PreNAS: 선호하는 원샷 학습 효율적인 신경망 아키텍처 검색

왕하이빈 [*1] 세 게 [*1] 헤센 첸 [1] 선 시우유 [1]

추상적인

사전 훈련된 모델의 광범위한 적용은
원샷 신경 구조 검색(NAS) 에서 일회성 훈련의 추세를 주도합니다 . 하지만,
거대한 샘플 공간 손상 내에서 훈련
개별 서브넷의 성능과 최적의 모델을 찾기 위해 많은 계산이 필요합니다. 본 논문에서는 원샷 훈련에서 대상 모델을 강조하는 검색 없는 NAS 접근 방식인 Pre NAS를 제시합니다. 특히 샘플 공간이 크게 줄어듭니다.

제로 코스트 셀렉터에 의해 사전에 웨이트 쉐어링 원샷 트레이닝이 수행됩니다.
업데이트 충돌을 완화하기 위해 선호하는 아키텍처 . 광범위한 실험을 통해 입증되었습니다.
PreNAS는 Vision Transformer와 컨볼루션 아키텍처 모두에서 최첨단 원샷 NAS 경쟁 제품보다 지속적으로 우수한 성능을 발휘합니다.
중요한 것은 즉각적인 전문화가 가능하다는 것입니다.
제로 검색 비용으로. 코드는 다음에서 사용할 수 있습니다.
https://github.com/tinyvision/PreNAS.

1. 소개

딥 러닝은 광범위한 분야에서 상당한 발전을 이루었습니다.
이미지 분류를 포함한 작업 수 (Tan & Le, 2019; Dosovitskiy et al., 2021), 음성인식 (Hsu et al., 2021) 및 자연어 처리 (Devlin et al., 2019).
그들의 성공에 기여하는 핵심 요소 중 하나는
결정적인 영향을 미칠 수 있는 모델 아키텍처 설계
최종 성능에 영향을 미칩니다. 그러나 수동 설계
모델 아키텍처는 종종 시간이 많이 걸리고
다양한 작업에 걸쳐 상당한 전문 지식을 제공합니다. 결과적으로 거기에
자동화된 설계 프로세스에 대한 필요성이 증가하고 있습니다.

신경망 아키텍처 검색(NAS)은 유망한 접근 방식입니다.
최적의 것을 찾는 것을 목표로 과제를 해결하기 위해

*동등 기부 1Alibaba Group, Beijing, China. 코레
지원: Xiuyu Sun <xiuyu.sxy@alibaba-inc.com>.

제 40차 국제기계학회 간행물
학습, 호놀룰루, 하와이, 미국. PMLR 202, 2023.

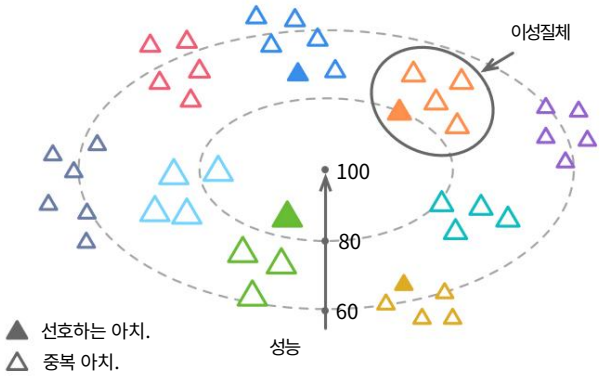


그림 1. PreNAS는 먼저 다음을 통해 선호하는 아키텍처를 식별합니다.
이성체로 표시되는 유사한 아키텍처 내의 비용이 없는 프록시,
그런 다음 선호하는 항목에만 집중 원샷 학습을 수행합니다.
더 나은 컨버전스를 달성하기 위한 아키텍처. 삼각형의 크기
아키텍처의 다른 리소스 소비를 의미합니다.

최소한의 인간 개입으로 주어진 작업에 대한 아키텍처 . One-shot NAS는 다음을 수행하는 일종의 NAS 제품군입니다.
알려진 잘 훈련된 가중치 공유 모델을 통해 검색
슈퍼넷으로. 가능한 서브넷을 교육하고 평가하여
하나의 슈퍼넷만 사용하는 검색 공간 내에서 원샷 NAS는 매번 훈련하고 평가하는 다른 NAS 방법에 비해 계산 비용을 크게 줄입니다.

처음부터 개별적으로 모델링합니다. One-shot NAS는 강력한 성능을 입증했으며 두 가지 모두에 적용되었습니다.
컨볼루션 신경망(CNN) 및 트랜스포머 기반 아키텍처 (Chitty-Venkata et al., 2022).

고성능 모델을 안정적으로 검색하기 위해,
원샷 NAS 솔루션은 대부분 중복 학습 전략을 사용합니다. 교육 단계 동안 과다한 서브넷
방대한 검색 공간(109 또는 심지어
훨씬 더 큰) 슈퍼넷을 반복적으로 업데이트합니다. 시
검색 단계에서 충분한 후보 서브넷이 채워집니다.
최적의 아키텍처에 순위를 매기십시오. 이 중복 전략은
실제로 좋은 성능을 달성하고 NAS의 불확실성을 줄였지만 최첨단 기술을 제한할 수도 있습니다.
원샷 NAS의 돌파구. 요약하면 다음과 같습니다.
우려해야 할 두 가지 특정 문제. (i) 검색 효율성.
지정된 리소스 제약 조건 하에서 각 검색에는 다음이 필요합니다.
수천 모델의 평가, 즉 시간

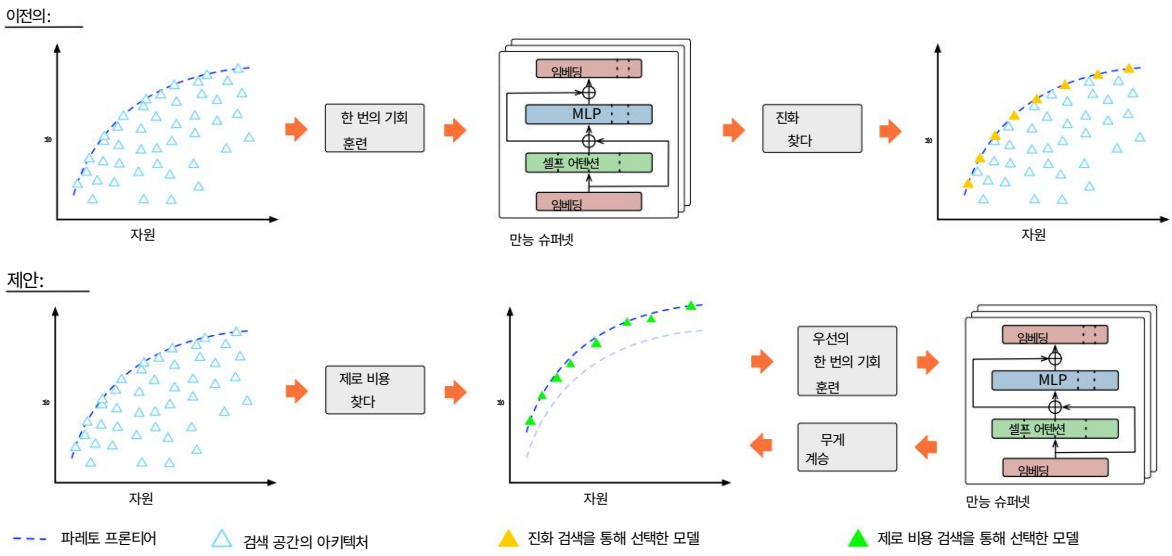


그림 2. PreNAS의 그림. 이전의 one-shot NAS는 one-shot 훈련 시 검색 공간의 모든 아키텍처를 샘플링합니다.
진화 검색에서 더 나은 평가를 위한 슈퍼넷. 대신 PreNAS는 먼저 무료 프록시를 통해 대상 아키텍처를 검색하고 다음으로
수퍼넷에 선호하는 원샷 트레이닝을 적용합니다. PreNAS는 선호하는 원샷 학습의 이점을 얻은 Pareto Frontier를 개선하고
제로 비용 검색에서 고급 선택된 아키텍처로 모델을 제공하여 교육 후 검색이 필요하지 않습니다.

소모적이고 자원 집약적입니다 (Cummings et al., 2022).
(ii) 훈련 효과. 방대한 교육 공간은 슈퍼넷 내의 학습 가능한 매개변수
를 적응하기 어렵게 만듭니다.
다양한 서브넷 (Liu et al., 2022). 결과 그라데이션
진동은 훈련 수렴을 손상시키고
궁극의 성능 (Gong et al., 2022).

이 작업은 새로운 선호 학습 전략을 제시합니다.
원샷 NAS, 즉 PreNAS의 잠재력을 충족합니다. 실용적인 응용을 고려하여 NAS
방법은 주어진 파레토 프론티어에서 고성능 모델을 선택합니다.

그림 1과 같이 리소스 제약이 있습니다. 따라서 우리는
최적의 리소스에 효율적으로 액세스할 수 있도록 다양한 리소스로 고품질
아키텍처를 미리 식별하도록 영감을 받았습니다.
거의 검색이 필요 없는 전문화를 달성합니다. 이에 따라 선택된
후보자는 좁은 샘플링을 형성합니다.
선호하는 공간이라고 합니다. 학습하는 동안 수퍼넷은 선호하는 샘플
공간에서 최적화됩니다.
기울기를 줄이는 훈련 효능 집중
충돌하고 따라서 잠재적으로 더 나은 수렴에 도달할 수 있습니다.
상태.

