CV이론 복습 문제 답안

1. FCN(Fully Connected Network)의 단점을 설명하고 CNN의 장점을 설명해주세요

Fully Connected Network(완전 연결 신경망)는 **모든 입력 뉴런이 모든 출력 뉴런과 연결되어 있어 파라미터 수가 매우 많아지는 단점**이 있습니다. 이로 인해 **계산 비용이 증가**하고, **과적합(overfitting)의 위험**이 커지며, 특히 **이미지와 같은 공간적 정보를 효과적으로 처리하지 못합니다**.

반면, Convolutional Neural Networks(CNN)는 **가중치 공유와 지역적 연결성을 통해 파라미터 수를 줄이면서도 공간적 패턴을 효과적으로 학습**할 수 있습니다. CNN은 계층적 특징 추출을 통해 **저수준부터 고수준까지 다양한 특징**을 효과적으로 학습하여 이미지 분류, 객체 인식 등에서 높은 성능과 효율성을 발휘합니다.

채점 기준:

FCN에서는 픽셀 단위, 이미지의 공간적 정보 반영이 어려움, 오버피팅 위험, 계산비용이 큼 -> 이 내용들 중 2개 이상을 정확히 작성하실 수 있으면 pass!

CNN에서는 low level ~ high level feature를 효과적으로 학습 가능하다, 필터를 통해 입력 이미지에서 관계성 있는 패턴 학습이 가능하다, 파라미터 수를 줄이면서도 효과적인 학습이 가능하다, 이미지 처리에서 높은 정확도 -> 이 내용들 중 2개 이상을 정확히 작성하실 수 있으면 pass!

---

2. CNN에서 'Receptive field'는 무엇을 의미하는지 설명해주세요

'Receptive field'(수용 영역)는 CNN의 특정 뉴런이 **입력 이미지의 어떤 영역을 참조하여 활성화되는지를 나타내는 개념**입니다. 이는 **해당 뉴런이 관찰할 수 있는 입력 데이터의 공간적 범위**를 의미하며, **네트워크의 깊이, 필터의 크기, 스트라이드(stride), 패딩(padding) 등의 요소에 의해 결정**됩니다. 예를 들어, 초기 층에서는 작은 영역만을 관찰하지만, 네트워크가 깊어질수록 더 큰 영역을 참조하게 됩니다. **큰 receptive field는 이미지의 전역적 패턴이나 객체 간의 관계를 인식하는 데 유리**하며, 이는 **복잡한 이미지 분석 작업에서 중요한 역할**을 합니다. 따라서, **receptive field의 크기는 모델이 학습할 수 있는 공간적 정보의 범위를 결정짓는 중요한 요소로 작용**합니다.

채점 기준:

뉴런이 입력 이미지의 어떤 영역을 참조하여 활성화되는지를 나타냈다는 내용
뉴런이 관찰할 수 있는 입력 데이터의 공간적 범위라는 내용
네트워크 깊이, 필터 크기, 스트라이드, 패딩 등에 의해 결정된다는 내용
receptive field의 크기는 모델이 학습할 수 있는 공간적 정보의 범위를 결정짓는 요소로 작용한다는 내용

이 내용들 중 2개 이상을 정확히 작성하실 수 있으면 pass!

---

3. 깊은 신경망의 기울기 소실 및 폭발 문제 최소화 방법과 원리를 서술해주세요

방법: 깊은 신경망에서 기울기 소실(vanishing gradients)과 기울기 폭발(exploding gradients) 문제를 최소화하기 위한 대표적인 방법 중 하나는 Residual Networks(ResNet)를 사용하는 것입니다. ResNet은 네트워크의 깊이가 깊어질수록 발

생할 수 있는 학습의 어려움을 극복하기 위해 고안된 구조입니다.

