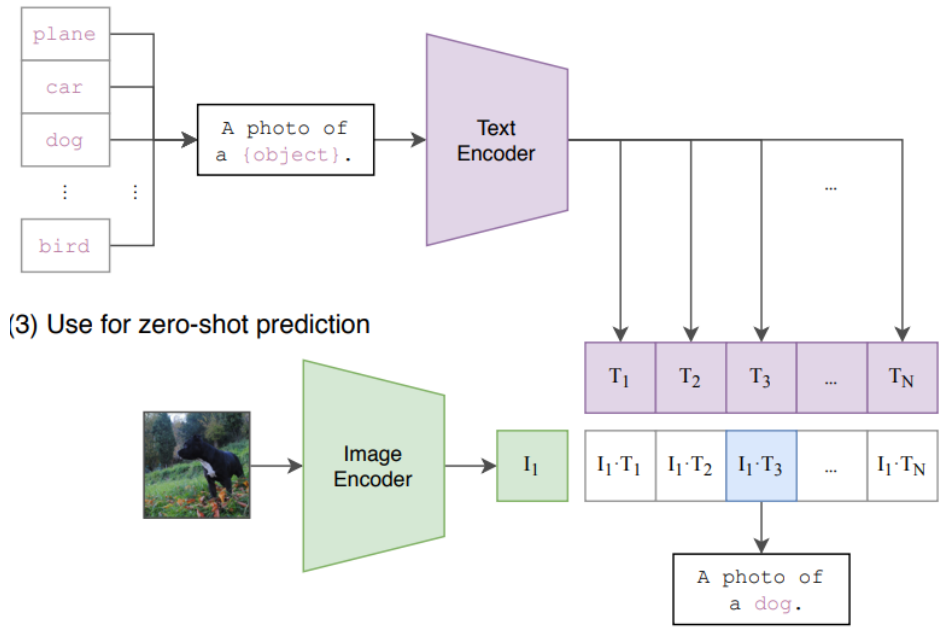


Figure: Zero-shot Prediction

Implementation Structure

Reproduction of results on the **CIFAR-100** dataset using the `ViT-B/32` model.

- **Clip.py**: Basic Zero-Shot implementation.
- **Clip_ensembling.py**: Advanced "Prompt Ensembling" technique.

Prompt Engineering & Ensembling

Problem: A single word ("apple") is ambiguous and out-of-distribution compared to training sentences.

Solution (Ensembling): Using multiple templates to smooth out noise.

```
templates = [
    "a photo of a {}.",
    "a bad photo of a {}.",
    "a rendering of a {}.",
    "the embroidered {}."
    # 12 variations ...
]
```

Averaging text embeddings before classification.

Reproduction Results (ViT-B/32)

The obtained results confirm those of the paper for the Base model:

Method	**CIFAR-100 Accuracy**
Random Chance	1.00%
Basic Zero-Shot	**62.93%**
Prompt Ensembling	**63.72%**

Conclusion of Reproduction:

- Successfully validated the paper's claims.
- Validated Prompt Engineering as a proxy for hyperparameter tuning.

Hypothesis and Methodology

Context: CLIP was trained on **English-only** datasets.

Hypothesis: Given the scale of the web (400M pairs), the model must have seen other languages implicitly.

Project Extension (Scaling to ViT-L/14@336px):

- **Translation** of 100 CIFAR classes (EN $\rightarrow$ FR, DE).
- **Multilingual Prompts**: *"Une photo de {label}"*, *"Ein Foto von {label}"*.
- **Trilingual Fusion**: Averaging embeddings of EN + FR + DE.