교육 없이 고품질 아키텍처를 식별하는 것은
사소하지 않은 작업. NAS 연구의 또 다른 라인, 제로샷
NAS는 이와 관련하여 실행 가능한 솔루션을 제공할 수 있습니다. 그만큼
제로샷 NAS의 핵심은 상관관계가 높은 빠른 프록시 역할을 하
는 제로 비용 평가 메트릭을 설계하는 데 있습니다.
실제 성능에. 이 작업에서는 one-shot NAS와 zero-shot
NAS 간의 연결을 창의적으로 설정하여
둘의 장점을 최대한 활용하십시오. 핵심 혁신
디자인 기능으로 스코어링 중심 프록시를 강화하는 것입니다.

이성질체 트랜스포머를 처리할 수 있습니다. 저평가되었지만
NAS의 중요한 현상은 여러 아키텍처가
동일한 리소스 소비를 소유할 수 있으며 이는 Transformer
아키텍처에서 특히 일반적입니다. 표준
Vision Transformers는 균일한
블록 전체에 치수를 삽입하여 블록을 재배열하여 새로운 블록을 형성
할 수 있다는 흥미로운 속성을 생성합니다.
일정한 매개변수를 유지하면서 아키텍처와
계산. 따라서 평가 프록시는 다음과 같이 수정됩니다.
이성질체 구조를 제거하는 능력을 소유합니다. 그림 1
설명된 프로세스에 대한 개략도를 제공합니다.
전반적으로 원샷과 제로샷 NAS의 조합
PreNAS는 최적의 빠른 검색을 달성할 뿐만 아니라
신경 구조뿐만 아니라 잘 훈련된 매개변수도 얻습니다.
효율적으로 즉시 사용 가능한 모델 전문화가 가능합니다.

주요 기여는 다음과 같이 요약됩니다.

- 원샷과 제로샷 NAS를 결합한 새로운 학습 패러다임인 PreNAS를 제안하여 훈련 및 검색
  선호하는 샘플 내에서 최적의 아키텍처를 위해
  공간, 검색 효율성 및 교육 향상을 목표로
  효능.
- 복합 아키텍처 선택 메커니즘은 특히 구조적 문제 해결에 중점을 두고 고품질 아키텍처를 구성하고 평가하도록 설계되었습니다.

  트랜스포머 아키텍처의 이성질체.
- 광범위한 실험과 분석이 수행됩니다.
  PreNAS의 효과를 확인하여
  다양한 작업에서 뛰어난 성능으로 CNN 및 ViT 아키텍처를 모두 검색
  할 수 있습니다 .

2. 배경 및 관련 작업

2.1. 원샷 NAS

원샷 NAS에서는 사전 정의된 검색 공간 내에서 서브넷을 공동으로 최적화하기 위해 가중치 공유 슈퍼넷이 구축되고 훈련됩니다. A 는 아키텍처 검색 공간이고 W는 모든 학습 가능한 가중치라고 합니다. 원샷 훈련의 목적은 최적의 가중치를 얻는 것입니다.

$$W^* = \underset{W}{\text{인수 분}} \quad E_{\alpha \in A}[L_{train}(\alpha \mid W_\alpha, D_{train})], \tag{1}$$

여기서 서브넷 α 는 검색 공간 A 에서 균일하게 샘플링되고 Wα 는 상속된 가중치의 해당 부분이며 Ltrain 은 훈련 데이터 세트에 정의된 손실 함수입니다. 잘 훈련된 슈퍼넷에서 주어진 리소스 제약 조건 c에서 최상의 모델 α는 서브넷 순위를 다음과 같이 검색하여 검색합니다.

$$\alpha^* = \underset{\alpha \in A}{\text{인수 최대값}} \; P_{val}(\alpha \mid W^*_\alpha, D_{val}) \text{ st } r(\alpha) < c, \tag{2}$$

여기서 Pval은 유효성 검사 데이터 세트의 성능을 평가 하고 r(·)은 리소스 소비를 보고합니다. 특수화 중에 검색 단계를 여러 번 수행하여 다양한 리소스 제약 조건에 대한 배포 모델을 얻을 수 있습니다.

원샷 NAS (Brock et al., 2018) 는 처음에 각 모델의 훈련 비용을 상각하기 위해 제안됩니다 .
정확도와 서브넷 품질의 상위 상관관계를 강조하므로 (Su et al., 2022) 최상의 아키텍처가 식별된 후 추가 재교육 단계가 필요합니다 (Wu et al., 2019; Li et al., 2020). . 최근 연구 에서는 컨볼루션 네트워크 (Yu & Huang, 2019; Cai et al., 2020; Yu et al., 2020), Transformer (Wang et al., 2020b; Chen et al., 2021b) 및 이들의 하이브리드 네트워크 (Gong et al., 2022). 그러나 검색 공간에서 모든 서브넷의 성능 순위를 유지하는 것과 단일 슈퍼넷 에서 대상 서브넷의 높은 성능을 달성하는 것을 모두 고려하기는 어렵습니다 . FocusFormer (Liu et al., 2022) 는 식의 균일 샘플러 대신 아키텍처 샘플러를 훈련합니다 . 1, 그러나 문제를 완전히 해결하지는 못합니다. 대신 Pre NAS는 사전에 제로샷 NAS를 통해 순위 작업을 원샷 교육에서 분리합니다 .

2.2. 제로샷 NAS

one-shot NAS의 장점은 아키텍처와 최적화된 매개 변수를 동시에 반환할 수 있다는 점입니다. 사전 훈련 된 모델 이 필요한 시나리오 에 더 적합 합니다 . _ 알., 2021). 그러나 아키텍처만 필요한 경우 원샷 NAS는 너무 번거롭습니다.

Zero-shot NAS는 비용이 많이 드는 모델 교육 없이 고품질 아키텍처를 찾고자 합니다 . 다양한 비용이 들지 않는 프록시는 아키텍처 품질을 효율적으로 추정하기 위해 제안되었습니다.
일반 목적은 다음과 같이 공식화될 수 있습니다.

$$\alpha^* = \underset{\alpha \in A}{\text{최대 인수}} \; P_{score}(\alpha) \text{ st } r(\alpha) < c, \tag{삼}$$

여기서 Pscore는 대규모 데이터 세트에 대한 교육 및 검증 없이 신속하게 아키텍처의 점수를 매길 수 있습니다 .

예를 들어 Mellor et al. (2021)은 훈련되지 않은 네트워크에서 데이터 포인트 간의 활성화 중첩에 의해 비용이 없는 프록시를 개발했습니다 . Lin et al. (2021)은 네트워크 표현력을 나타내는 Zen-Score를 제안합니다 . 한편, 이전 연구는 그래디언트 노름 (Abdelfattah et al., 2021), SNIP (Lee et al., 2019), GraSP (Wang et al., al., 2020a), SynFlow (Tanaka et al., 2020), NASI (Shu et al., 2022a). 또한 Shu et al. (2022b)는 서로 다른 그래디언트 기반 제로 비용 프록시 간의 연결을 이론적으로 증명하고 HNAS를 제안하여 기존의 교육이 필요 없는 NAS 알고리즘을 지속적으로 향상시킵니다. Xu et al. (2021a) RoBERTa 에 제안된 그래디언트 기반 프록시를 적용합니다 (Liu et al., 2019). TF-TAS (Zhou et al., 2022) 는 특히 ViT (Dosovitskiy et al., 2021) 에 대한 최초의 그래디언트 기반 제로 비용 프록시입니다 . Javaheripi et al. (2022)는 놀랍게도 autoregressive Transformers의 디코더 매개변수 수가 작업 성능 과 높은 상관관계가 있음을 발견했습니다.

일반성을 잃지 않고 대표적인 그래디언트 기반 프록시인 SNIP (Lee et al., 2019) 가 이 작업에서 채택되었습니다. SNIP의 스코어링 기능은 다음과 같이 정의됩니다.

$$P_{SNIP}(\alpha) = \sum_{\theta_i \in \theta_\alpha} \left| \frac{\partial L}{\partial \theta_i} \odot \theta_i \right|. \tag{4}$$

우리는 θ를 사용하여 훈련된 가중치 W와 구별하기 위해 무작위로 초기화된 매개변수를 나타냅니다. 손실 함수 L 을 정의하려면 순방향 추론을 수행하기 위해 최소 하나의 미니 배치 데이터가 필요합니다 . SNIP는 고품질 아키텍처를 선택할 수 있는 능력이 있는 한 비용이 들지 않는 다른 프록시로 자유롭게 대체할 수 있습니다 . 그림 2 에서 볼 수 있듯이 제로 비용 프록시로 샘플링된 모델은 일반적으로 Pareto Frontier에서 약간의 편차만 있으면 충분합니다.

3. 프리나스

이 섹션에서는 기존 원샷 NAS의 기본 잠재력을 밝히는 것으로 시작하여 그림 2와 같이 기능을 완전히 해제하기 위해 새로운 선호 학습 전략을 소개합니다 . 그런 다음 Trans 이전 아키텍처의 이성질체 문제에 대해 논의 하고 적응된 제로 비용 프록시를 자세히 설명합니다.
마지막으로 성능 분포의 균형을 맞추기 위해 교육 공정성을 해결할 것을 제안합니다 .

