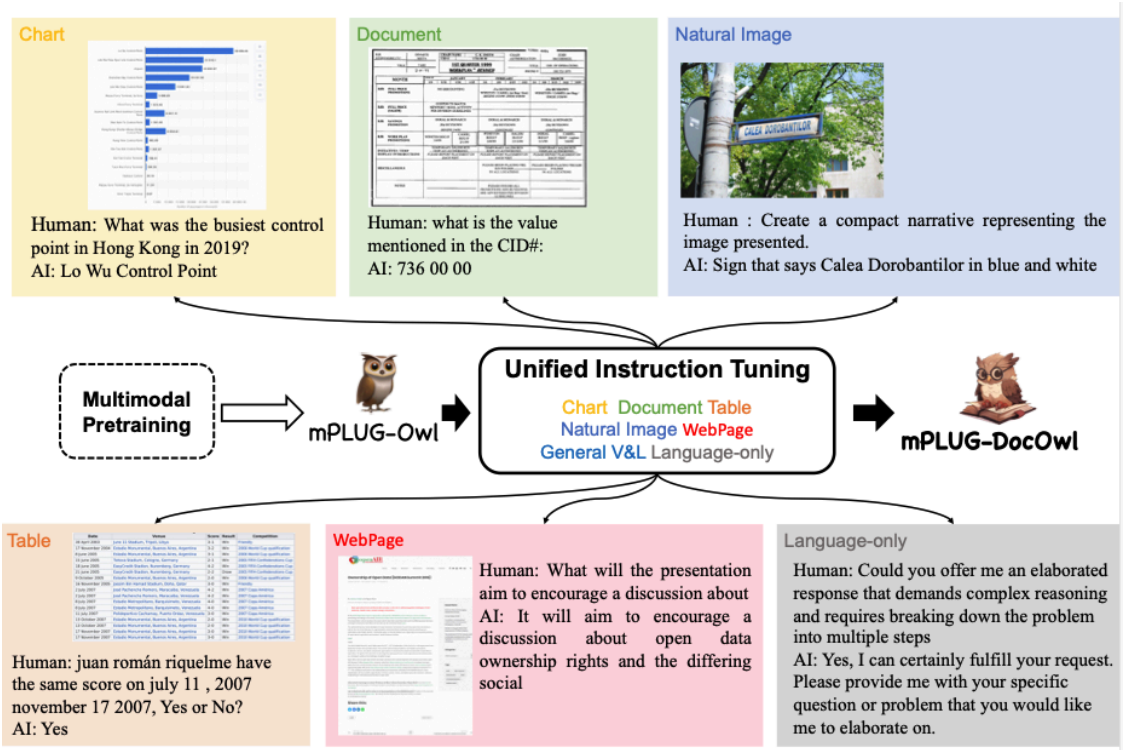
Enhancing Visual Document Understanding with Contrastive Learning in Large Visual-Language Models (DoCo)

Tencent YouTu Lab, 2024.02

문제정의: VLM 이라면 LLM에 vision encoder를 통해 정보를 전달할텐데, 이 때 문서에 대한 이해도가 높은 vision encoder를 만들어서 정보를 제공하자!!!

Introduction

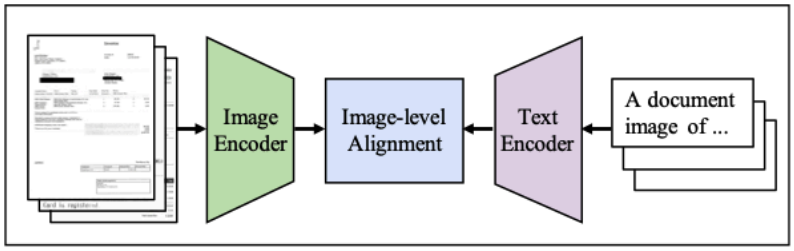
- 최근 LVLM (Large Visual-Language Models)가 등장하면서, VDU (Visual Document Understanding) 분야 연구가 활발하게 이루어지고 있음
  - mPLUG-DocOwl: document **image comprehension**에 특화 된 데이터셋 구축해서 학습