원리: ResNet은 skip connection을 도입하여 입력 신호가 네트워크의 여러 층을 직접 통과하도록 합니다. 이러한 skip connection은 입력과 출력 사이에 직접적인 경로를 제공함으로써 기울기가 역전파되는 과정에서 소실되거나 폭발하는 것을 방지합니다. 즉, 네트워크가 깊어지더라도 신호가 직접 전달되어 기울기의 크기가 일정하게 유지되며, 이는 안정적인 학습을 가능하게 합니다. 이러한 구조 덕분에 ResNet은 매우 깊은 네트워크에서도 효과적으로 학습할 수 있으며, 기울기 소실 및 폭발 문제를 크게 완화할 수 있습니다.

채점 기준:

Residual Network, skip connection 을 정확히 포함
skip connection 이 입출력 사이에 직접적인 경로를 제공해 기울기 소실 및 폭발을 방지한다고 작성

위 내용을 정확히 작성하실 수 있으면 pass!

---

4. Transformer의 Multi-head attention은 구조 내에서 어떤 역할을 수행하는지 서술해주세요

https://www.youtube.com/watch?v=eMlx5fFNoYc&t=2s (참고하시면 좋습니다)

Transformer의 Multi-head Attention은 모델이 **입력 데이터의 다양한 부분 간의 관계를 동시에 학습할 수 있도록 하는 핵심 메커니즘**입니다. 이는 **여러 개의 독립적인 Attention 헤드를 병렬로 운영**하여, **각 헤드가 입력의 다른 부분에 집중할 수 있게 합니다**. 예를 들어, **하나의 헤드는 문장의 문법적 관계를, 다른 헤드는 의미적 관계를 학습**할 수 있습니다. 이러한 **다중 헤드 구조는 모델이 다양한 표현 공간에서 정보를 추출하고 통합할 수 있게 하여, 더 풍부하고 다층적인 특징을 학습할 수 있도록 합니다**. 결과적으로, Multi-head Attention은 Transformer가 **복잡한 패턴과 관계를 효과적으로 이해하고 처리할 수 있게 해주며, 이는 자연어 처리, 번역, 이미지 처리 등 다양한 분야에서 높은 성능을 발휘하는 데 기여**합니다.

채점 기준:

입력 데이터의 다양한 부분 간의 관계를 동시에 학습할 수 있도록 한다
여러 개의 독립적인 어텐션 헤드가 병렬 운영된다
각 헤드가 입력의 다른 부분에 집중하도록 한다
멀티 헤드 구조를 통해 모델이 다양한 표현 공간에서 정보를 추출, 통합 -> 풍부한 특징 학습

위 내용을 정확히 작성하실 수 있으면 pass!

---

5. Transformer는 기존의 CNN과 어떤 점에서 다른지 inductive bias 관점에서 설명해주세요

Transformer는 기존의 Convolutional Neural Networks(CNN)과는 근본적으로 다른 접근 방식을 채택합니다. **Transformer는 Self-Attention 메커니즘을 사용하여 입력 데이터의 전역적 관계를 동시다발적으로 학습**합니다. 이는 인간의 inductive bias, 즉 학습 과정에서 자연스럽게 형성되는 가정이나 선호도와 비교할 때, **특정 구조적 가정을 최소화**하는 점에서 차별화됩니다.

CNN은 공간적 패턴에 대한 가정, 즉 이미지의 국부적인 특징을 중시합니다. 이러한 접근 방식은 인간이 시각적 패턴을 인식하거나 시간적 흐름을 이해하는 방식과 유사한 inductive bias를 반영합니다.

반면 Transformer는 이러한 특정한 구조적 가정보다는 데이터 간의 관계를 자유롭게 학습할 수 있도록 설계되어, 보다 유연하고 일반화된 패턴 인식이 가능합니다. 이는 인간이 특정 방식에 얽매이지 않고 다양한 패턴을 인식하는 능력과 유사한 면을 가지고 있습니다.

채점 기준:

어떤 테스크를 진행할 때 CNN은 특정 테스크에 특화된 작업을 수행하도록 인간이 설정한 문제 풀이 방식입니다. 반면 Transformer는 대량의 데이터 내에서 모델이 데이터 간 관계를 파악하기에 보다 Global한 정보를 추출할 수 있죠.