3.1. 선호 학습이 포함된 원샷 NAS

앞서 언급한 바와 같이 기존의 one-shot NAS는 검색 효율성과 학습 효과 측면에서 부족한 점이 있습니다.
훈련 단계에서 Eq. 1은 아키텍처 품질과 일치하는 성능에 도달하도록 모든 서브넷 α∈A 가 샘플링되고 최적화될 것으로 예상합니다. 그러나 검색 공간 A 는 거의 무한대로 크며 중복 교육 전략도 모든 서브넷 아키텍처를 통과할 수 없습니다. 더 중요한 것은 학습 가능한 가중치 W가 고품질 서브넷과 저품질 서브넷 에서 공유되기 때문에 충돌하는 적응으로 인해 최종 수렴 상태가 약해집니다. 게다가, 후보 날짜 아키텍처 그룹에서 최상의 아키텍처를 검색하기 위해 성능 평가기 Pval은 Eq. 2는 단일 검색에서 수천 번 실행될 수 있습니다.

이 검색 시간 오버헤드는 임의 검색 (Bender et al., 2018; Li & Talwalkar, 2019), 진화 알고리즘 (Real et al., 2019; Guo et al., 2020) 및 강화 학습과 같은 개선된 전략으로도 불가피합니다. (Pham et al., 2018; Tan et al., 2019).

검색 효율성을 높이는 관점에서 평가할 후보의 수를 최소화해야 합니다.

이 처리의 유익한 부작용 은 더 적은 수의 서브넷에서 집중 업데이트를 허용 하고 잠재적으로 각 모델이 더 나은 수렴 상태에 도달할 수 있다는 것입니다. 따라서 우리는 다음을 만족하는 훨씬 더 작은 선호 검색 공간 A 내에서 원샷 학습을 수행할 것을 제안합니다.

$$A \subset A \text{ 및 } |A| \ll |아| . \tag{5}$$

축소된 검색 공간은 충분한 품질을 갖추어야 하며 주어진 리소스 제약 조건에서 찾은 최상의 아키텍처가 전체 검색 공간 에서 이론적으로 최적인 아키텍처와 비교할 수 있습니다 .

$$\max_{\tilde{\alpha}\in\tilde{A}} \text{Pval}(\tilde{\alpha}) \approx \max_{\alpha\in A} \text{Pval}(\alpha) \tag{6}$$

대상

$$r(\tilde{\alpha}) < c \text{ 및 } r(\alpha) < c. \tag{7}$$

Pval 의 가중치 및 데이터 세트 기호는 단순화를 위해 생략되었습니다 .

보다 명확하게, 선호하는 검색 공간은 다양한 리소스 제약 조건 하에서 고품질 아키텍처를 선택하여 형성됩니다 .

$$A = S(A[c] , N) \mid \forall c \in C . \tag{8}$$

선택기 S는 유효한 하위 집합 에서 상위 N개의 최상의 아키텍처를 선택 하고 A[c]는 다음과 같이 정의됩니다.

$$A[c] = \{ \alpha \mid \alpha \in A \text{ st } r(\alpha) < c \} . \tag{9}$$

제약 조건 C는 리소스 소비에 대한 일련의 제한 사항입니다 (예: 모델 크기 또는 FLOP,

$$C = (a, a + \varepsilon, a + 2\varepsilon, \cdots, b), \tag{10}$$

여기서 ε 은 이산화 마진이고 a, b 는 검색 공간의 하한 및 상한입니다 .

선호하는 검색 공간에서 슈퍼넷은 가중치 얽힘 방식으로 훈련 되지만 (Chen et al., 2021b) 고품질 아키텍처에만 집중합니다. 선호하는 원샷 교육 프로세스는 다음과 같이 다시 작성할 수 있습니다.

$$W^* = \underset{W}{\text{인수 분}} \quad E_{\alpha \quad B\{A\}} [L_{train}(\alpha \mid W_\alpha, D_{train})]. \tag{11}$$

고품질 서브넷에 집중된 최적화는 열악한 아키텍처와의 가중치 공유를 피하고 개별 성능이 우수하다는 결과를 가져옵니다. 이는 섹션 4.2 에 보고된 최신 결과 에서 확인할 수 있습니다.

전문화 단계에서 주어진 리소스 제약 하에서 최적의 아키텍처를 직접 얻을 수 있습니다.

$$A^* = A. \tag{12}$$

이 절차는 많은 수의 후보자에 대한 비용이 많이 드는 검색 및 평가를 필요로 하지 않습니다 . 대부분의 경우 비용이 들지 않는 전문화를 달성하며 일반적인 용도로 충분합니다. 드문 세분화된 사용 사례의 경우 몇 개의 서브넷 순위를 지정하는 데 약간의 추가 오버헤드만 있습니다.

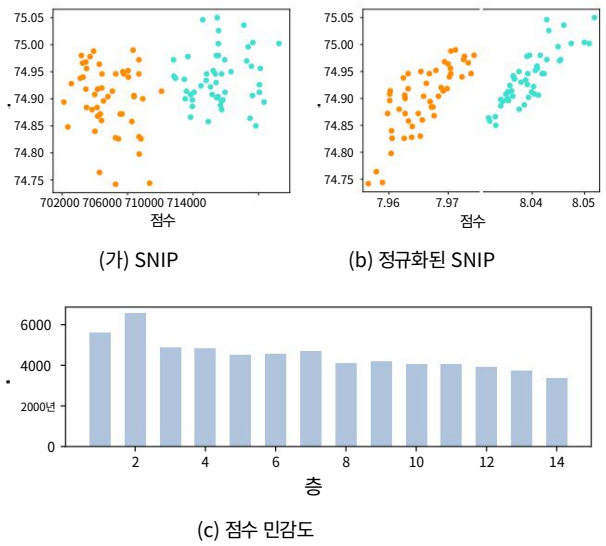
3.2. 제로 비용 변압기 선택기

Proxy Confusion PreNAS의 기초는 데이터 기반 교육 전에 고품질 아키텍처를 선택하는 데 있습니다 . 8. 이것은 사소한 작업이 아니며 우리는 제로샷 NAS 연구에서 프록시 기술에 의존합니다.

순진한 아이디어는 자격 있는 서브넷을 검색하기 위해 SNIP (Lee et al., 2019) 와 같은 그래디언트 기반 프록시를 사용하는 것입니다 .
예비 실험 후 우리는 SNIP와 다른 유사한 프록시가 다양한 용량의 아키텍처에서 잘 작동한다는 것을 발견했습니다 . 계산된 점수는 실제 정확도와 높은 상관 관계가 있습니다 . 그러나 동일한 리소스 소비로 아키텍처를 구분하는 데 혼란이 발생합니다 .

정의 3.1 (변압기 이성질체). i번째 블록 의 구성 이 βi인 L-블록 트랜스 포머 아키텍처 α := (β1, β2, ..., βL) 에 대해 , α를 재정렬함으로써 생성된 모든 아키텍처는 이성질체이다.

이성질체 이해 우리는 혼란 현상을 Transformer 아키텍처의 이성질체에 기인한다고 생각합니다.
SNIP는 원래 CNN 아키텍처용으로 제안되었으며 모델 크기와 높은 상관관계를 나타냅니다. 그러나 트랜스 전자 이성질체는 일정한 모델 크기와 FLOP를 가지고 있어 SNIP 점수를 속입니다. 그림 3(a)는 두 그룹의 이성질체를 나타내며 SNIP와 정확도 간에 유의미한 상관관계가 없음이 분명합니다 . 슈퍼넷의 블록당 점수를 조사하면 그림 3(c) 에서 하위 블록의 단위 기여도가 상대적으로 더 높다는 것을 알 수 있습니다.



(가) SNIP (b) 정규화된 SNIP

(c) 점수 민감도

그림 3. (a) SNIP 점수는 두 이성질체에서 다양한 서브넷의 테스트 정확도와 관련이 없습니다. (b) 계층 정규화를 사용한 SNIP 점수는 두 이성질체에서 다양한 서브넷의 테스트 정확도와 상당한 양의 상관관계를 가집니다. (c) MLP 비율이 3.5에서 4로 증가함에 따라 각 계층에서 SNIP 점수의 성장 .

따라서 SNIP는 일반적으로 평범한 성능을 나타내는 더 넓은 하위 블록이 있는 아키텍처를 선택하는 경향이 있습니다.
자세한 분석은 부록 A를 참조하십시오 . 더 낮은 블록의 약간의 이점은 임베딩 폭과 블록 수를 조정하여 쉽게 지배 할 수 있다는 점에 주목할 가치가 있습니다 . 이는 SNIP가 다른 리소스 소비에서 여전히 작동하는 이유를 설명합니다.

정규화된 이성질체 프록시 위의 분석을 바탕으로 계층별로 SNIP를 정규화하는 자연스러운 솔루션을 제안합니다 . 정규화된 SNIP의 수정된 스코어링 기능은 다음과 같이 정의됩니다.

$$\text{피}^{\text{표준}}_{\text{한조각}} = \sum_{l=1}^{\text{엘}} \left| \frac{\partial L}{\partial \theta_l} \odot \theta_\alpha^{\text{l}} \right| \Big/ \left| \frac{\partial L}{\partial \theta_l} \odot \theta_l \right|, \quad (13)$$

엘 여기서 θ 의 임의로 초기화된 매개변수는 의 상속된 매개변수임을 나타 l번째 블록 의 슈퍼넷 및 θ 서브넷 $_\alpha^{\text{l}}$ 냅니다.
α. 그림 3(b) 에서 볼 수 있듯이 정규화된 SNIP 점수는 이성체에 대한 정확도와 양의 상관관계를 보여줍니다. 각 이성질체 그룹이 많은 아키텍처를 포함할 수 있으며 모든 아키텍처에 P 놈을 적용하는 것은 비효율적이라는 점은 주목할 만합니다. 임의 검색 또는 진화 $_{\text{한조각}}$ 알고리즘과 같은 다른 근사 기술은 속도를 높이는 데 도움이 될 수 있지만 항상 최상의 결과를 얻는 것을 보장하지는 않습니다. 따라서 우리는 O(1) 복잡도 에서 최고 득점 이성질체를 탐욕스럽게 구성하는 P 놈 에 기반한 효율적인 할당 알고리즘을 제공합니다 . 알고리즘은 부록 B 에 자세히 설명되어 $_{\text{한조각}}$ 있습니다.