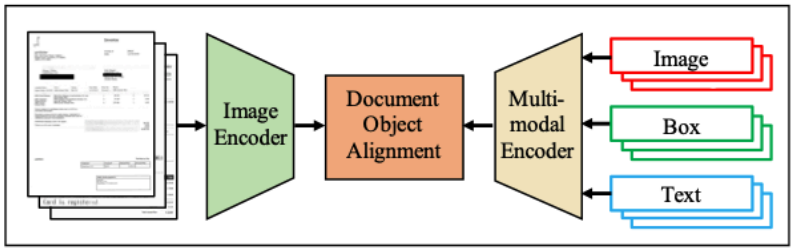
- LLaVAR: text-rich 이미지에서 text recognition에 특화 된 작업 학습

  - UniDoc: 대규모 멀티모달 데이터셋을 통합해서 학습
    - 이 모델은 텍스트 중심의 모델은 아닌 것 같음
- 하지만 이런 연구도 이미지 내 텍스트의 세부사항을 정확하게 포착하지는 못한다? □ 잘하는거같은뎅;;;ㅋㅋ;;..
- 왜 그렇냐면… 일단 이미지 내 텍스트를 디테일하게 분석해내지는 못했고… 보통 CLIP 기반으로 인코더를 만드는데, CLIP은 OCR 수준의 텍스트 분석보다는 이미지가 내포하는 객체 중심으로 학습된 모델이었음
- 그렇기 때문에, 오히려 CLIP 기반의 이미지 인코더를 이용해 VDU로 접근하는게 이상하고, VDU에 필요한 정확하고 세밀한 visual feature를 추출할 수 없음: fine-grained collapse issue

- 본 연구에서는 Document Object Constrastive Learning, DoCo 방법론을 제안함! □ CLIP 컨셉인데, 데이터가 VDU를 위한 데이터



(a) Image-level instance discrimination (CLIP)

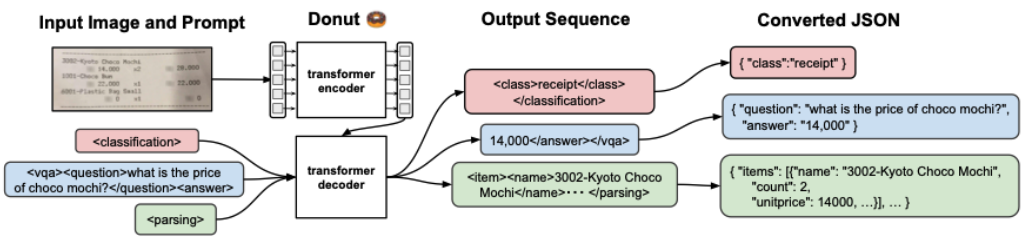


(b) Document object discrimination (Our DoCo)

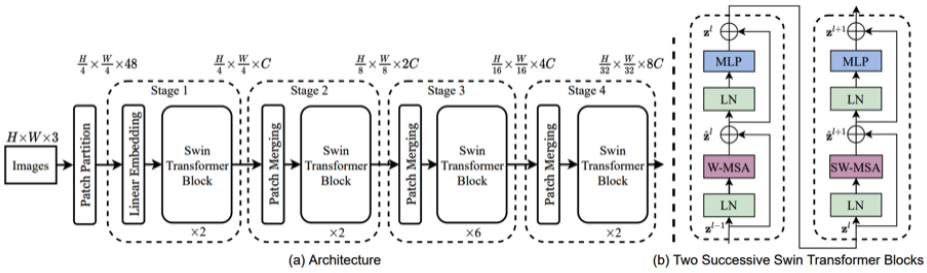
- 구체적으로는 멀티모달 인코더를 활용해서 문서 객체의 이미지, 레이아웃, 텍스트 등을 추출해서 image encoder와 constrastive learning 수행하는 방법!
- 텍스트 내용은 OCR 엔진을 활용
- 특히, 기존 VLM에서의 학습처럼… DoCo 방법론으로 pre-train 된 모델을 갈아끼우면 되니까, Plug-and-play 방식의 강점을 가질 수 있다.