Inductive bias 관점에서 CNN에 비해 Transformer는 고유한 inductive bias가 부족합니다. 그렇기에 충분하지 못한 양의 데이터로 학습하면 일반화가 잘 이루어지지 않죠

CNN = 특정 테스크 특화(이미지 분류 등), 이외 테스크에는 적용이 어려움(방법론이 정해짐)

Transformer = 데이터 전체의 관계를 학습, 이미지 분류 및 자연어 처리 등 다양한 테스크에 활용됨

위 내용을 작성하실 수 있으면 pass!

---

6. Positional embedding이란 무엇인지 설명해주세요

Positional Embedding은 Transformer와 같은 모델에서 입력 시퀀스의 각 요소에 위치 정보를 부여하여, 모델이 순서 정보를 인식할 수 있도록 하는 기법입니다. Transformer는 기본적으로 순서에 대한 정보를 내재적으로 가지지 않기 때문에, 입력 데이터의 순서를 모델이 이해할 수 있도록 도와주는 역할을 합니다.

이러한 위치 정보는 주로 사인-코사인 함수(sine-cosine functions)를 사용하여 생성되거나, 학습 가능한 매개변수로 구현됩니다. 사인-코사인 함수를 사용하는 방법은 각 위치에 대해 고유한 주기성을 부여하여, 모델이 위치 간의 관계를 쉽게 학습할 수 있도록 합니다. 학습 가능한 방식은 모델이 데이터에 최적화된 위치 정보를 직접 학습하도록 하여, 더 유연한 위치 인코딩을 가능하게 합니다. 결과적으로, Positional Embedding은 Transformer가 입력 시퀀스의 순서를 효과적으로 인식하고, 이를 기반으로 더 정확한 예측과 분석을 수행할 수 있게 합니다.

채점 기준:

입력 시퀀스의 각 요소에 위치 정보를 부여한다
모델이 순서 정보 인식할 수 있도록 한다
사인-코사인 함수 활용한다
주기성을 부여하여 위치 관계를 학습할 수 있도록 한다
입력 시퀀스의 순서를 효과적으로 인식하여 더 정확한 예측 및 분석이 가능하다

위 내용 중 2개 이상 작성하실 수 있으면 pass!

---

7. Classification token은 어떤 역할을 수행하는지 설명해주세요

Classification Token([CLS] 토큰)은 Transformer 기반 모델, 특히 BERT와 같은 모델에서 입력 시퀀스의 시작에 추가되는 특별한 토큰입니다. 이 토큰은 전체 입력 시퀀스의 종합적인 표현을 학습하는 역할을 담당합니다. 모델은 [CLS] 토큰을 통해

입력 데이터의 전체적인 의미와 맥락을 파악하게 되며, 이를 기반으로 분류 작업을 수행합니다.

예를 들어, 문장 분류(task)에서는 [CLS] 토큰의 최종 출력 벡터가 문장의 전체적인 특징을 요약한 벡터로 사용되며, 이를 통해 문장이 어떤 카테고리에 속하는지를 예측하게 됩니다. 이와 같은 방식으로 [CLS] 토큰은 입력 시퀀스의 통합된 정보를 효과적으로 추출하고, 이를 바탕으로 다양한 분류 작업에 활용될 수 있습니다. 따라서, [CLS] 토큰은 모델이 입력 데이터의 전반적인 의미를 이해하고, 이를 기반으로 정확한 예측을 수행하는 데 중요한 역할을 합니다.

채점 기준:

입력 시퀀스의 시작에 추가되는 특별한 토큰이다
CLS 토큰을 통해 입력 데이터의 전체적인 의미와 맥락을 파악한다
입력 시퀀스의 통합 정보를 효과적으로 추출하고 이를 바탕으로 다양한 작업에 활용한다

위 내용 중 2개 이상 작성하실 수 있으면 pass!