표 1. Vision Transformer용 PreNAS의 검색 공간. AutoFormer (Chen et al., 2021b)에 이어 서로 다른 매개변수 척도에 걸쳐 세 개의 슈퍼넷이 구축되었습니다 . 각 변수 요인의 삼중 항은 각각 하한, 상한 및 증분 단계를 의미합니다 . QKV 치수, 헤드 수 및 MLP 비율은 레이어마다 다릅니다.

	Supernet-Tiny	Supernet-소형	Supernet-베이스
임베드 치수	(192, 240, 24)	(320, 448, 64)	(528, 624, 48)
QKV Dim	(192, 256, 64)	(320, 448, 64)	(512, 640, 64)
MLP 비율	(3.5, 4, 0.5)	(3, 4, 0.5)	(3, 4, 0.5)
머리 번호	(3, 4, 1)	(5, 7, 1)	(8, 10, 1)
깊이 번호	(12, 14, 1)	(12, 14, 1)	(14, 16, 1)
매개변수 범위	5.4 – 10.5M	13 – 34M	42 – 76M

마지막으로 Eq.의 제로 비용 선택기 S 8 은 복합 2단계 작업입니다. 먼저 이성질체에 P 규범을 적용하여 대표자를 선출한 $_{\text{한조각}}$ 다음 PSNIP를 적용하여 주어진 리소스 제약 조건에서 선호하는 상위 N개의 아키텍처를 선택합니다 .

3.3. 성능 균형

학습 공정성에 대해 논의하고 학습 성능을 서브넷 전체에 보다 공정하게 분배하는 방법을 시연합니다 . 식에 표시된 바와 같이 . 1, 일반적인 원샷 훈련은 광대한 검색 공간, 즉 α U{A}에서 균일하게 서브넷을 샘플링합니다.
우리가 선호하는 검색 공간 A에서 후보 아키텍처는 리소스 소비 측면에서 고르게 분포되어 있습니다.
그러나 A 에 대한 균일한 샘플링은 서브넷 측면에서 공정한 업데이트 날짜를 가져오지 않습니다 . 임베딩 차원, 블록 수, 어텐션 헤드 수와 같은 변수 구성 요소는 동일하게 나타나지 않으며, 그 중 for mer 2가 더 큰 영향을 미칩니다. 따라서 우리는 A 의 아키텍처를 임베딩 차원 과 깊이로 그룹화하고 이 보정된 분포에 대해 균일한 샘플링을 수행할 것을 제안합니다. 즉, 식에서 α B{A} 11.

4. 실험

4.1. 설정

검색 공간 우리는 Transformer 블록의 5가지 변수 요소를 포함하여 Auto Former (Chen et al., 2021b) 와 동일한 검색 공간을 사용합니다 . 1. 우리는 PreNAS의 패러다임을 Tab에서 BigNAS (Yu et al., 2020) 와 동일한 검색 공간으로 CNN에 적용합니다. 2.

구현 세부 정보 우리는 timm (Wightman, 2019) 라이브러리 의 개선 사항을 사용하여 PyTorch (Paszke et al., 2019) 프레임워크 에 PreNAS를 구현했습니다. 슈퍼넷은 Auto Former (Chen et al., 2021b) 에 설명된 레시피에 따라 훈련됩니다.

표 2. MobileNetV2 기반 PreNAS의 검색 공간. BigNAS (Yu et al., 2020) 에 따라 빌딩 블록에는 다음이 포함됩니다.
바닐라 컨볼루션 및 역 병목 현상 잔차 블록(MB Conv) (Sandler et al., 2018).

스테이지	오퍼레이터	해상도	#채널	#레이어	커널
	전환수	192 – 320	32 – 40		삼
1	MBConv1	96 – 160	16 – 24	1	1 – 삼
2	MBConv6	96 – 160	24 – 32	2	2 – 삼
3	MBConv6	48 – 80	40 – 48	3	2 – 3, 5
4	MBConv6	24 – 40	80 – 88	3	2 – 3, 5
5	MBConv6	12 – 20	112 – 128	4	2 – 3, 5
6	MBConv6	12 – 20	192 – 216	6	2 – 3, 5
7	MBConv6	6 – 10	320 – 352	6	1 – 3, 5
	전환	6 – 10	1280 – 1408	2	1 1

용이하게 하기 위해 사용됩니다.
RandAugment를 포함한 수렴 (Cubuk et al., 2020),
mixup (Zhang et al., 2018), CutMix (Yun et al., 2019), Random Erasing (Zhong et al., 2020), 확률적 깊이 (Huang et al., 2016), 반복 확대 (Berman et al., 2019; Hoffer et al., 2020) 및 라벨 평활화 (Szegedy et al., 2016년; 위안 외, 2020). 자세한 하이퍼파라미터는
부록 C에 제시되어 있습니다. 입력 이미지는 모두 크기가 조정됩니다.
224 × 244 로 변경 하고 16 × 16 크기의 패치로 나눕니다. 우리는
미니 배치 크기가 1024인 AdamW 옵티마이저를 사용합니다.
학습률은 처음에 1e-3으로 설정되고 2e-5로 감소합니다.
500 에포크의 코사인 스케줄러를 통해. 이산화 마진 ε은 1M으로 설정됩니다. 실험을 진행했고
NVIDIA A100 GPU에서 측정된 설계 시간.

4.2. 주요실적

NAS ViT와 비교 PreNAS의 결과는
탭에 제시. 3. DeiT (Touvron et al.,
2021) 세 가지 사양에서 최상의 결과를 비교하고,
즉, PreNAS-Tiny, PreNAS-Small 및 PreNAS-Base, 각각
6M, 23M 및 54M의 매개변수 제한에 해당
각기. 최신 AutoFormer (Chen et al., 2021b)
중복되는 원샷 교육 전략을 채택하고 최적의

진화 알고리즘에 의한 서브넷. 관찰할 수 있습니다
당사의 PreNAS는 상위 1위와 상위 5위 모두에서 AutoFormer를 능가합니다.
세 가지 리소스 제약 조건 모두에서 정확도. 게다가, 그것은
이러한 성과를 높은 수준으로 달성하는 것이 중요합니다.
경쟁력있는 검색 효율성. 특히, AutoFormer는 진화 알고리즘을 활용하며 최소 90 GPU가 필요합니다.
최적의 모델을 안전하게 제공합니다. 반면 PreNAS는
실제로 원샷 교육 후 검색이 필요 없으며 즉시
주어진 자원 제약 하에서 원하는 모델을 제공합니다.

수제 ViT와 비교 일부 경쟁력 있음
인간이 정교하게 설계한 ViT 모델은
또한 DeiT (Touvron et al., 2021),
ConViT (d'Ascoli et al., 2021), TNT (Han et al., 2021) 및

표 3. Ima geNet의 다양한 Vision Transformer 비교 . ‡컨볼루션 및 트랜스포머 블록의 하이브리드 모델.

모델	TOP-1 (%)	TOP-5 (%)	#PARAMS (중)	플롭 (G)
LVT ‡	74.8	92.6	5.5	0.9
DEIT-TI	72.2	91.1	5.7	1.2
CONVIT-TI	73.1	91.7	6.0	1.0
TNT-TI	73.9	91.9	6.1	1.4
오토포머-TI	74.7	92.6	5.9	1.3
프레나스-TI	77.1	93.4	5.9	1.4
BOTNET-S1-59 ‡	81.7	95.8	33.5	7.3
DEIT-S	79.8	95.0	22.1	4.7
CONVIT-S	81.3	95.7	27.0	5.4
TNT-S	81.5	95.7	23.8	5.2
T2T-VIT-14	81.7	-	21.5	6.1
오토포머-S	81.7	95.7	22.9	4.9
프레나스-에스	81.8	95.9	22.9	5.1
BOTNET-S1-110 ‡	82.8	96.3	55	11
DEIT-B	81.8	95.6	86	18
CONVIT-B	82.4	95.9	86	17
오토포머-B	82.4	95.7	54	11
프레나스-B	82.6	96.0	54	11

T2T-ViT Yuan et al. (2021) 은 순수한 ViT 아키텍처인 반면
LVT (Yang et al., 2022) 및 BoTNet (Srinivas et al., 2021)
컨볼루션과 트랜스 포머 블록 으로 구성된 하이브리드 아키텍처입니다 . PreNAS는 정확성을 높이는 동시에
수동 설계보다 매개변수와 FLOP가 적습니다.
특히 더 큰 규모의 모델에서. 예를 들어, PreNAS Base는

DeiT-B의 매개변수는 63%, FLOP는 61%에 불과합니다. 하이브리드 아키텍처가 컨볼루션을 활용하여 기능을 효과적으로 다운샘플링한다는 점은 주목할 만합니다.
계산 감소에 유리합니다. 놀랍게도 Pre NAS는

유사하거나 더 적은 매개변수 및 FLOP.

요약하면, 실험 결과는 PreNAS의 효과를 확실하게 보여줍니다. 순수한 최첨단 비전 트랜스포머를 능가할 뿐만 아니라 컴팩트한 하이브리드 아키텍처로 100% 수준의 패리티에 도달합니다.