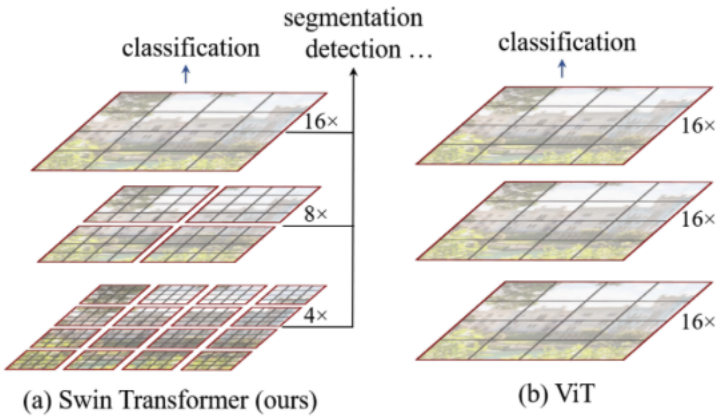
Related works

- VDU는 크게 2가지 파트로 나뉘어짐
  - 외부 OCR 엔진을 활용한 접근
    - LayoutLMv3: 이미지 패치 + OCR 간의 정렬을 이용해 학습
    - UDOP: 이미지와 텍스트 정보를 통합해서 임베딩
  - 외부 OCR 엔진 사용하지 않는 경우
    - Donut: 이미지에서 텍스트를 읽는 사전 학습 작업을 수행 (e2e model)



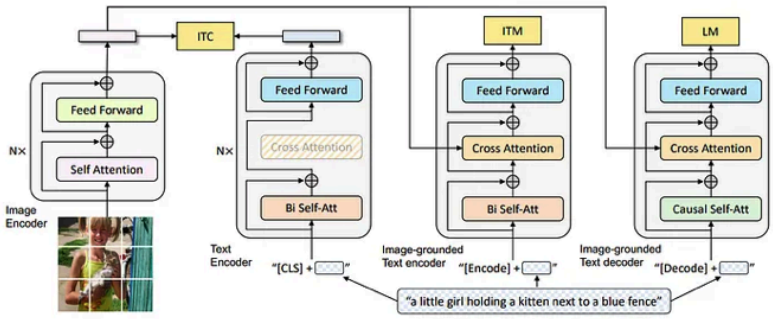
- Visual encoder를 사용해서 이미지를 패치로 자르고 Swin Transformer 사용





- VDU 관점에서는 char 단위와 문서의 layout을 함께 이해해야 하니까, 이런 구조로 인코딩하면 목적이 달성될 듯?

- 그래서 기존의 LVLM은…
  - 이미지 인코더를 두고.. 이미지 인코더의 출력 값과, LLM에 들어가는 token 간의 관계를 줄여줘서 일종의 token화 □ LLM이 학습한 토큰 형태로 변환됨을 유도하는 것



Architecture as in BLIP paper

Let's take a look at the pretraining objectives that is concerned with each of the modules:

1. *Image-Text Contrastive Loss(ITC Loss):* similar to CLIP, the encoders are trained to generate similar representations for similar image and text pairs and different representations for negative input pairs.
2. *Image-Text Matching Loss(ITM Loss):* this helps the text encoder to learn multi-modal representation that captures the fine-grained alignment between vision and language. This is a binary classification task where the loss outputs 1 for a positive image-text pair and 0 for a negative pair.
3. **Language Modeling Loss(LM Loss):** this loss helps the text decoder generate text descriptions for the corresponding input image.