---

8. CAM(Class activation mapping)에서 FC layer 대신 어떤 레이어를 사용하나요?

Class Activation Mapping(CAM)에서는 전통적인 Fully Connected(FC) 레이어 대신 Global Average Pooling(GAP) 레이어를 사용합니다. GAP 레이어는 각 특징 맵(feature map)의 평균 값을 계산하여, 클래스별로 중요한 영역을 추출하는 데 효과적입니다. FC 레이어는 모든 뉴런이 서로 연결되어 있어 많은 파라미터를 필요로 하고, 공간적 정보를 유지하지 못하는 반면, GAP 레이어는 각 특징 맵의 공간적 정보를 보존하면서도 파라미터 수를 크게 줄일 수 있습니다.

CAM에서는 GAP 레이어를 통해 추출된 평균 값과 각 클래스에 대한 가중치를 결합하여, 입력 이미지에서 특정 클래스와 관련된 중요한 영역을 시각화할 수 있습니다. 이러한 접근 방식은 모델이 어떤 부분에 주목하여 결정을 내리는지를 직관적으로 이해할 수 있게 해주며, 모델의 해석 가능성을 높이는 데 기여합니다. 결과적으로, GAP 레이어는 CAM에서 공간적 정보를 효과적으로 활용하면서도 효율적인 계산을 가능하게 합니다.

채점 기준:

CAM 에서는 Global Average Pooling 사용한다는 내용이 있다면 pass!

---

9. ViT에서 CLS 토큰은 무엇인지 설명해주세요

Vision Transformer(ViT)에서 CLS 토큰은 입력 이미지의 패치 시퀀스의 맨 앞에 추가되는 특별한 토큰으로, 전체 이미지의 종합적인 특징을 학습하는 역할을 합니다. ViT는 이미지를 고정된 크기의 패치로 분할한 후, 각 패치를 일렬로 나열하여 Transformer의 입력으로 사용합니다. 이때, CLS 토큰을 추가함으로써 모델이 전체 이미지의 전반적인 정보를 요약하고 이를 기반으로 최종 분류 작업을 수행할 수 있게 합니다.

CLS 토큰의 최종 출력 벡터는 이미지 전체의 특징을 집약한 표현으로 사용되며, 이는 분류기(classifier)에 입력되어 이미지가 속하는 클래스를 예측하게 됩니다. 이러한 방식은 자연어 처리에서 [CLS] 토큰이 문장의 전체적인 정보를 요약하는 역할을 하는 것과 유사합니다. CLS 토큰을 통해 ViT는 이미지의 전역적인 특성을 효과적으로 학습하고, 이를 기반으로 높은 정확도의 이미지 분류를 수행할 수 있습니다.

채점 기준:

입력 이미지의 패치 시퀀스 맨 앞에 추가된다
전체 이미지의 종합 특징을 학습한다
CLS 토큰의 최종 출력 벡터는 이미지 전체의 특징을 집약한 것이다
classifier에 입력으로 들어가 이미지가 어느 클래스에 속하는지 예측한다

위 내용 중 3개 이상 작성하시면 pass!

---

10. Fully connected layer와 Fully convolutional layer의 차이를 서술해주세요

Fully Connected Layer(완전 연결층): Fully Connected Layer는 모든 입력 뉴런이 모든 출력 뉴런과 연결되는 구조를 가지고 있습니다. 이는 고차원적인 특징을 학습하는 데 강력한 능력을 발휘하지만, 파라미터 수가 매우 많아져 계산 비용이 증가하고, 과적합(overfitting)의 위험이 높아집니다. 또한, 공간적 정보를 유지하지 못하고 단순히 벡터화된 입력을 처리하기 때문에 이미지의 공간적 구조나 패턴을 반영하는 데 한계가 있습니다. 이러한 이유로, 주로 네트워크의 마지막 부분에서 분류 작업을 수행할 때 사용됩니다.

Fully Convolutional Layer(완전 합성곱층): 반면, Fully Convolutional Layer는 공간적 구조를 유지하면서 지역적 특징을 학습할 수 있는 장점을 가지고 있습니다. 이는 합성곱 필터를 사용하여 입력 데이터의 공간적 관계를 반영하며, 입력 크기에 유연하게 대응할 수 있습니다. Fully Convolutional Layer는 주로 이미지 세분화(segmentation)와 같은 작업에서 사용되며, 공간적 정보를 보존하면서도 효율적으로 특징을 추출할 수 있습니다. 또한, 파라미터 수가 상대적으로 적어 계산 비용이 낮고, 다양한 입력 크기에 대해 일관된 성능을 발휘할 수 있습니다. 따라서, Fully Convolutional Layer는 공간적 정보를 유지하면서도 효율적인 학습이 필요한 다양한 컴퓨터 비전 작업에서 중요한 역할을 합니다.