4.3. 분석 및 절제 연구

선호 아키텍처의 품질 PreNAS에서 선호 아키텍처 A 의 품질은 매우 중요하며
전체 프레임워크. 따라서 우리는 아키텍처를
제로 비용 선택기 S 에 의해 선택되고 (후속 원샷 훈련 없이) 처음부터 훈련하여 단독으로 검증합니다.
그 효과. 결과는 탭에 표시됩니다. 4 와 모두
방법은 동일한 300 epoch에 대해 공정하게 훈련됩니다.
레시피. Tiny 및 Small 검색 공간 모두에서 훈련된
PreNAS의 정확도는 다른 경쟁사보다 우수합니다.

표 4. 처음부터 훈련된 서브넷의 성능 비교
(300 에포크 및 1024 배치 크기). SNIP는 교체하여 적용됩니다.
선택기 S의 첫 번째 단계에서 정규화된 SNIP. 아래 첨자
ZS는 PreNAS의 제로샷 셀렉터만 사용한다는 의미입니다.

방법	TOP-1 (%)	TOP-5 (%)	#PARAMS (중)	시간 (시간)
SNIP-TI 74.4 92.4 AUTOFORMER-TI 74.4			6.0	0.1
92.4 TF-TAS-TI 75.2 92.7 PRENAS-TI ZS			5.9	415
74.8 92.6			6.2	12
			6.0	0.1
SNIP-S	81.2 95.7 81.5		23	0.3
오토포머-S	95.6 81.4 95.7		23	667
TF-TAS-S	81.5 95.8		24	17
PRENAS-S ZS			23	0.3

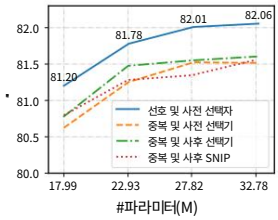


그림 4. ImageNet 정확도
제로 비용의 다양한 루틴
프록시 및 원샷 교육
작은 공간.

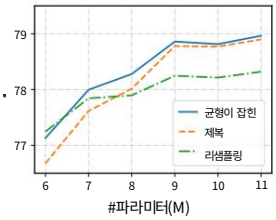
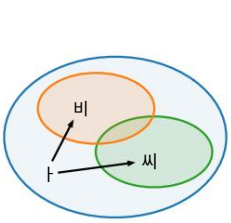


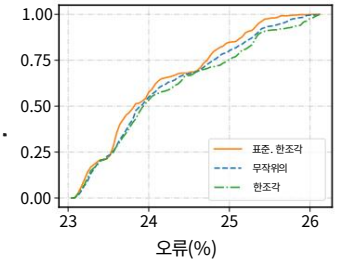
그림 5. ImageNet의 정확도
다양한 샘플링 전략
원샷 훈련 중
작은 공간.

실제로 더 유리한 TF-TAS-Ti를 제외하고
매개변수. 우리는 PreNAS의 이점이 선호도를 제거하는 이성질체 프록
시의 정상화에 있다고 생각합니다.
SNIP에서 더 넓은 하위 블록용. 예를 들어, MLP 비율
SNIP-Ti의 하위 계층에서 상위 계층까지 4.0,
각각 4.0, 4.0, 4.0, 4.0, 4.0, 4.0, 4.0, 4.0, 3.5, 3.5, 3.5,
4.0, 4.0, 4.0, 4.0, 4.0, 3.5, 3.5, 4.0, 4.0, 3.5,
PreNAS-Ti에서 4.0, 4.0. 효율성 측면에서 보면,
우리의 2단계 셀렉터는 탁월한 제로샷으로 유능합니다.
대리. 검색 시간은 다음과 같은 수준에서 거의 0입니다.
SNIP 및 아키텍처 품질은 비슷하거나 심지어
원샷 AutoFormer에서 진화 검색보다 낫습니다. TF TAS (Zhou et al., 2022) 는 최근 제안된 제로샷
트랜스포머 전용 프록시. 홍보했지만
AutoFormer에 비해 검색 효율이 48배 향상되었습니다.
를 통해 아키텍처를 제공하는 데 여전히 최소 12시간이 소요됩니다.
대규모 순방향 추론. 반대로 PreNAS는
미니 배치를 통한 단 하나의 전방 및 후방 전파
모든 서브넷은 초기화된 가중치를 직접 상속하고
슈퍼넷에서 그래디언트.

Preferred Training의 조사 PreNAS는 zero-shot selector와 선호하는 one-shot training의 조합 입니다 . 제안된 선택자를 다음과 결합할 수 있습니다.
기존의 중복 원샷 교육. 그림 4 는



(a) 검색 공간



(b) 오차 분포

그림 6. 다양한 검색 공간에서의 모델 품질. (a) 공간 A 는
초기 전체 공간, B는 다음을 적용하여 정제된 검색 공간입니다.
식에서와 같이 이성질체에 대한 계층 정규화 SNIP. 13, C 가 선택됨
정규화 없이 원본 SNIP로. (b) 오류율 통계
다른 공간에서 ImageNet 분류. 우리는 무작위로 샘플링
각 공간에 있는 500개의 모델이 아래에서 오류 분포를 검사합니다.
중복 원샷 교육. 누적 비율은 특정 값 미만의 오류가 있는 샘플링된 모델의 백분율을 나타냅니다.

몇 가지 가능한 대안. "선호 및 사전 선택자"는
PreNAS의 레시피. "중복 및 사전 선택기"란
중복 방식으로 슈퍼넷을 훈련하고 선택
임의로 초기화된 매개변수를 사용하여 사전에 서브넷,
"중복 및 사후 선택기"는 서브넷이
나중에 잘 훈련된 매개변수를 사용하여 슈퍼넷에서 선택합니다. 그림은 선호하는 효과를 확인합니다.
융합교육. 특히, 동일한 항목을 선택한 상태에서
"preferred & pre-selector" 및 "redundant & pre-selector"의
서브넷, 선호 훈련은 정확도를 약
중복 서브넷과 비교하여 각 서브넷에 대해 0.5%. 에 의해
고품질 아키텍처에 집중, 선호
교육에서 업데이트 충돌이 덜 발생하므로
선택한 모든 서브넷에 대한 보다 적절한 최적화.

정규화된 프록시의 효과 twp 단계 선택기 S에서 정규화된 이성질체 프록시의 중요성을 추가 로 분석합니다 . 첫 번째 단계에서 고유한 대표를 보존합니다.
표준화된 방법을 사용하여 Transformer 이성질체의 각 그룹에서
SNIP 점수 P 놈 스닙, 세련된 검색 공간 B를 다음과 같이 형성합니다.
그림 6(a)에서. 비교를 위해 원본 SNIP를 적용합니다.
검색을 얻기 위한 첫 번째 단계에서 계층 정규화 없이
space C. 우리는 RegNet (Radosavovic) 의 기준을 따릅니다.
et al., 2020) 정제된 검색의 품질을 특징짓기 위해
공백. 우리는 무작위로 500개의 아키텍처를 샘플링합니다.
성능 분포를 분석하기 위해 전체 검색 공간 A.
B 와 C 의 해당 이성질체가 선택되어 유지됩니다.
공정한 비교를 위한 상수 매개변수. 에 표시된 바와 같이
그림 6(b)에서 곡선은 B 의 품질이 A를 커버함을 의미합니다.
C와 C의 모델은 최악의 성능을 보입니다.

성능 균형의 효과 원샷 훈련 중에 옵티마이저는 학습 가능한 가중치를 업데이트합니다.
각 반복의 검색 공간에서 선택된 서브넷. ㅏ

표 5. ImageNet 사전 교육을 통해 다양한 다운스트림 작업에 대한 전이 학습.

모델	#PARAMS	CIFAR-10	CIFAR-100	꽃	자동차	애완동물	INAT-19
EFICIENTNET-B5	3천만	98.7	91.1	98.5	-	-	-
그라핏 RESNET-50 25M		-	-	98.2	92.5	-	75.9
VIT-B/16	86M	98.1	87.1	89.5	-	93.8	-
VIT-L/16	307M	97.9	86.4	89.7	-	-	-
DEIT-B	86M	99.1	90.8	98.4	92.1	-	77.7
VITAE-S	24M	98.8	90.8	97.8	91.4	94.2	76.0
DEARKD-S	22M	98.4	89.3	97.4	91.3	-	-
프레나스-에스	23M	99.1	91.2	97.6	92.2	94.9	76.4

표 6. ImageNet에서 CNN 아키텍처의 비교 결과.
PreNAS의 패러다임은 BigNAS 검색 공간에 적용됩니다.

그룹 모델		TOP-1 (%)	플롭 (중)
200M	비그나스-S 76.5 프레나스-S CNN		242
	77.4		237
400M	BIGNAS-M 78.9 PRENAS-M CNN		418
	79.9		413
600M	BIGNAS-L 79.5 PRENAS-L CNN		586
	80.0		528
1000M	BIGNAS-XL 80.9 PRENAS-XL CNN		1040
	81.4		986

기본 샘플링 방법은 후보자를 균일하게 선택하는 것입니다.
동일한 확률로 선호 공간 A 의 아키텍처 . 그러나 이것은 최적
이 아닌 최적화로 이어질 수 있습니다.
가변 빌딩 요소가 훈련됨에 따라 특정 가중치의
다른 주파수로. 그림 5에서 볼 수 있듯이 균일한 샘플링은 작
은 모델의 성능을 저하시킵니다.
변수 요인의 흔하지 않은 선택. 대조적으로 우리는
또한 사용된 리샘플링 방법을 실험했습니다.
중복 원샷 교육, 즉 무작위로 새로운 구성
분해된 검색 공간에서 가변 요소를 최소 구성 요소로 사용하여 아키텍처.
이것은 작은 모델을 개선하는 데 도움이 되지만 대부분의 서브넷 성능은

최적화를 위한 서브넷의 증가로 인해 성능이 저하됩니다. 균형 잡힌 샘플링
은 가장
중요한 임베딩 차원과 깊이가 상당히 업데이트됩니다.
따라서 지속적으로 우수한 성능을 달성합니다.