-

- Pix2Struct: screenshot parsing을 컨셉

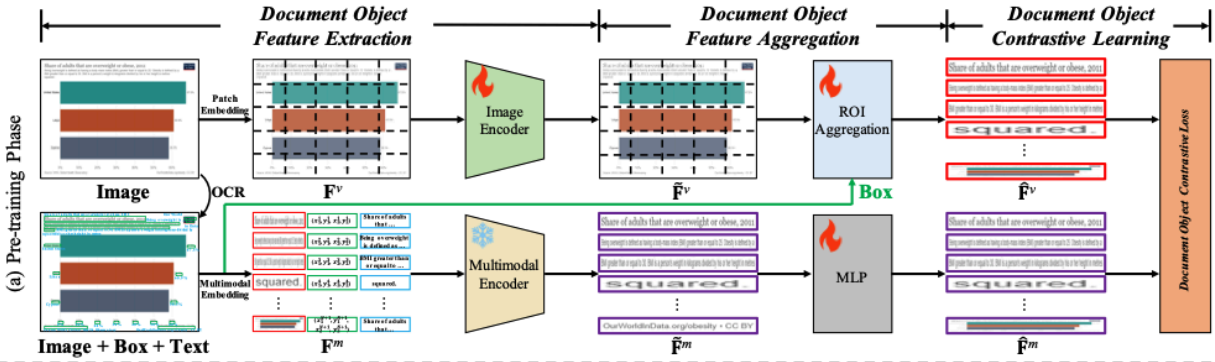
Ground-truth Parse

```
<<<<Coronavirus Update! We are open and ready to help you.>
   <We are conducting most of our appointments via phone to help prevent
    the spread of the virus.>>
  <Chapter 13 Coronavirus Update>>
 <<img_src=Logoo img_alt=Stamps & Stamps Attorneys At Law>
  <img_src=Phone>
  <Contact for a free Initial Consultation>
  <<Call Us> <(937) 247-6447>>
  <<Text Us> <(937) 265-6418>>>
 <<Home> <About> <Articles> <Videos> <Testimonials> <Tax Relief> <News>
  <Podcasts> <Rate Us> <Contact>>
 <<We can provide the guidance you need to get through stressful family>
  <disputes with your rights and interests intact.>>
 <<<img_src=Bankruptcy img_alt=Bankruptcy Overview>
   <<Bankruptcy> <Overview>>>
  <img_src=Criminal-Defense1
   img_alt=Criminal Defense & Traffic Offenses>>>
```

- Pixel 입력을 html로 치환하는 것을 목적으로 함
- 그냥.. ViT로.. (차이점은 해상도 조절할 때, 비율 왜곡 없이..)

Method

- Pre-training phase



- 일단, 두 종류의 인코더를 사용하는데 (1) 이미지 인코더, (2) 멀티모달 인코더
- (1) 이미지 인코더

- patch encoding 방식임
- 주어진 문서를 이미지로 바라보고 패치 수행
- 일반적인 ViT 모델 구조를 가지되, CLIP 모델의 weight로 initializing 하고 학습 수행
- ROI Aggregation
  - 이미지 인코딩 이후, 패치 간 관계성을 학습해야함!