채점 기준:

Fully Connected Layer ; 모든 입력 뉴런이 모든 출력 뉴련과 연결, 파라미터 수 매우 많다, 계산 비용 크다, 과적합, 공간 정보를 유지 못함

Fully Conv Layer: 공간 구조를 유지하면서도 지역적 특징 학습, 합성곱 필터 사용, 파라미터 수 상대적으로 적음, 효율적 학습

위 내용들이 2개 이상씩 들어가면 pass!

---

11. Zero-shot과 Few-shot에 대해 각각 설명해주세요

Zero-shot 학습: Zero-shot 학습은 모델이 학습 과정에서 **전혀 보지 못한 클래스나 테스크를 수행할 수 있는 능력**을 의미합니다. 이는 모델이 **기존에 학습한 지식을 바탕으로 새로운 상황이나 개념을 일반화하여 적용하는 것을 목표**로 합니다. 예를 들어, 텍스트 설명만을 통해 새로운 이미지를 인식하거나, 전혀 새로운 객체를 분류하는 작업이 이에 해당합니다. Zero-shot 학습은 대규모 멀티모달 데이터셋과 강력한 일반화 능력을 필요로 하며, 자연어 처리와 컴퓨터 비전의 통합적인 접근을 통해 구현됩니다.

Few-shot 학습: Few-shot 학습은 **매우 적은 수의 샘플, 보통 몇 개에서 수십 개의 예시만으로 새로운 클래스나 테스크를 학습하는 능력**을 의미합니다. 이는 메타러닝(meta-learning)이나 전이 학습(transfer learning) 기법을 통해 구현됩니다. **Few-shot 학습은 모델이 기존에 학습한 지식을 효과적으로 재활용하여 새로운 작업에 빠르게 적응할 수 있도록 도와줍니다.** 예를 들어, 단 몇 장의 이미지로 새로운 객체를 인식하거나, 소수의 예시로 새로운 언어 패턴을 학습하는 작업이 이에 해당합니다. Few-shot 학습은 데이터가 제한된 상황에서 모델의 유연성과 효율성을 높이는 데 중요한 역할을 합니다.

채점 기준:

zero shot : 한 번도 못 본 데이터에서도 테스크를 수행
few shot: 매우 적은 수의 샘플로 클래스, 테스크 학습 -> 과제 수행

위 내용이 포함되면 pass!

---

12. Multimodal이란 무엇인지 설명해주세요

Multimodal(멀티모달)은 텍스트, 이미지, 오디오 등 서로 다른 유형의 데이터를 동시에 처리하고 통합하는 방식을 의미합니다. 이는 인간이 다양한 감각을 통해 정보를 인식하고 처리하는 방식과 유사하게, 머신러닝 모델이 여러 종류의 데이터를 결합하여 보다 풍부하고 다층적인 정보를 활용할 수 있도록 합니다. 예를 들어, 텍스트와 이미지를 동시에 이해하여 이미지에 대한 설명을 생성하거나, 오디오와 비디오를 결합하여 더 정확한 감정 인식을 수행하는 작업이 이에 해당합니다.

멀티모달 접근 방식은 다양한 감각 정보를 결합함으로써 단일 모달리티에서 얻을 수 없는 깊이 있는 이해와 예측을 가능하게 합니다. 이는 의료 영상 분석, 자율 주행, 인간-컴퓨터 상호작용 등 다양한 분야에서 활용되며, 복잡한 현실 세계의 문제를 효과적으로 해결하는 데 기여합니다. 또한, 멀티모달 모델은 각 모달리티의 강점을 상호 보완적으로 활용하여, 더 높은 정확도와 신뢰성을 제공할 수 있습니다.