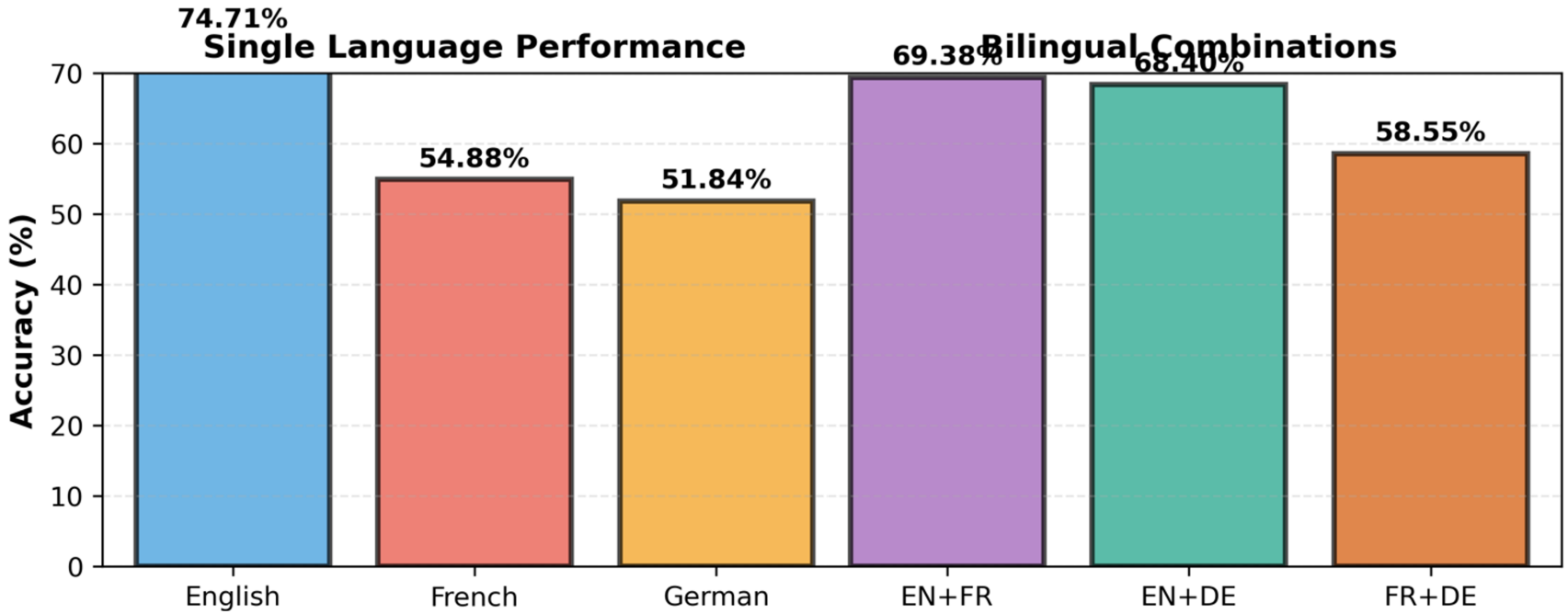
Results: Multilingual Capabilities (ViT-L/14)

Single Language Performance:

- **English**: 74.71% (Baseline)
- **French**: 54.88%
- **German**: 51.84%

Fusion Performance:

- **FR + DE**: 58.55%
  (> French or German alone)
- **EN + FR + DE**: 66.32%



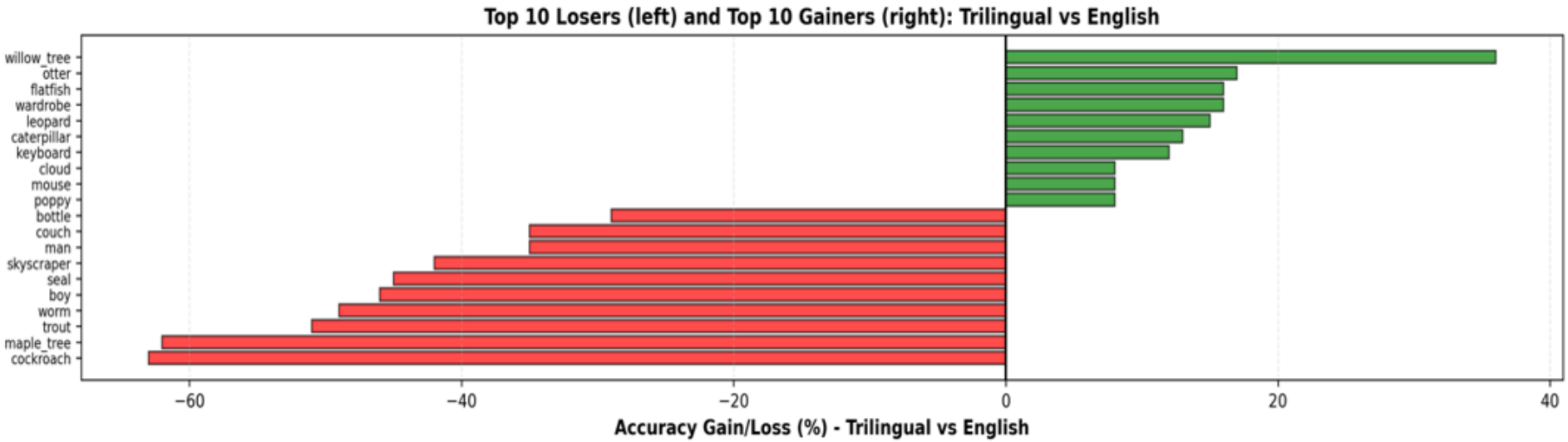
Note: The FR+DE combination outperforms individual languages, proving that linguistic diversity reduces noise even without the dominant language.

Multilingualism: Gains and Losses

Why use multilingualism if English is better?

Specific Use Case: Disambiguation

- **Top Gainers** (Trilingual vs English):
  - Classes: `Willow Tree`, `Otter`, `Flatfish`.
  - Gain of up to +30% on certain difficult classes.
- **Interpretation**: Multilingual prompts act as a disambiguation tool. Concepts ambiguous in English might be precise in French or German.



Conclusion

1. **Successful Reproduction:** Validated the paper's Zero-Shot performance (63% on ViT-B/32).
2. **Scaling Up:** Switching to ViT-L/14 drastically improves results (75%).
3. **Implicit Multilingualism:** Proven capability even without explicit training.
4. **Key Finding:** Mixing languages (e.g., FR+DE) improves performance by providing multiple semantic viewpoints.

Thank you for your attention!

Code available on GitHub :
https://github.com/MV-13/CLIP-Contrastive-Language-Image-Pre-training