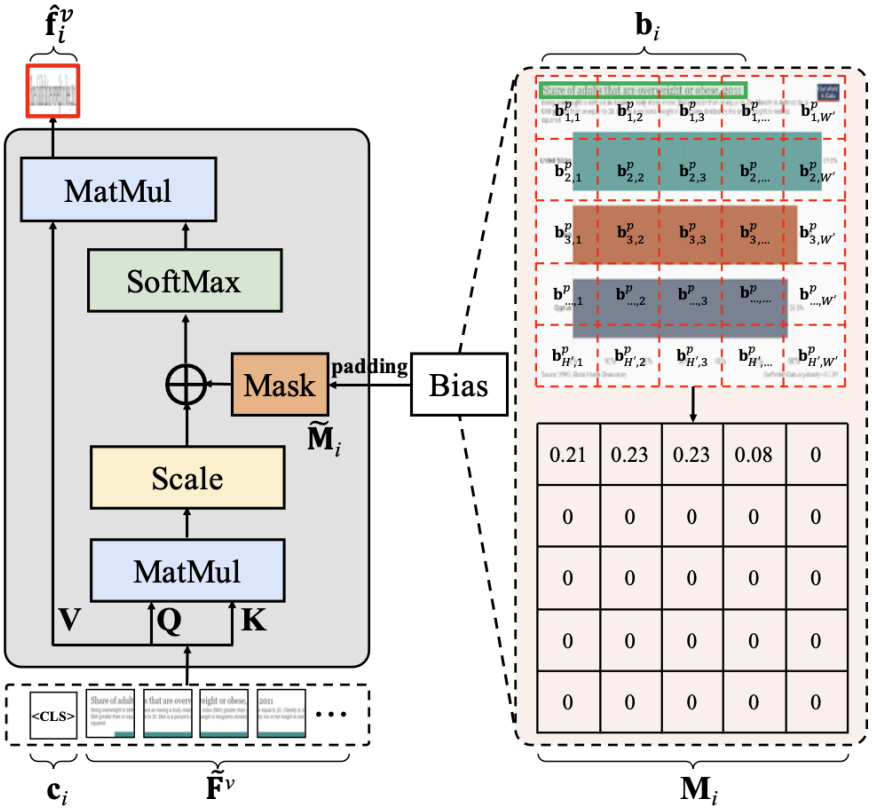
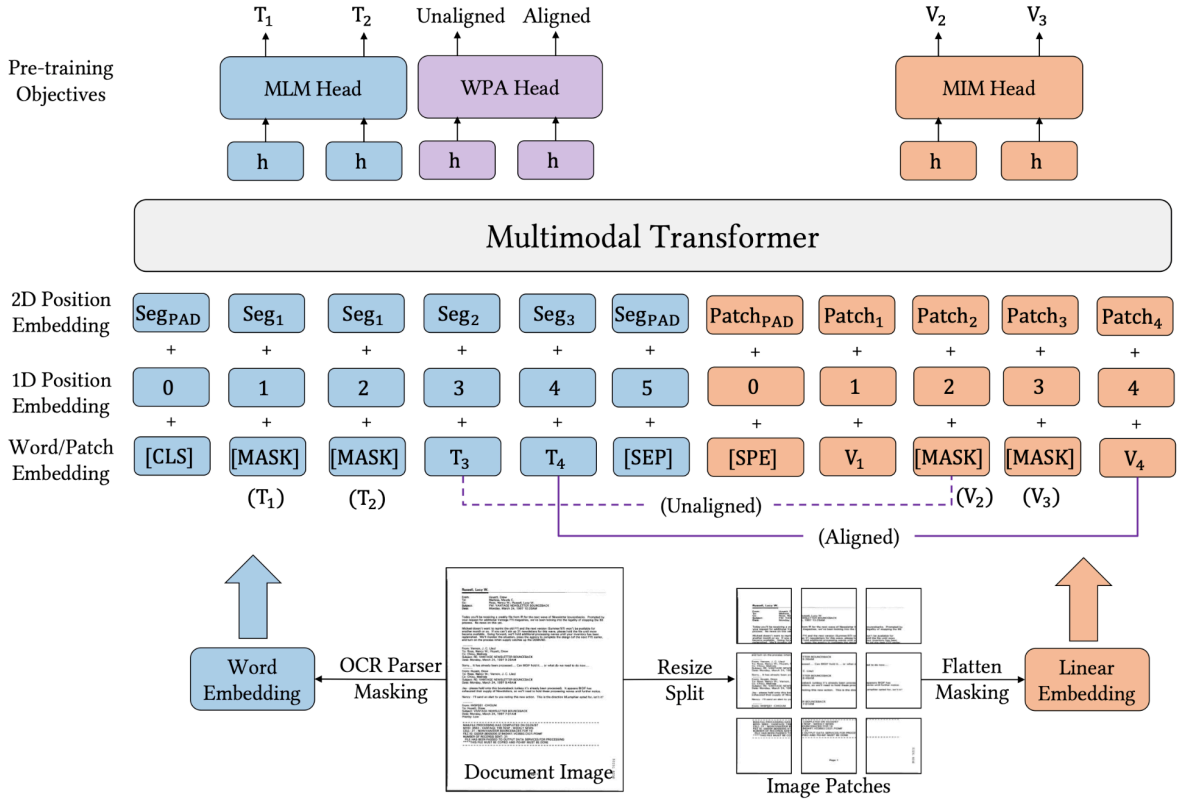


Figure 3. The proposed ROI Aggregation. The dashed red grid represents the image patch features, and the solid green region denotes the bounding box region. The overlap between each patch and the given region is calculated and serves as the attention mask for the visual aggregation of the document objects. Best viewed in color.