4.4. 전이 학습 결과

딥러닝에서는 일반화의 힘이 중요하기 때문에
모델, 우리는 다양한 다운스트림에서 PreNAS를 추가로 평가합니다.
전이 학습 능력을 측정하는 작업. 수행원
컨벤션, 비교 모델은 Ima에서 사전 훈련됩니다.
geNet을 선택한 다음 대상 데이터 세트에서 미세 조정합니다. 보여진 바와 같이
탭에서 5, CIFAR-10/100 (Krizhevsky
& Hinton, 2009), Flowers-102 (Nilsback & Zisserman,
2008), Stanford Cars (Krause 외, 2013), Oxford-IIIT
Pets (Parkhi et al., 2012) 및 iNaturalist 2019 (Horn et al.,
2018). 우리의 작은 모델은 ViT 및 DeiT를 능가할 수 있습니다.
매개변수가 몇 배 더 적은 여러 데이터 세트에서
PreNAS-S는 비슷한 규모의 ViTAE-S (Xu
et al., 2021b) 및 DearKD-S (Chen et al., 2022) 가 있으며
convnet 모델과 동등합니다.

4.5. CNN 결과

PreNAS의 주요 초점은 Vision Transformer이지만 one-shot NAS
에 대한 선호 학습의 개념은
일반적이며 다양한 아키텍처에 적용될 수 있습니다. 여기
우리는 CNN 아키텍처에 대한 적응을 시연합니다.
그 다양성과 확장성을 증명합니다. Big NAS (Yu et al., 2020) 를 실험적
기준으로 선택합니다 .
CNN 아키텍처용으로 널리 사용되는 단일 단계 NAS 프레임워크입니다.
PreNAS의 패러다임은 BigNAS로 쉽게 마이그레이션할 수 있으며,
주요 노력은 트랜스 포머 블록의 검색 공간을 컨볼루션 레이어로 대체하
는 것입니다. 실험된
검색 공간은 탭에 자세히 설명되어 있습니다. 2. 다운 샘플링 특성으로
인해 CNN 아키텍처에서 이성질체 가 거의 나타나지 않기 때문에 ,
식의 정규화 단계. 13은 건너뛸 수 있습니다. 벤치 마크 결과는 탭에 표시됩니
다. 6. 당사의 PreNAS는
다양한 모델 크기에 대해 0.5 % ~ 1.0% 향상된 정확도
유사한 FLOP 측면에서 BigNAS보다.

5. 결론

본 논문에서는 훈련 공간을 줄이는 원샷 신경 구조 탐색 방법인 PreNAS
를 제안하였다.
유망한 이성질체를 식별하고 추가 가지치기를 통해
희소화를 통한 훈련 공간. 이를 통해 PreNAS
보다 집중된 세트에서 원샷 학습을 수행할 수 있습니다.
제로 비용 사전 검색을 통해 학습 후 무료로 검색할 수 있습니다. 우리의

실험은 PreNAS가 생성할 수 있음을 보여줍니다.

다른 NAS 방법에 비해 검색 시간이 크게 단축된 매우 정확한 모델입니다. 우리는 PreNAS가 NAS의 효율성과 효과를 개선하는 유망한 접근 방식을 대표한다고 믿으며 향후 연구에서 PreNAS의 잠재력을 더 탐구할 수 있기를 기대합니다 .

참조

Abdelfattah, MS, Mehrotra, A., Dudziak, L. 및 Lane, ND 경량 NAS를 위한 비용이 들지 않는 프록시. 2021년 ICLR(International Conference on Learning Representations) 에서 .

Bender, G., Kindermans, P., Zoph, B., Vasudevan, V. 및 Le, QV 원샷 아키텍처 검색 이해 및 단순화. ICML(International Conference on Machine Learning), 80권, pp. 549–558, 2018.

Berman, M., Jegóu, H., Vedaldi, A., Kokkinos, I. 및 Douze, M. MultiGrain: 클래스 및 인스턴스에 대한 통합 이미지 임베딩 . arXiv 프리프린트 arXiv:1902.05509, 2019.

Brock, A., Lim, T., Ritchie, JM, and Weston, N. SMASH: 하이퍼넷 작업을 통한 원샷 모델 아키텍처 검색 . 2018년 ICLR(International Conference on Learning Representations)에서 .

Cai, H., Gan, C., Wang, T., Zhang, Z., and Han, S. Once for-All: 하나의 네트워크를 교육하고 효율적인 배포를 위해 전문화합니다 . 2020년 ICLR(International Conference on Learning Representations)에서 .

Chen, H., Lin, M., Sun, X. 및 Li, H. NAS-Bench-Zero: 제로샷 신경 구조 검색을 이해하기 위한 대규모 데이터 세트. https://openreview.net/forum?id=hP-SILoczR, 2021a.

Chen, M., Peng, H., Fu, J. 및 Ling, H. AutoFormer: 시각적 인식을 위한 변환기 검색. In IEEE/CVF International Conference on Computer Vision (ICCV), pp. 12250–12260, 2021b.

Chen, X., Cao, Q., Zhong, Y., Zhang, J., Gao, S. 및 Tao, D. DearKD: 비전 변환기를 위한 데이터 효율적인 초기 지식 증류 . 컴퓨터 비전 및 패턴 인식(CVPR)에 관한 회의 , pp. 12042–12052, 2022.

Chitty-Venkata, KT, Emani, M., Vishwanath, V. 및 So mani, AK 변압기에 대한 신경 아키텍처 검색: 설문 조사. IEEE 액세스, 10:108374–108412, 2022.

Cubuk, ED, Zoph, B., Shlens, J. 및 Le, QV Ran daugment: 검색 공간을 줄인 실용적인 자동 데이터 확대 . 컴퓨터 비전 및 패턴 인식 워크숍(CVPR)에 관한 회의, pp. 3008–3017, 2020.

Cummings, D., Sarah, A., Sridhar, SN, Szankin, M., Munoz, JP 및 Sundaresan, S. 양식 전반에 걸쳐 신경 아키텍처 검색을 가속화하기 위한 하드웨어 인식 프레임워크. arXiv 프리프린트 arXiv:2205.10358, 2022.

d'Ascoli, S., Touvron, H., Leavitt, ML, Morcos, AS, Biroli, G. 및 Sagun, L. ConViT: 소프트 컨벌루션 유도 바이어스로 비전 변환기 개선 . ICML (International Conference on Machine Learning), 139권, pp. 2286–2296, 2021.

Devlin, J., Chang, M., Lee, K. 및 Toutanova, K. BERT: 언어 이해를 위한 깊은 양방향 변환기 사전 훈련 . 전산 언어학 협회 북미 지부 컨퍼런스 : 인간 언어 기술(NAACL-HLT), 4171–4186, 2019페이지.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. 및 Houlsby, N. 이미지는 16x16 단어의 가치가 있습니다: Transformers for image recognition at scale. 2021년 ICLR(International Conference on Learning Representations)에서 .

Gong, C., Wang, D., Li, M., Chen, X., Yan, Z., Tian, Y., Chandra, V. 등. NASViT: 그래디언트 충돌 인식 슈퍼넷 교육을 통해 효율적인 비전 변환기를 위한 신경 아키텍처 검색 . 2022년 ICLR(International Conference on Learning Representations)에서 .

Guo, Z., Zhang, X., Mu, H., Heng, W., Liu, Z., Wei, Y. 및 Sun, J. 균일한 샘플링을 사용한 단일 경로 원샷 신경 구조 검색 . 유럽 컴퓨터 비전 회의(ECCV), 12361권, 544–560페이지 , 2020년.

Han, K., Xiao, A., Wu, E., Guo, J., Xu, C. 및 Wang, Y. 변압기의 변압기. 신경 정보 처리 시스템(NeurIPS)의 발전, pp. 15908–15919 , 2021.

Hoffer, E., Ben-Nun, T., Hubara, I., Giladi, N., Hoefler, T. 및 Soudry, D. 배치 보강: 인스턴스 반복을 통해 일반화 개선. 컴퓨터 비전 및 패턴 인식(CVPR)에 관한 회의 , pp. 8126–8135, 2020.

Horn, GV, Aodha, OM, Song, Y., Cui, Y., Sun, C., Shepard , A., Adam, H., Perona, P. 및 Belongie, SJ 비자연주의 종 분류 및 탐지 데이터 세트 . 컴퓨터 비전 및 패턴 인식 (CVPR) 에 관한 회의, pp. 8769–8778, 2018.

Hsu, W., Bolte, B., Tsai, YH, Lakhotia, K., Salakhutdinov, R. 및 Mohamed, A. HuBERT: 자기 감독 연설

은닉 유닛의 마스킹된 예측에 의한 표현 학습 . IEEE ACM 트랜스. 오디오 스피치 랭. 프로세스., 29:3451–3460, 2021.

Huang, G., Sun, Y., Liu, Z., Sedra, D. 및 Weinberger, KQ 확률적 깊이가 있는 Deep 네트워크. 컴퓨터 비전에 관한 유럽 회의(ECCV), 9908권, 646–661페이지, 2016년.

