# VL2Lite: Task-Specific Knowledge Distillation from Large Vision-Language Models to Lightweight Networks

Title: VL2Lite: Task-Specific Knowledge Distillation from Large Vision-Language
Models to Lightweight Networks

Authors: Jang, Jinseong, Chunfei Ma, and Byeongwon Lee

Publication: CVPR 2025



# Abstract

- 대규모 모델은 성능은 좋지만 계산량이 너무 많아, 자원이 한정된 환경에 배포하기 어려움
- 가벼운(Lightweight) 모델이 필요하지만, 단순히 크기만 줄이면 복잡한 데이터를 처리하는 능력이 떨어져 정확도가 낮아지는 문제가 발생
- image classification 데이터셋에서 **최대 7%의 성능 향상**

# Introduction

- VLM은 방대한 멀티모달 데이터로 학습되어 **고차원적 표현 능력**을 지니지만, **막대한 계산 비용과 메모리 요구량** 때문에 실시간 응용이나 엣지 디바이스에서는 사용이 현실적으로 어려움
- **가벼우면서도 정확한 모델을 만드는 방법**이 여전히 중요한 과제

VL2Lite는 이러한 한계를 해결하기 위해 다음과 같은 접근을 제시합니다:

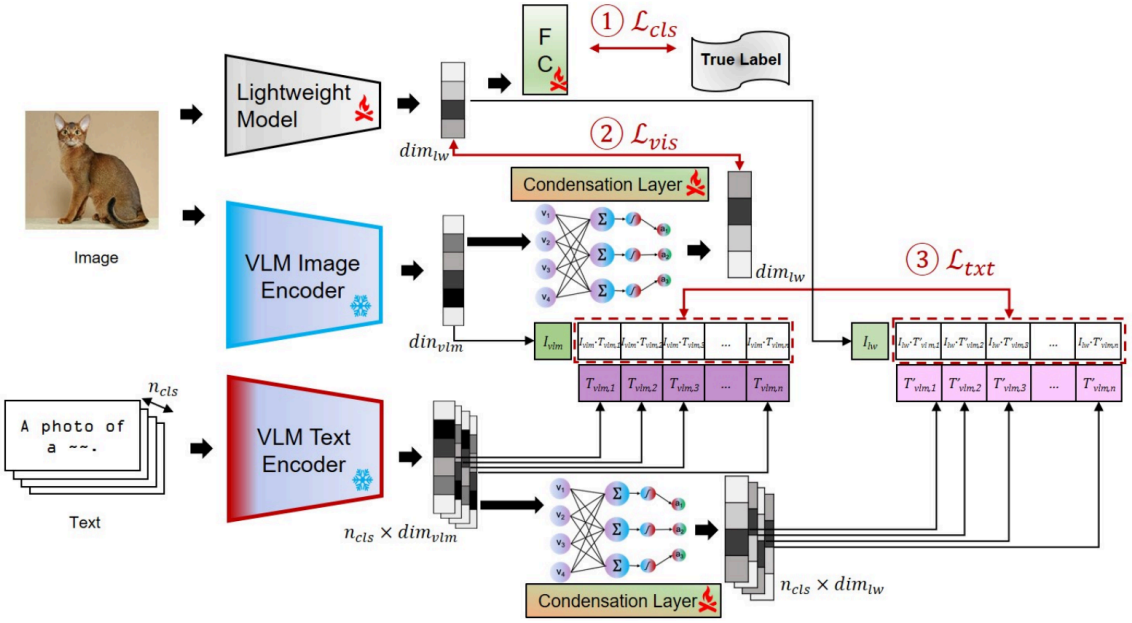
1. **One-step Knowledge Distillation**
    - 별도의 Teacher 학습 없이, 이미 사전 학습된 VLM(예: CLIP 계열)으로부터 멀티모달 지식을 **직접 Student에게 전이**합니다.
    - 즉, **VLM → Lightweight Network**로 바로 증류하며, 학습 과정이 단 한 단계로 단축됩니다.
2. **멀티모달 지식 통합**
    - VLM이 가진 **시각적(visual)** + **언어적(linguistic)** 표현을 동시에 활용합니다.
    - 이를 위해 **prompt engineering**을 사용해 의미 있는 텍스트 표현을 추출하고, 이를 Student 모델에 통합하여 시각 과제에서도 인식 성능을 향상시킵니다.
3. **지식 응축 계층 (Knowledge Condensation Layer)**
    - VLM의 고차원 표현을 Student 모델의 저차원 특성 공간에 맞게 변환하는 커스텀 계층을 도입했습니다.
    - 이 계층은 정보를 압축하면서도 의미를 보존하여 효율적인 지식 전달을 가능하게 합니다.
4. **합성 손실 함수 (Composite Loss)**

    VL2Lite는 학습 과정에서 다음 세 가지 손실을 동시에 최적화합니다:

    - **Task Loss**: 실제 분류 과제의 정답 예측 손실
    - **Visual KD Loss**: 이미지 표현 수준의 지식 증류 손실
    - **Linguistic KD Loss**: 텍스트 표현 수준의 지식 증류 손실

      이를 통해 **분류 학습**과 **지식 증류**를 한 번에 수행합니다.