채점 기준:

서로 다른 유형의 데이터를 동시에 처리하고 통합하는 방식이라는 내용이 들어갔다면 pass!

---

13. CLIP은 어떤 원리로 동작하는지 설명해주세요

CLIP(Contrastive Language-Image Pretraining)은 **텍스트와 이미지 쌍을 대규모로 학습**하여, **두 모달리티 간의 연관성을 파악하는 모델**입니다. CLIP은 대규모 텍스트-이미지 데이터셋을 사용하여, 텍스트 설명과 해당 이미지 간의 관계를 학습합니다. 이 과정에서 **Contrastive Learning 기법을 적용하여, 동일한 쌍은 유사도를 높이고, 다른 쌍은 유사도를 낮추는 방향으로 모델을 최적화합니다.**

구체적으로, CLIP은 텍스트와 이미지를 각각 별도의 인코더(예: 텍스트는 Transformer, 이미지는 CNN 또는 ViT)를 통해 임베딩 공간으로 매핑합니다. 그런 다음, 텍스트 임베딩과 이미지 임베딩 간의 유사도를 계산하여, 올바른 쌍일수록 높은 유사도를 가지도록 학습합니다. 이를 통해 CLIP은 텍스트와 이미지 간의 의미적 연관성을 효과적으로 이해하게 되며, 다양한 멀티모달 작업에서 유연하게 활용될 수 있습니다. 예를 들어, 텍스트 설명을 기반으로 이미지를 검색하거나, 이미지에 대한 설명을 생성하는 등의 응용이 가능합니다.

채점 기준:

텍스트와 이미지 쌍을, 대규모로 학습하여, 두 모달리티 간 연관성을 파악한다.
Contrastive Learning 을 사용하여 동일한 쌍의 유사도는 높이고, 다른 쌍의 유사도는 낮춤을 통해 최적화를 진행한다

위 내용이 들어가면 pass!

---

14. VAE에 대해 설명해주세요

VAE(Variational Autoencoder)는 데이터의 잠재 공간(latent space)을 확률적으로 모델링하여 새로운 데이터를 생성하는 생성 모델입니다. VAE는 두 개의 주요 구성 요소인 Encoder와 Decoder로 구성됩니다. Encoder는 입력 데이터를 잠재 분포로 매핑하여 잠재 변수(latent variables)를 생성하고, Decoder는 이러한 잠재 변수를 다시 원래 데이터 공간으로 재생성합니다.

VAE의 학습 과정에서는 Reconstruction Loss와 Kullback-Leibler(KL) Divergence를 최적화합니다. Reconstruction Loss는 Decoder가 재생성한 데이터가 원래 입력 데이터와 얼마나 유사한지를 측정하며, 모델이 입력 데이터를 정확하게 재구성할 수 있도록 합니다. KL Divergence는 Encoder가 생성한 잠재 분포가 사전에 정의된 정규 분포와 얼마나 다른지를 측정하여, 잠재 공간이 구조화되고 일반화될 수 있도록 유도합니다.

이러한 구조 덕분에 VAE는 단순한 재구성뿐만 아니라, 잠재 공간에서의 샘플링을 통해 새로운 데이터를 생성할 수 있습니다. 예를 들어, 학습된 VAE 모델은 잠재 공간에서 샘플링된 변수를 사용하여 기존 데이터와 유사하지만 새로운 이미지를 생성할 수 있습니다. VAE는 이미지 생성, 데이터 보강, 노이즈 제거 등 다양한 응용 분야에서 활용되며, 잠재 공간을 통한 데이터의 구조적 이해와 생성 능력에서 강력한 성능을 발휘합니다.

채점 기준:

데이터의 잠재 공간을 확률적으로 모델링하여 새로운 데이터를 생성하는 생성 모델
Encoder와 Decoder
Encoder는 입력 데이터를 잠재 분포로 매핑하여 잠재 변수 생성
Decoder는 잠재 변수를 다시 원래 데이터 공간으로 재생성
Reconstruction Loss
KL Divergence

위 내용이 전부 들어가면 pass!