Javaheripi, M., de Rosa, GH, Mukherjee, S., Shah, S., Religa, TL, Mendes, CCT, Bubeck, S., Koushanfar, F. 및 Dey, D. LiteTransformerSearch: 훈련이 필요 없는 신경 구조 효율적인 언어 모델을 검색합니다.
신경 정보 처리 시스템(NeurIPS)의 발전 , 2022.

Krause, J., Stark, M., Deng, J. 및 Fei-Fei, L. 세분화된 분류를 위한 3d 객체 표현. 컴퓨터 비전 워크숍 에 관한 IEEE 국제 회의 (ICCV 3dRR-13), pp. 554–561, 2013.

Krizhevsky, A. 및 Hinton, G. 작은 이미지에서 여러 계층의 기능 학습 . 기술 보고서, 토론토 대학, 2009.

Lee, N., Ajanthan, T. 및 Torr, PHS SNIP: 연결 민감도에 기반한 단일 샷 네트워크 프루닝. 2019년 ICLR(International Conference on Learning Representations) 에서 .

Li, C., Peng, J., Yuan, L., Wang, G., Liang, X., Lin, L. 및 Chang, X. 지식 증류를 통한 블록 단위 감독 신경 구조 검색. 컴퓨터 비전 및 패턴 인식(CVPR)에 관한 회의, pp. 1989–1998, 2020.

Li, L. 및 Talwalkar, A. 신경 구조 검색을 위한 무작위 검색 및 재현성. UAI(Artificial Intel ligence)의 불확실성에 관한 35차 회의 진행 , 115권, 367–377페이지, 2019년.

Lin, M., Wang, P., Sun, Z., Chen, H., Sun, X., Qian, Q., Li, H. 및 Jin, R. Zen-NAS: A zero-shot nas for 고성능 이미지 인식. 컴퓨터 비전에 관한 IEEE/CVF 국제 회의(ICCV), pp. 347–356 , 2021.

Liu, J., Cai, J., and Zhuang, B. FocusFormer: 아키텍처 샘플러를 통해 필요한 것에 집중합니다. arXiv 프리프린트 arXiv:2208.10861, 2022.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. 및 Stoyanov, V. RoBERTa: 견고하게 최적화된 bert pretraining ap proach입니다. arXiv 프리프린트 arXiv:1907.11692, 2019.

Mellor, J., Turner, J., Storkey, AJ 및 Crowley, EJ 교육 없이 신경 아키텍처 검색. 국제에서
기계 학습에 관한 회의(ICML), 139권, pp. 7588–7598, 2021.

Nilsback, M. 및 Zisserman, A. 많은 클래스에 대한 자동화된 꽃 분류. In Indian Conference on Computer Vision, Graphics & Image Processing (ICVGIP), pp. 722–729, 2008.

Parkhi, OM, Vedaldi, A., Zisserman, A. 및 Jawahar, CV Cats and dogs. 컴퓨터 비전 및 패턴 인식(CVPR)에 관한 회의, pp. 3498–3505, 2012.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., 및 Chintala, S. PyTorch: 명령형 스타일의 고성능 딥 러닝 라이브러리입니다. 신경 정보 처리 시스템의 발전 32, pp. 8024–8035, 2019.

Pham, H., Guan, M., Zoph, B., Le, Q. 및 Dean, J. 매개변수 공유를 통한 효율적인 신경 구조 검색. ICML (International Conference on Machine Learning), pp. 4095–4104, 2018.

Radosavovic, I., Kosaraju, RP, Girshick, R., He, K. 및 Dollar, P. 네트워크´설계 공간 설계. 컴퓨터 비전 및 패턴 인식 (CVPR) 에 관한 회의 , pp. 10428–10436, 2020.

Real, E., Aggarwal, A., Huang, Y. 및 Le, QV 이미지 분류기 아키텍처 검색을 위한 정규화 진화. 인공 지능에 관한 제 33차 AAAI 회의 (AAAI), pp. 4780–4789, 2019.

Sandler, M., Howard, AG, Zhu, M., Zhmoginov, A. 및 Chen, L. 반전 잔차 및 선형 병목 현상: 분류, 감지 및 세분화를 위한 모바일 네트워크 . CoRR, abs/1801.04381, 2018.

Shu, Y., Cai, S., Dai, Z., Ooi, BC 및 Low, BKH
NASI: 초기화 시 레이블 및 데이터에 구애받지 않는 신경 구조 검색 . ICLR(International Conference on Learning Representations) 에서 2022a.

Shu, Y., Dai, Z., Wu, Z. 및 Low, BKH 그래디언트 기반 훈련이 필요 없는 신경 구조 검색 통합 및 강화. arXiv 프리프린트 arXiv:2201.09785, 2022b.

Srinivas, A., Lin, T., Parmar, N., Shlens, J., Abbeel, P. 및 Vaswani, A. 시각적 인식을 위한 병목 현상 변환기. 컴퓨터 비전 및 패턴 인식(CVPR)에 관한 회의 , pp. 16519–16529, 2021.

Su, X., You, S., Xie, J., Zheng, M., Wang, F., Qian, C., Zhang, C., Wang, X. 및 Xu, C. ViTAS: 비전 트랜스포머 건축 검색. 컴퓨터 비전에 관한 유럽 회의 (ECCV), pp. 139–157, 2022.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. 및 Wojna, Z. 컴퓨터 비전을 위한 초기 아키텍처 재고. 컴퓨터 비전 및 패턴 인식(CVPR) 컨퍼런스, 2818–2826, 2016.

Tan, M. 및 Le, QV EfficientNet: 컨볼루션 신경망을 위한 모델 스케일링 재고. ICML(International Conference on Machine Learning), 97권, pp. 6105–6114, 2019.

Tan, M., Chen, B., Pang, R., Vasudevan, V., Sandler, M., Howard, A. 및 Le, QV MnasNet: 모바일용 플랫폼 인식 신경 아키텍처 검색 . 컴퓨터 비전 및 패턴 인식(CVPR) 컨퍼런스 , 2820–2828, 2019.

Tanaka, H., Kunin, D., Yamins, DLK 및 Ganguli, S. 시냅스 흐름을 반복적으로 보존하여 데이터 없이 신경망 가지치기 . 신경 정보 처리 시스템(NeurIPS) 의 발전 , 2020.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A. 및 Jegou, H. 교육 데이터 효율적인 이미지 변환기 및 주의를 통한 증류. ICML(International Conference on Machine Learning), 139권, pp. 10347–10357, 2021.

Wang, C., Zhang, G. 및 Grosse, RB는 기울기 흐름을 보존하여 훈련 전에 당첨 티켓을 뽑습니다. ICLR(International Conference on Learning Representations)에서 2020a.

Wang, H., Wu, Z., Liu, Z., Cai, H., Zhu, L., Gan, C. 및 Han, S. HAT: 효율적인 자연어 처리를 위한 하드웨어 인식 변환기. 전산 언어학 협회(ACL) 연례 회의에서, pp. 7675–7688 , 2020b.

Wang, R., Bai, Q., Ao, J., Zhou, L., Xiong, Z., Wei, Z., Zhang, Y., Ko, T. 및 Li, H. LightHuBERT: 가볍고 구성 가능 단일 숨겨진 단위 BERT를 사용한 음성 표현 학습. International Speech Communication Association (INTERSPEECH) 연례 회의에서 , pp. 1686–1690, 2022.

Wightman, R. Pytorch 이미지 모델. https://깃허브. com/rwightman/pytorch-image-models, 2019.

Wu, B., Dai, X., Zhang, P., Wang, Y., Sun, F., Wu, Y., Tian, Y., Vajda, P., Jia, Y. 및 Keutzer, K. FBNet: 차별화 가능한 신경 구조 검색을 통한 하드웨어 인식 효율적인 convnet 설계. 컴퓨터 비전 및 패턴 인식(CVPR)에 관한 회의, pp. 10734–10742, 2019.

Xu, J., Zhao, L., Lin, J., Gao, R., Sun, X. 및 Yang, H. KNAS: 녹색 신경 구조 검색. ICML(International Conference on Machine Learning), 139권, pp. 11613–11625, 2021a.

Xu, Y., Zhang, Q., Zhang, J. 및 Tao, D. ViTAE: 고유 유도 바이어스를 탐색하여 고급 비전 변압기 . 신경 정보 처리 시스템(NeurIPS) 의 발전 , pp. 28522–28535, 2021b.

Yang, C., Wang, Y., Zhang, J., Zhang, H., Wei, Z., Lin, Z. 및 Yuille, 향상된 셀프 어텐션을 갖춘 AL Lite 비전 트랜스포머. 컴퓨터 비전 및 패턴 인식(CVPR)에 관한 회의, pp. 11988–11998, 2022.

Yin, Y., Chen, C., Shang, L., Jiang, X., Chen, X., and Liu, Q. AutoTinyBERT: 사전 훈련된 효율적인 언어 모델을 위한 자동 하이퍼 매개변수 최적화.

전산 언어학 협회 (ACL) 연례 회의에서, pp. 5146–5157, 2021.

Yu, J. 및 Huang, TS 보편적으로 가늘어지는 네트워크 및 개선된 교육 기술. 컴퓨터 비전에 관한 IEEE/CVF 국제 회의(ICCV), pp. 1803–1811, 2019.

Yu, J., Jin, P., Liu, H., Bender, G., Kindermans, P., Tan, M., Huang, TS, Song, X., Pang, R. 및 Le, Q. BigNAS : 대규모 단일 단계 모델로 신경 구조 검색을 확장합니다. 컴퓨터 비전에 관한 유럽 회의 (ECCV), 12352권, 702–717페이지, 2020년.