- 패치와 객체(OCR로 나온 box)는 서로 어긋날 수 있음
- i 번째 패치가 가지는 객체 1,2, 3 .. 간의 overlap을 살펴봄
- 패치 별로 어느정도의 면적이 객체와 겹치는지 비율을 계산함 → overlap 비율

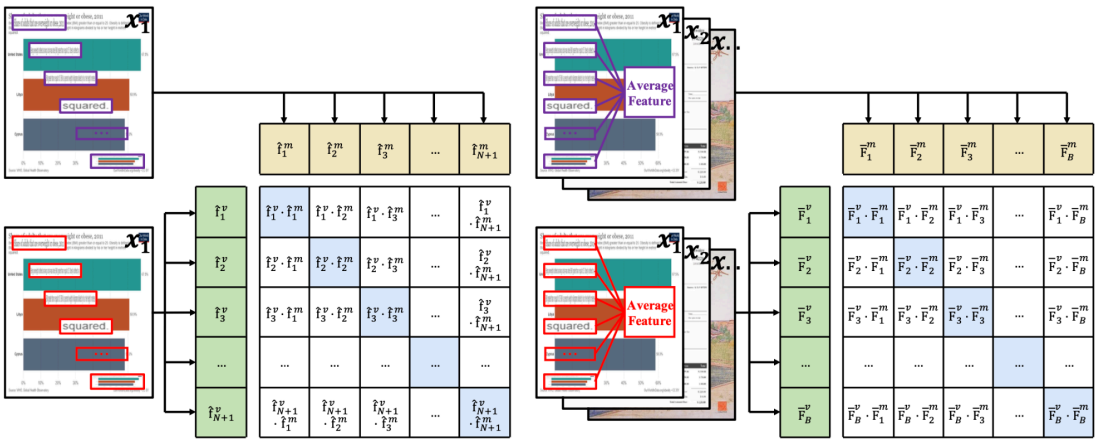
- 이 overlap의 비율 값을 attention mask로 정의하여, 많이 겹치는 비율을 가진 패치에는 더 높은 가중치를 부여함 → bias를 만들어냄
- 즉, 일반적인 gpt로 생각한다면, next token을 안보여주는 masking을 계단식으로 순차적 적용하는데, 비슷한 너낌으로 겹침 정도가 큰 패치에 높은 가중치를 주고, 나머지는 무시하도록 학습을 유도하는 것

- (2) 멀티모달 인코더
  - OCR을 돌려서 나온 {문장 이미지 + bounding box (x1, y1, x2, y2) + 텍스트(OCR 결과)}
  - LayoutLMv3의 multi-modal transformer를 사용했다



  - 일단 텍스트는.. RoBERTa의 word embedding을 initial로 사용하고
  - 2D position embedding은 bounding box를 의미
  - 이미지는 linear projection을 해서 feeding 했다고 함
  - 각 이미지는 visual vocabulary를 통해서 각각의 픽셀은 독립된 token으로 tokenizing 됨
  - 학습은
    - masked language modeling (span masking)
    - masked image modeling

