모델 비교 분석 보고서

모델 리스트

|  |  |  |  |
| --- | --- | --- | --- |
| **모델** | **Bottleneck 반복 수** | **Backbone 구조** | **Neck 구조** |
| v8n\_2442 | [2, 4, 4, 2] | C2f | C2f |
| v8n\_4664 | [4, 6, 6, 4] | C2f | C2f |
| v8n\_6886 | [6, 8, 8, 6] | C2f | C2f |
| v8n\_b8n5\_2442 | [2, 4, 4, 2] | C2f | C3 |
| v8n\_b8n5\_4664 | [4, 6, 6, 4] | C2f | C3 |
| v8n\_b8n5\_6886 | [6, 8, 8, 6] | C2f | C3 |
| v5n\_2442 | [2, 4, 4, 2] | C3 | C3 |
| v5n\_4664 | [4, 6, 6, 4] | C3 | C3 |
| v5n\_6886 | [6, 8, 8, 6] | C3 | C3 |
| v5n\_b5n8\_2442 | [2, 4, 4, 2] | C3 | C2f |
| v5n\_b5n8\_4664 | [4, 6, 6, 4] | C3 | C2f |
| v5n\_b5n8\_6886 | [6, 8, 8, 6] | C3 | C2f |

성능 표

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Npu/gpu | Npu map | Gpu map |
| V8n\_b5n8\_2442 | 0.554335895 | 0.303 | 0.5466 |
| V8n\_b5n8\_4664 | 0.53974665 | 0.294 | 0.5447 |
| v5n\_b8n5\_4664 | 0.579024035 | 0.318 | 0.5509 |
| v5n\_b8n5\_2442 | 0.55545471 | 0.306 | 0.5473 |
| v5n\_4664 | 0.604196816 | 0.334 | 0.5528 |
| V8n\_4664 | 0.599498926 | 0.335 | 0.5588 |
| V8n\_2442 | 0.53162732 | 0.295 | 0.5549 |
| v5n\_2442 | 0.544481358 | 0.295 | 0.5418 |

1. 실험 개요
   1. 배경 : 모델의 네트워크 구조를 수정하여 GPU 상에서 성능 결과가 좋아도 NPU 상에서의 성능이 꼭 좋다고 할 수 없다. 일반적으로 GPU는 부동소수점 연산에 최적화되어 있어 8bit 양자화 연산 시에도 성능 저하가 크지 않다. 하지만 NPU의 경우 연산 능력 및 구조가 GPU와 다르기 때문에 8bit 양자화를 적용하였을 때 특정 모델 구조에서는 성능 저하가 크게 나타날 수 있다. 따라서 YOLO 모델의 다양한 구조가 GPU 대비 NPU에서 얼마나 성능 차이를 유발하는지 분석하여 NPU에 적합한 구조를 찾기 위해 실험할 필요가 있다.
   2. 목적 : 모델의 다양한 구조들이 각각 GPU 대비 NPU에서 성능 차이에 어떤 영향을 미치는지 분석하고, 특정 구조 및 설정이 오차를 유발하는 요인을 찾는다. 특히 Bottleneck 반복 수와 Backbone & Neck에서의 여러 실험을 통해 Bottleneck과 C2f, C3가 성능 차이에 영향을 얼마나 미치는지 확인한다. 이를 통해 GPU에서는 성능이 우수하나 NPU에서는 하락하는 구조를 식별한다. 이로써 8bit 연산에 적합하거나 그렇지 않은 구조를 구분하고, NPU 성능의 최적하에 기여할 수 있는 구조적 특징을 도출한다.
2. 실험 기준
   1. 비교 기준 :
      1. Bottleneck 반복 수 : [2,4,4,2] vs. [4,6,6,4] vs [6,8,8,6]
         1. Bottleneck이 사용되는 레이어에서 [2,4,4,2], [4,6,6,4], [6,8,8,6] 세 가지 반복 수를 설정한다. Bottleneck은 모델의 연산 복잡도와 성능에 중요한 영향을 미치는 요소로, 반복 횟수에 따라 모델의 추론 시간 및 성능이 달라질 수 있다. 이를 통해 Bottleneck 반복 수가 GPU 대비 NPU 성능 저하에 미치는 영향을 확인하고자 한다.
      2. Backbone & Neck 구조 : (C2f, C2f) vs (C3, C3) vs (C2f, C3) vs (C3, C2f)
         1. Backbone과 Neck에서 C2f와 C3를 조합하여 (C2f, C2f), (C3, C3), (C2f, C3), (C3, C2f) 네 가지 조합을 설정한다. 각 조합이 NPU와 GPU에서의 성능에 미치는 영향을 파악하기 위해 설정한 것이다. 예를 들어, C2f는 상대적으로 경량화된 구조이며 C3는 복잡도가 C2f보다는 높은 구조이다. 따라서 이 조합을 통해 NPU에서 8비트 양자화 시 성능 저하가 크게 발생하는 조합을 찾고자 한다.
   2. 설명 : 각 기준이 성능에 미치는 영향을 분석하고, 이를 통해 NPU와 GPU에서 어떤 구조가 성능 차이를 일으키는지 파악한다. 실험의 목적은 GPU에서 우수한 성능을 보이지만 NPU에서는 성능 저하가 큰 구조를 찾아내어, 해당 구조가 8비트 연산에 적합하지 않다는 결론을 도출하는 것이다.
3. 실험 세팅
   1. 모델 구조 : 비교하는 모델의 세부적인 구조 설명 (예: v8n\_2442 vs v8n\_4664, v5n\_2442 vs v5n\_4664 등).
   2. 양자화 및 구현 방법 : 모델을 8비트로 양자화하여 실험하는 과정과 각 모델에서 사용된 기법을 기술합니다.
4. 실험 결과
   1. 모델 정보 : 전체 모델 결과
   2. 모델 성능 : 각 모델의 NPU와 GPU에서의 성능 결과를 정리하여 표와 그래프로 표시합니다. (예: FPS, mAP 등).
5. 결과 및 비교
   1. 성능 비교 : GPU와 NPU 간의 성능 차이가 큰 모델과 적은 모델을 구분하여 설명합니다.
   2. 오차 분석 : 특정 구조적 요소가 성능 차이에 미치는 영향과 그 원인을 분석합니다.
6. 분석 및 결론
   1. 결론 요약 : GPU 대비 NPU에서 성능 차이가 적은 모델 구조를 도출하고, 이를 통한 개선 방향을 제시합니다.
   2. 향후 방향 : 추가 실험이나 분석이 필요한 요소와 향후 연구 방향을 설명합니다.
7. 참고 문헌