# Method




## 증류 목표

- Large-Scale VLM이 가지고 있는 고퀄리티 정보를 사용해서 우리가 원하는 특정 task의 classification 성능을 높이자.
- 사용자가 딱 원하는 작은 도메인, 데이터가 제한된 특정 과제에서의 분류 성능을 높이자.
- 학습이 오래 걸리더라도 서비스 시점에서는 연산 비용을 작게 가져가자.

## 손실 함수

- L_cls: task-specific classification loss
  - lightweight model의 분류 성능
- L_vis: visual representation capability loss (KD)

where $\mathbf{I}'_{\text{vlm}}$ represents the transformed feature vector of the VLM after the condensation projection.

Similarly, the lightweight model produces its own feature representation:

$$\mathbf{I}_{\text{lw}} = \text{LW}(x), \tag{3}$$

where $\text{LW}(x)$ is the output of the lightweight model's encoder for input $x$.

We compute the pairwise Euclidean distances $d_{\text{lw}}$ in the lightweight model's feature space as:

$$d_{\text{lw}}(i,j) = \left\| \mathbf{I}_{\text{lw}}(x_i) - \mathbf{I}_{\text{lw}}(x_j) \right\|, \tag{4}$$

This metric reflects the spatial relationship between features as captured by the lightweight model.

Similarly, the pairwise Euclidean distances $d'_{\text{vlm}}$ between the transformed VLM feature vectors are:

$$d'_{\text{vlm}}(i,j) = \left\| \mathbf{I}'_{\text{vlm}}(x_i) - \mathbf{I}'_{\text{vlm}}(x_j) \right\|, \tag{5}$$

The visual knowledge distillation loss $\mathcal{L}_{\text{vis}}$ is defined as the sum of discrepancies between the pairwise distances of the lightweight model and those of the transformed VLM features, using the smooth L1 loss function for robustness against outliers:

$$\mathcal{L}_{\text{vis}} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} L1_{\text{smooth}}(d_{\text{lw}}(i,j) - d'_{\text{vlm}}(i,j)), \tag{6}$$

where the smooth L1 loss function $L1_{\text{smooth}}(\delta)$ is defined as:

- L_txt:

$$\cos_{\text{vlm}} = \frac{\mathbf{I}_{\text{vlm}}^{\top} \mathbf{T}_{\text{vlm}}}{\|\mathbf{I}_{\text{vlm}}\| \|\mathbf{T}_{\text{vlm}}\|}, \tag{10}$$

$$\cos_{\text{lw}} = \frac{\mathbf{I}_{\text{lw}}^{\top} \mathbf{T}'_{\text{vlm}}}{\|\mathbf{I}_{\text{lw}}\| \|\mathbf{T}'_{\text{vlm}}\|}, \tag{11}$$

The linguistic knowledge distillation loss function is formulated as:

$$\mathcal{L}_{\text{txt}} = T^2 \cdot \text{KL}\left(\text{softmax}\left(\frac{\cos_{\text{lw}}}{T}\right), \text{softmax}\left(\frac{\cos_{\text{vlm}}}{T}\right)\right), \tag{12}$$

# 모델 선정

## Teacher Model (VLM)

- ConvNeXt-XXLarge (OpenCLIP)
- ViT-Large

## Student Model

- ResNet-18, ResNet-34, EfficientNet-B0, MobileNet-V2, ShuffleNet-V2
- Swin Transformer-Tiny
- with ImageNet-pretrained
- SEED, DisCo

# Condensation Layer

- two-layer MLP + activation

# ETC

- A4000 * 4
- SEED/DisCo ResNet-50×2 백본으로 학습
- SSL + VLM KD: 모든 데이터셋에서 일관된 +2~7% 향상
- 소량의 데이터에서도 효과적
- 일반화된 task에서는 효과가 떨어짐
- L_txt loss가 제일 효과적 → inference 시 txt condensation 사용도 가능