Yuan, L., Tay, FEH, Li, G., Wang, T. 및 Feng, J. 레이블 스무딩 정규화를 통해 지식 증류를 다시 방문합니다. 컴퓨터 비전 및 패턴 인식(CVPR) 컨퍼런스, 3902–3910, 2020.

Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z., Tay, FEH, Feng, J. 및 Yan, S. Tokens-to-token ViT : ImageNet에서 처음부터 비전 변환기를 교육합니다. IEEE /CVF 국제 컴퓨터 비전 회의 (ICCV), pp. 538–547, 2021.

윤성, 한동, 천성, 오성주, 유영주, 최재진 CutMix: 지역화 가능한 기능으로 강력한 분류자를 교육하는 정규화 전략입니다 . 컴퓨터 비전(ICCV) 에 관한 IEEE/CVF 국제 컨퍼런스, pp. 6022–6031, 2019.

Zhang, H., Cisse, M., Dauphin, YN 및 Lopez-Paz, D. mixup: 경험적 위험 최소화를 넘어. 2018년 ICLR(International Conference on Learning Representations) 에서 .

Zhong, Z., Zheng, L., Kang, G., Li, S., and Yang, Y. Random 소거 데이터 증대. AAAI 인공 지능 회의(AAAI), pp. 13001–13008, 2020.

Zhou, Q., Sheng, K., Zheng, X., Li, K., Sun, X., Tian, Y., Chen, J. 및 Ji, R. 교육이 필요 없는 변압기 아키텍처 검색. 컴퓨터 비전 및 패턴 인식(CVPR)에 관한 회의, pp. 10884–10893, 2022.

A. 변압기에서 가중치 공유 전략을 사용한 SNIP에 대한 통계 분석

전체 검색 공간에서 서브넷을 상당히 최적화하므로 기존의 원샷 교육 후 서브넷을 분석합니다.
성능과 모델 품질의 일관성을 유지합니다. 서로 다른 2000개의 서브넷을 임의로 샘플링합니다.
가중치를 상속하는 모델 크기는 슈퍼넷을 형성하고 SNIP 점수가 성능과 0.86 Kendall 상관 관계가 있음을 확인합니다.
모델 크기와 성능의 상관 관계 0.85보다 약간 높습니다. 나쁜 이유를 더 활용하기 위해
전체 검색 공간에서 SNIP의 성능, 서브넷에서 SNIP의 성능을 분석하는 데 전념합니다.
동일한 모델 크기로.

표 7. Case 1과 Case 2에서 분석된 서브넷의 통계
서브넷의 수와 정확도 범위는 최악의 서브넷에서 그룹의 최상의 서브넷까지입니다.

사례 1	그룹 ID 모델	1	2	삼	4	5	6	7	8	9	10	11
	번호. 12		66	220	495	792	924	792	495	220	66	12
	Acc. 스팬 그	0.334	0.542	0.730	0.842	1.008	0.958	0.944	0.820	0.762	0.378	0.336
사례 2	룹 ID 모델 번	1	2	삼	4	5	6	7	8	9	10	11
	호. 12		66	220	495	792	924	792	495	220	66	12
	Acc. 기간	0.280	0.322	0.332	0.372	0.394	0.382	0.386	0.372	0.296	0.240	0.160

다음 두 가지 경우에 분석을 집중합니다.

(Case 1) 임베딩 차원을 192로, 깊이를 12로, 멀티 헤드 어텐션 블록의 헤드 수를 4로 고정합니다.
MLP 비율이 3.5 또는 4인 모든 서브넷을 고려하십시오. 서브넷을 탭으로 그룹화합니다. 7 그룹이 어디에 있는지 보여줍니다.
ID는 4 MLP 비율의 레이어 수와 같습니다.

(Case 2) 각 레이어에서 임베딩 차원을 192, 깊이를 12, MLP 비율을 4로 고정하고 모든 서브넷을 고려합니다.
멀티 헤드 주의 블록에 3개 또는 4개의 헤드가 있습니다. 서브넷을 탭으로 그룹화합니다. 7 은 그룹 ID가
다중 헤드 어텐션 블록에서 4개의 헤드가 있는 레이어의 수.

탭으로. 7은 헤드 수, 특히 MLP 비율의 큰 분포를 찾는 것이 가치가 있습니다.
모델 크기가 동일한 서브넷에서 성능이 확연히 다르기 때문입니다. 게다가 검색 공간이 줄어듭니다.
전체 검색 공간과 동일한 범위의 모델 크기를 유지하면서 집중적으로. 그러나 SNIP 프록시가 실패함을 발견했습니다.
다음 쇼와 같은 일을 할 때.

평균 Best Rank(mBR) (Chen et al., 2021a) 는 다음에서 제로 비용 프록시의 성능을 측정하기 위해 제안됩니다.
서로 다른 네트워크 집합이며 프록시가 선택한 최상의 네트워크의 실제 품질을 강조합니다.
우리의 상황을 위해. 다음과 같이 공식화할 수 있습니다.

$$\mathrm{mBR} = \frac{\sum_{g}(rg - 1)}{\sum_{g}(|g| - 1),}$$

여기서 g는 우리 상황에서 그룹을 나타내고, rg 는 프록시가 선택한 서브넷 정확도의 실제 순위입니다.
식에서 3 및 |g| 이 그룹의 서브넷 수입니다. 프록시가 항상 최상의 네트워크를 선택하는 경우 mBR = 0 이지만
최악의 네트워크의 경우 mBR = 1입니다.

표 8. 사례 1 및 사례 2에서 SNIP의 mBR

		mBR
사례 1	한조각	0.986
	레이어 표준이 있는 SNIP. 0.099	
사례 2	한조각	0.825
	레이어 표준이 있는 SNIP. 0.180	

탭으로. 8 에서 볼 수 있듯이 SNIP는 Case 1과 Case 2 모두에서 거의 최악의 서브넷을 선택합니다. 그런 다음 SNIP 점수를 비교합니다.
슈퍼넷의 각 계층에서 기울기가 약간 더 높기 때문에 하위 계층에서 더 높은 SNIP 점수가 있음을 알 수 있습니다.

이로 인해 SNIP는 수퍼넷과의 가중치 및 기울기 공유 전략에서 더 넓은 하위 계층이 있는 서브넷을 선택하게 됩니다.
이를 고려하여 계층 정규화를 Eq로 하는 SNIP 프록시를 제안합니다. 13회 공연. SNIP의 작은 mBR
Tab에서 계층 정규화를 사용하는 프록시. 8 효과를 확인합니다.

B. 탐욕스러운 알고리즘

Ag ∈ G 의 이성질체 구조 (여기서 G 는 Transformer 검색 공간 A의 분할임)는 동일합니다.
레이어의 동일한 구조로 인해 모든 레이어에 크기, 깊이, 총 헤드 수 및 총 MLP 비율 포함
변압기의. 따라서 각 레이어에 헤드와 MLP 비율을 할당하는 방법만 고려하면 됩니다.
최대 P 놈 의 아키텍처 한조각 각 Ag 에서 여기서는 알고리즘 1 의 헤드 할당 과
MLP 비율은 비슷합니다. 각 레이어에서 헤드의 수와 MLP 비율의 두 가지 선택만 있을 때,
그리디 알고리즘에 의해 유도된 아키텍처는 정확히 최적화된 아키텍처입니다.

알고리즘 1 이성질체 구조에 대한 헤드의 욕심 많은 할당

입력: 레이어 번호 n, 총 머리 수 h, 머리 수 하한 d 및 머리 수 상한 u.
출력: 각 레이어 alc의 헤드 할당.
각 레이어 l에 대해 alc를 alc[l] = d로 초기화합니다.
할당할 왼쪽 머리의 수를 왼쪽 t = h n d로 계산합니다.
i = 1 에서 왼쪽 으로 t 할 때
  alc[l] < u 인 경우 다음을 통해 계층 정규화를 사용하여 각 계층 l 에서 (alc[l] + 1)번째 헤드 의 SNIP 점수 P[l]을 계산합니다.
  등식 13, 그렇지 않으면 P[l] = 0.
  현재 = 인수 maxlP[l]
  alc[cur] = alc[cur] + 1
끝
반환: alc

C. 정규화 및 데이터 확대

3개 슈퍼넷의 학습 용량은 최소 5.4M에서 최대 5.4M까지의 매개변수로 크게 다릅니다.
76M. 따라서 Supernet-Tiny에 대한 정규화 및 데이터 증가를 적절하게 줄이고 그에 따라
과적합 또는 과소적합을 피하기 위한 Supernet-Base의 크기. 자세한 하이퍼 매개변수 설정은
탭. 9. 가치 용어는 주로 timm을 따릅니다 (Wightman, 2019).

표 9. 훈련 정규화 및 데이터 증대의 하이퍼 매개변수.

기법	Supernet-Tiny	슈퍼넷-소형	슈퍼넷 기반
무게 감쇠 레이블	0.02	0.05	0.05
스무딩 Stoch. 깊이	0.1	0.1	0.1
반복 8월	✓	✓	✓
	✓	✓	✓
Mix switch prob		0.5	0.5
Mixup alpha	0	0.8	0.8
Mixup 모드		원소	원소
Cutmix alpha	0	1	1
Rand Augment	m9-n2-mstd0.5-inc1 AutoAug		m10-n3-mstd0.5-inc1
Erasing		v0r-mstd0.5	
prob Erasing	0.25	0.25	0.25
count	1	1	2