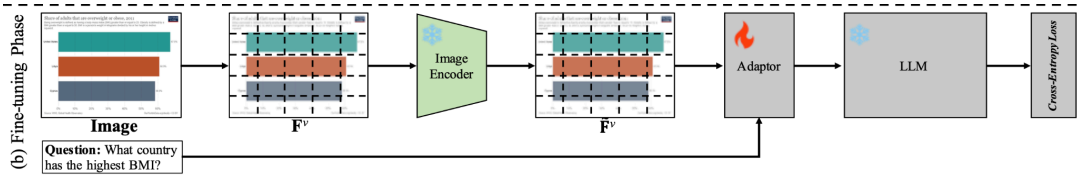
        - BEiT에서 따왔다는데…
        - 이미지 토큰의 40%를 마스킹하고, 이걸 reconstruction하는 걸 목적으로 학습
    - word-patch alignment
        - 이미지 패치랑 word vector 간의 상관관계를 학습시키기 위함
        - aligned 됐는가, 안됐는가를 분류하는 모델을 부착
        - aligned는 텍스트랑 이미지 둘 다 unmasked일 때 할당하고
        - unaligned는 텍스트가 마스크되면 할당 → 모델이 마스크 된 텍스트 단어와 이미지 패치 간의 관계는 학습 안되도록
        - 2개짜리 MLP 레이어를 넣고 이진분류로 학습
- Document Object Contrastive Learning
    - 이제 이미지와 텍스트 정보(멀티모달)를 합치는 학습이 필요함
    - (1) Intra-DoCo
        - 문서를 이미지로 관찰
        - 시각 정보와 멀티모달 정보 사이의 관계성을 학습하는 것을 목적
        - 즉, 이미지 인코더를 통해 출력 된 임베딩 벡터와, 멀티모달 인코더를 통해 추출 된 임베딩 벡터 간의 유사도를 최대로 하도록 학습함
    - (2) Inter-DoCo
        - 서로 다른 이미지 간의 관계를 학습하는 모델
        - 배치 B에 속해있는 이미지 n개
        - 각 이미지 마다 속해있는 객체의 평균값: 문서 1개에 대한 전역 벡터 → 이미지 벡터, 멀티모달 벡터 획득
    - DoCo의 loss는 위 두 학습 방법론의 합으로 구성



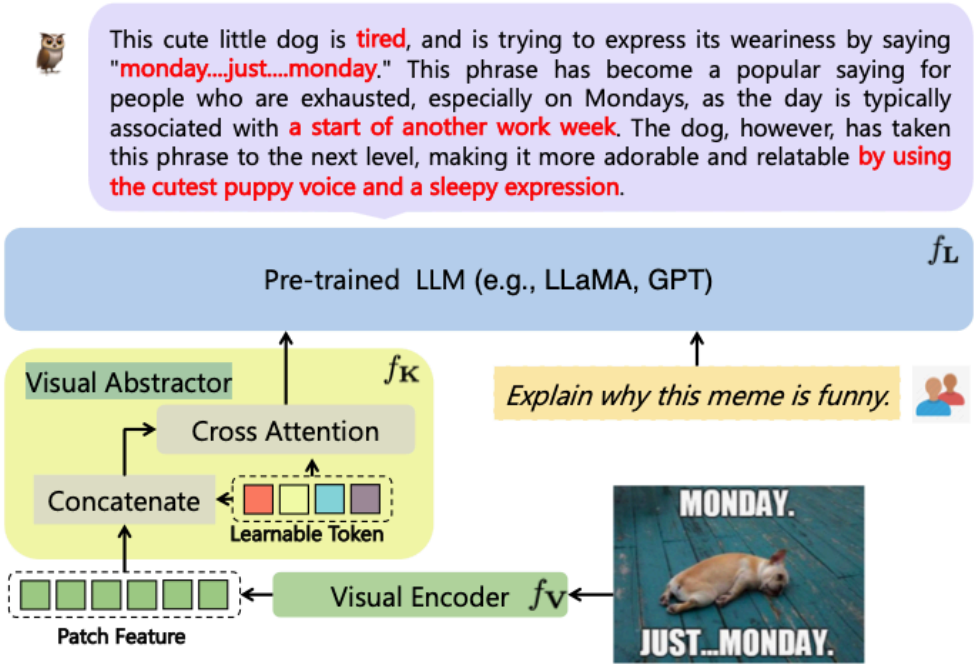
(a) Intra-DoCo (b) Inter-DoCo

- Training phase

- pre-training
  - 위에서 언급한대로.. 근데 multi-modal 인코더는 freezing 시켜둠
- fine-tuning



  - Qwen-VL-Chat, mPLUG-Owl 모델을 백본으로 사용하고
  - image encoder를 떼다가 활용



  - Learnable Token → nn.embedding:
  - token의 개수가 100개고, patch feature 벡터가 768이면
  - 100, 768

실험 결과는 논문으로 함께 봐요