AI504 19강 정리

- Image to Text
  - Seq2seq: 텍스트 입력, 텍스트 출력
  - Image to sequence: 이미지 입력, 텍스트 출력
    - Ex) 주어진 이미지를 텍스트로 묘사하기(Image Captioning)
  - Encoder – Decoder
    - Seq2seq
      - Encoder: RNN
      - Decoder: RNN
    - Image to sequence
      - Encoder: CNN
      - Decoder: RNN
- Show and Tell
  - A Neural Image Caption Generator
  - 처음으로 도메인 지식없이 Neural Image Caption을 한 논문
    - No Object Detection, Language Modeling, Description Templates
    - Not text ranking, but pure generation
  - 구조
    - 이미지 입력 -> CNN -> RNN -> 텍스트 출력
    - RNN의 첫 입력 <Start>전에 -1번째 입력으로 CNN의 출력을 넣음
    - RNN hidden layer의 initialization은 0으로 함
- Show, Attend and Tell
  - Neural Image Caption Generation with visual Attention
  - 이미지 입력 -> CNN -> RNN with Attention over image -> 단어 단위로 생성하여 텍스트 출력
  - 모델이 단어를 생생할 때 이미지의 관련 있는 부분에 "attending"한다

- Context Vector c_i
  - Seq2seq에서는 encoder의 hidden layer들과 h_{i-1}에 의해 결정됨
  - Show, Attend and Tell에서는 encoder가 RNN이 아니므로 hidden layer가 없음
    - CNN중간의 Convolution Layer -> output이 n*n*channel 형태 (VGG16에서 9번째 conv layer, output: 14*14*512)
    - 해당 레이어의 output의 n*n 부분을 flatten하여 n^2*channel로 바꿈 (196*512) -> ①
    - 이 형태와 Seq2seq encoder의 hidden layer의 형태가 유사함
    - => ①과 h_{i-1}에 의해 c_i가 결정됨
- Technical detail
  - RNN의 initial hidden state, h_0는 미리 학습되어 있음
    - ①의 평균을 구해 512차원의 벡터로 나타낸 것을 MLP에 넣어 만듦

- Text to Image
  - 텍스트로 조건이 주어졌을 때 GAN을 이용해 이미지를 생성
  - 구조
    - RNN으로 텍스트를 encode
    - GAN으로 code를 이미지로 decode(generate)
      - Deconvolution을 사용해 upsample (DC-GAN)
    - GAN의 학습을 위해 decoder 뒤에 Discriminator Network 추가
  - 학습방법
    - Discriminator의 일이 복잡함
      - 진짜(높은 품질의) 이미지 + 알맞은 텍스트 -> Real (s_r)
      - 가짜(낮은 품질의) 이미지 + 알맞은 텍스트 -> Fake (s_w)
      - 진짜(높은 품질의) 이미지 + 틀린 텍스트 -> Fake (s_f)
      - 가짜(낮은 품질의) 이미지 + 틀린 텍스트 -> Fake
    - Discriminator에 들어가는 케이스는 위의 세 가지

- 첫 번째는 1, 나머지는 0이 나와야 함
- => Loss = $\log(s_r) + (\log(1 - s_w) + \log(1 - s_f)) / 2$ (1과 0의 비율을 맞추기 위해 뒤의 둘은 더해서 반으로 나눔)