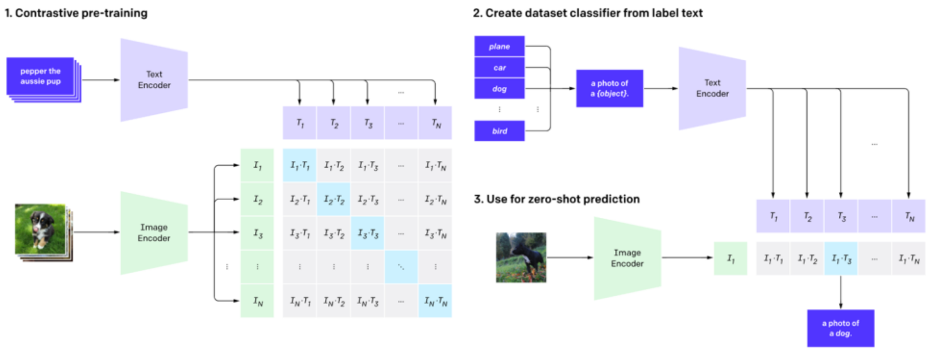
Open World Image Semantic Segmentation

- Introduction to **CLIP** (Contrastive Language-Image Pre-training)
- Aligns the CLS token of image and text encoder via InfoNCE
- Zero-shot image classification



$$\mathcal{L}_{\mathbf{x}} = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{e^{\phi(\mathbf{x_i},\mathbf{y_i})}}{\sum_{j=1}^{k} e^{\phi(\mathbf{x_i},\mathbf{y_j})}}\right)$$

$$\mathcal{L}_{\mathbf{y}} = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{e^{\phi(\mathbf{x_i},\mathbf{y_i})}}{\sum_{j=1}^{k} e^{\phi(\mathbf{x_j},\mathbf{y_i})}}\right)$$

$$f_v : \mathbb{R}^{C,H,W} \rightarrow \mathbb{R}^{D_v} \text{ and } f_t : \mathbb{R}^{l} \rightarrow \mathbb{R}^{D_t}$$

$$e_v : \mathbb{R}^{D_v} \rightarrow \mathbb{R}^{\bar{D}} \text{ and } e_t : \mathbb{R}^{D_t} \rightarrow \mathbb{R}^{\bar{D}}$$

$$\phi(\mathbf{x},\mathbf{y}) = \left(\frac{e_v(f_v(\mathbf{x}))}{|e_v(f_v(\mathbf{x}))|} \cdot \frac{e_t(f_t(\mathbf{y}))}{|e_t(f_t(\mathbf{y}))|}\right)$$

$$\mathcal{L}_{\text{InfoNCE}} = 1/2(\mathcal{L}_{\mathbf{x}} + \mathcal{L}_{\mathbf{y}}).$$

Attentive Mask CLIP

A-CLIP

- Main idea:
  - Drop irrelevant tokens!!
  - Drop almost >= 50%
- Random dropping degrades performance though
  - Masked Image Modeling (MIM) works with random dropping but not CLIP, why?

Okay, one more! Look at that tail! Not breeding plumage but everyday wear for the adult bird.

Drives at IITA Ibadan, Nigeria showing the guest building at the background

Statue of Schiller in Schiller Park, German Village, Columbus, Ohio

Mystery solved: it's a people-counting camera balloon!

"Crown Mary" at New Holland Dock

Santa Barbara Chalk Festival

The fountain in front of The Kansas City Star at 18th and Grand

This is Alison a few days after her cleft palate repair - wearing the much dreaded arm restraints

Strong correlation to text

Attentive Mask

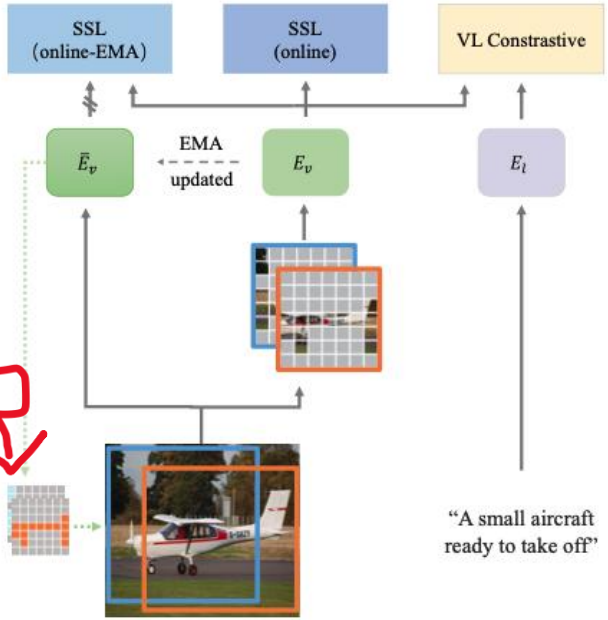
$$s_P = \frac{1}{HL}\sum_{l=1}^{L}\sum_{h=1}^{H}\text{Softmax}\left(\frac{\mathbf{f}_{lh}^{q}(CLS)\cdot\mathbf{f}_{lh}^{k}(P)}{\sqrt{C}}\right)$$

token

importance of token

where l denotes the layer index; h denotes the attention head index; $\mathbf{f}_{lh}^{q}(CLS)$ denotes the query embedding of the $[CLS]$ token at Layer l and Head h; $\mathbf{f}_{lh}^{k}(P)$ denotes the key embedding of Layer l and Head h for an image token at location P; C is the number of channels for the query and key embedding.

- Low strategy - discard low s_p score, high strategy: discard high s_p scores, mixture strategy
- EMA for attention score computation
  - A-CLIP-eff: half-res for speed as it does not need to be very accurate
- Shared EMA scored map for multiple views, efficiency
- Multiple masked views
  - Auxiliary contrastive learning
    - Between a masked view and EMA feature (BYOL)
    - Between masked views (SIMCLR)



Sigmoid Loss based CLIP

SGLIP

- Recall:

$$\mathcal{L}_{\mathbf{x}} = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{e^{\phi(\mathbf{x_i},\mathbf{y_i})}}{\sum_{j=1}^{k} e^{\phi(\mathbf{x_i},\mathbf{y_j})}}\right)$$

$$\mathcal{L}_{\mathbf{y}} = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{e^{\phi(\mathbf{x_i},\mathbf{y_i})}}{\sum_{j=1}^{k} e^{\phi(\mathbf{x_j},\mathbf{y_i})}}\right)$$

Inefficient and prevents large batch size (usually 32768, N positive pairs, NxN – N negative pairs)

SGLIP

- Instead of global normalization as in CLIP, separate pair processing
- A lot of negative pairs causing a lot of imbalance during optimization
    - Massive over-correction in the training results
- Added a bias term and temperature
- Z_{ij} is the label (1 for positive pair and -1 otherwise)
- Batch size can go up to a million, but typically 32k

$$-\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\sum_{j=1}^{|\mathcal{B}|}\underbrace{\log\frac{1}{1+e^{z_{ij}(-t\mathbf{x}_i\cdot\mathbf{y}_j+b)}}}_{\mathcal{L}_{ij}}$$

Allows for chunking

Attendance

Assignment Due 9/18

- Dataset: Human Action Recognition (HAR)
  - Same train/test split
- Fine tune on HAR and compare CLIP, A-CLIP and SGLIP
- Report performance, speed and memory consumption
  - SGLIP is able to load very large batch size, use 32k
  - For CLIP, A-CLIP (4096)
- Report, code, video to be submitted