CLIP и vision-language — краткий справочник

Одна идея, функции потерь, размеры моделей и рецепт для видео — на одной странице.

ОДНА ИДЕЯ

Vision-language модель помещает изображения и текст в одно общее пространство чисел (эмбеддингов). Совпадающие пары «картинка-подпись» оказываются рядом; несовпадающие — далеко. Близость = скалярное произведение двух строк.

КАК ОБУЧАЕТСЯ CLIP

• Два энкодера: для изображений (Vision Transformer) и для текста (Transformer).

• Обучен на 400 миллионах пар «картинка-подпись» из веба (датасет OpenAI WIT).

• Контрастно: для батча из N пар считаем все N x N комбинаций; диагональ вверх, остальное вниз.

• Симметричная кросс-энтропия с обучаемой температурой. Ручная разметка не нужна.

ZERO-SHOT КЛАССИФИКАЦИЯ

Размечаем изображение только по НАЗВАНИЯМ категорий, без обучающих примеров. Кодируем 'a photo of a {class}' для каждого ярлыка, кодируем картинку, берём ближайший. Лучший CLIP дал ~76% top-1 на ImageNet без меток ImageNet.

CLIP против SigLIP

	CLIP (2021)	SigLIP / SigLIP 2
Потери	Softmax (батч связан)	Sigmoid (пары независимы)
Большие батчи	Много памяти	Дёшево — батчи больше
Zero-shot ImageNet	~76% (ViT-L/14@336)	~79-81% (SigLIP 2 Base)
Статус 2026	Legacy / совместимость	По умолчанию в новых VLM

РАЗМЕРЫ МОДЕЛЕЙ (OpenAI ViT)

Модель	Разр.	Параметры	Эмбеддинг
ViT-B/32	224	88M	512
ViT-B/16	224	86M	512
ViT-L/14	224	304M	768
ViT-L/14@336	336	304M	768

ОТ КАДРА К ВИДЕО

Семплируем кадры (≈1/сек), кодируем каждый через CLIP, усредняем эмбеддинги в один эмбеддинг клипа, сравниваем с текстом. 600 кадров x 5 мс = ~3 с GPU, чтобы 10-минутный клип стал искомым. X-CLIP / ViFi-CLIP добавляют временной шаг, когда важен порядок кадров.

ЛОВУШКА

CLIP схватывает суть, а не детали. Он слаб в подсчёте, чтении текста, пространственной раскладке и отрицании ('без машин'). Используйте для грубого сопоставления, поиска и тегирования — не для точного чтения.