**DynamicGate-MLP: 게이트 기반 동적 희소화 신경망 구조**

**초록**

본 논문에서는 다층 퍼셉트론(MLP)의 연산 효율성을 개선하기 위한 새로운 구조인 **DynamicGate-MLP**를 제안한다. 기존 MLP는 모든 입력–은닉–출력 연결이 항상 활성화되어, 불필요한 연산과 메모리 사용이 발생한다. 제안하는 DynamicGate-MLP는 각 연결 가중치에 대응하는 \*\*게이트 파라미터(gate logit)\*\*를 도입하여, 학습 과정에서 연결의 중요도를 학습적으로 결정한다. 순전파(forward)에서는 시그모이드 확률로부터 생성된 \*\*하드 게이트(hard gate)\*\*를 적용하여 특정 연결을 켜거나 끄고, 역전파(backward)에서는 \*\*Straight-Through Estimator(STE)\*\*를 사용해 그래디언트를 전달함으로써 게이트의 학습 가능성을 유지한다. MNIST 데이터셋을 대상으로 한 실험 결과, DynamicGate-MLP는 Baseline MLP와 유사한 정확도를 유지하면서도 활성 연결 수와 연산량(MACs)을 최대 70%까지 감소시킬 수 있었다. 본 연구는 MLP 계열 신경망의 효율적 추론 구조 설계와 모델 해석 가능성 강화에 기여한다.

**1. 서론**

딥러닝 모델은 이미지 인식, 음성 인식, 자연어 처리 등 다양한 분야에서 뛰어난 성능을 보였지만, 높은 계산량과 메모리 요구량은 여전히 문제로 지적된다. 특히 다층 퍼셉트론(MLP)은 구조가 단순하면서도 중요한 벤치마크 역할을 하지만, 모든 연결을 항상 활성화하는 dense 구조의 한계를 갖는다.

연산량과 파라미터 수를 줄이기 위한 다양한 기법들이 연구되어 왔다. 드롭아웃(Dropout)은 학습 과정에서 연결을 랜덤하게 비활성화하여 과적합을 방지하지만, 추론 시에는 모든 연결을 다시 사용하기 때문에 연산량 감소 효과가 없다. 프루닝(Pruning)은 학습이 끝난 뒤 불필요한 연결을 제거하는 방식으로, 추론 효율성은 개선되지만 학습 과정 중 희소성을 반영하지 못한다.

본 논문은 **학습과 추론 과정 모두에서 연결의 희소성을 반영할 수 있는 모델**, 즉 DynamicGate-MLP를 제안한다. 이 모델은 연결 단위로 학습 가능한 게이트를 부여하여, 학습 중 불필요한 연결은 점차 비활성화되고, 최종적으로는 효율적인 희소 구조로 변환된다.

**2. 관련 연구**

* **Dropout**: 랜덤 연결 차단으로 regularization 효과는 있지만 추론 효율 개선 없음.
* **Pruning**: 학습 후 불필요 연결 제거, 추론 효율은 있지만 학습 중 불필요 연결 반영 불가.
* **L0 regularization, Lottery Ticket Hypothesis**: 희소화 연구의 발전. 하지만 구현 복잡성, 안정성 문제 존재.
* **Dynamic neural network**: 조건부 연산 및 입력 기반 경량화 연구가 진행 중.

DynamicGate-MLP는 **연결 단위 게이트 학습**이라는 점에서 Dropout, Pruning, L0 regularization을 통합하는 특징을 가진다.

**3. 모델 제안: DynamicGate-MLP**

**3.1 기본 구조**

* 입력:
* 가중치:
* 게이트 파라미터:

**3.2 게이트 계산**

순전파 시에는 를 사용한다.

**3.3 Straight-Through Estimator (STE)**

이를 통해 순전파에서는 하드 마스크를, 역전파에서는 연속 확률로 그래디언트가 흘러가도록 한다.

**3.4 손실 함수**

Cross Entropy 손실에 게이트 확률 평균에 대한 L1 규제를 추가하여, 게이트가 불필요하게 켜지는 것을 억제한다.

**4. 실험**

**4.1 데이터셋**

* MNIST (손글씨 숫자 이미지, 28×28)

**4.2 실험 설정**

* Baseline MLP: 784→256→10
* DynamicGate-MLP: 동일 구조 + 게이트
* Optimizer: Adam (lr=1e-3)
* Epochs: 50

**4.3 결과**

* 정확도: Baseline 대비 거의 동일(98% 수준)
* 활성 연결 비율(r): 약 0.3~0.5 범위
* 추론 MACs: Baseline 대비 30~50% 수준
* 파라미터 수: 프루닝 후 약 60k~100k로 감소

**5. 논의**

* 학습 시에는 gate\_logit이 추가되므로 파라미터 수와 연산량이 Baseline 대비 증가.
* 그러나 추론 단계에서는 hard gate를 적용해 **작은 Dense MLP**로 변환 가능.
* Dropout과 달리 추론 효율을 가져가고, Pruning과 달리 학습 과정에서 희소화가 반영됨.
* β, τ 값에 따라 정확도와 연산 절감 간 trade-off 존재.

**6. 결론**

본 논문은 DynamicGate-MLP라는 새로운 게이트 기반 MLP 구조를 제안하였다. 이 모델은 학습 과정에서 불필요한 연결을 동적으로 비활성화하고, 추론 단계에서는 경량화된 네트워크로 변환되어 효율적인 연산을 가능하게 한다. MNIST 실험 결과, 정확도 손실 없이 최대 70% 연산 절감 효과를 확인하였다. 향후 연구로는 CIFAR-10/100, ImageNet 등 더 복잡한 데이터셋 적용, 게이트 스케줄링 기법 개발, 실제 추론 엔진에서의 속도 최적화가 필요하다.