(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2020년01월02일
(11) 등록번호 10-2061935
(24) 등록일자 2019년12월26일

(51) 국제특허분류(Int. Cl.)
G06N 3/08 (2006.01) *G06N 3/04* (2006.01)
(52) CPC특허분류
G06N 3/08 (2013.01)
G06N 3/0454 (2013.01)
(21) 출원번호 10-2017-0102269
(22) 출원일자 2017년08월11일
심사청구일자 2017년08월11일
(65) 공개번호 10-2018-0096469
(43) 공개일자 2018년08월29일
(30) 우선권주장
1020170023057 2017년02월21일 대한민국(KR)
(56) 선행기술조사문헌
Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531. 2015.
Romero, Adriana, et al. Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550. 2014.

(73) 특허권자
한국과학기술원
대전광역시 유성구 대학로 291(구성동)
(72) 발명자
김준모
대전광역시 유성구 대학로 291 (구성동, 한국과학기술원)
임준호
대전광역시 유성구 대학로 291 (구성동, 한국과학기술원)
주동규
대전광역시 유성구 대학로 291 (구성동, 한국과학기술원)
(74) 대리인
양성보

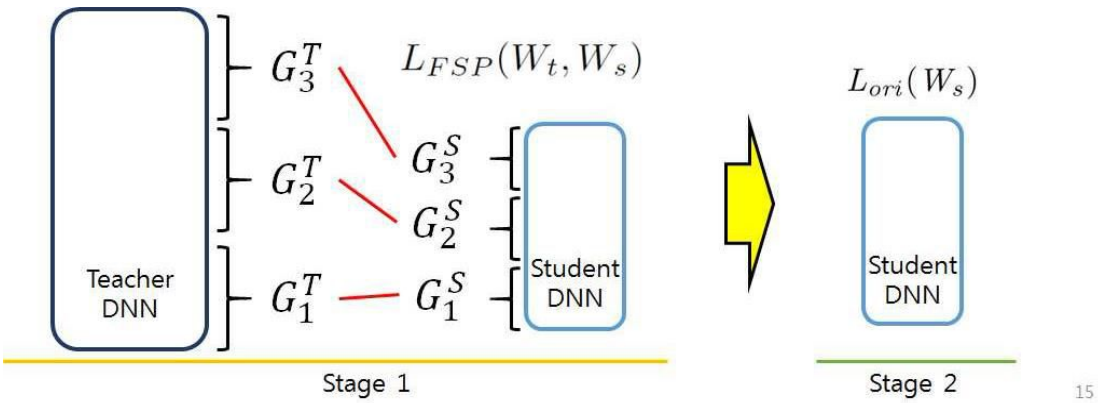
전체 청구항 수 : 총 6 항 심사관 : 서광훈

(54) 발명의 명칭 **딥 신경망을 이용한 정보 이전 방법 및 그 장치**

(57) 요 약

딥 신경망을 이용한 정보 이전 방법 및 그 장치가 개시된다. 본 발명의 일 실시예에 따른 정보 이전 방법은 이미 학습된 제1 딥 신경망의 레이어들 간 정보 흐름을 정의하는 단계; 및 상기 정의된 정보 흐름을 이용하여 학습시키고자 하는 제2 딥 신경망을 학습시키는 단계를 포함하고, 상기 정의하는 단계는 두 레이어 사이의 특성들 간의 흐름을 상기 정보 흐름으로 정의하며, 상기 흐름은 두 벡터 간의 방향성을 나타내는 내적을 사용한 매트릭스로 정의될 수 있다.

대 표 도



명 세 서

청구범위

청구항 1

정의부에서 이미 학습된 제1 딥 신경망의 레이어들 간 정보 흐름을 정의하는 단계; 및

학습부에서 상기 정의된 정보 흐름을 이용하여 학습시키고자 하는 제2 딥 신경망을 학습시키는 단계

를 포함하며,

상기 정의하는 단계는

두 레이어 사이의 특성들 간의 흐름을 상기 정보 흐름으로 정의하며,

상기 흐름은

두 벡터 간의 방향성을 나타내는 내적을 사용한 매트릭스로 정의되는 정보 이전 방법.

청구항 2

삭제

청구항 3

삭제

청구항 4

정의부에서 이미 학습된 제1 딥 신경망의 레이어들 간 정보 흐름을 정의하는 단계; 및

학습부에서 상기 정의된 정보 흐름을 이용하여 학습시키고자 하는 제2 딥 신경망을 학습시키는 단계

를 포함하며,

상기 학습시키는 단계는

상기 제2 딥 신경망의 정보 흐름이 상기 제1 딥 신경망의 정보 흐름과 똑같아지도록 유클리드 로스(Euclidean loss)를 이용하여 상기 제2 딥 신경망의 정보 흐름을 학습시키는 것을 특징으로 하는 정보 이전 방법.

청구항 5

제4항에 있어서,

상기 학습시키는 단계는

상기 정보 흐름이 학습된 제2 딥 신경망의 웨이트(weight)를 초기 웨이트로 하는 분류 로스(classification loss)를 사용한 분류 태스크를 수행함으로써, 상기 제2 딥 신경망을 학습시키는 것을 특징으로 하는 정보 이전 방법.

청구항 6

이미 학습된 제1 딥 신경망의 레이어들 간 정보 흐름을 정의하는 정의부; 및

상기 정의된 정보 흐름을 이용하여 학습시키고자 하는 제2 딥 신경망을 학습시키는 학습부

를 포함하며,

상기 정의부는

두 레이어 사이의 특성들 간의 흐름을 상기 정보 흐름으로 정의하며,

상기 흐름은

두 벡터 간의 방향성을 나타내는 내적을 사용한 매트릭스로 정의되는 정보 이전 장치.

청구항 7

삭제

청구항 8

삭제

청구항 9

이미 학습된 제1 딥 신경망의 레이어들 간 정보 흐름을 정의하는 정의부; 및

상기 정의된 정보 흐름을 이용하여 학습시키고자 하는 제2 딥 신경망을 학습시키는 학습부

를 포함하며,

상기 학습부는

상기 제2 딥 신경망의 정보 흐름이 상기 제1 딥 신경망의 정보 흐름과 똑같아지도록 유클리드 로스(Euclidean loss)를 이용하여 상기 제2 딥 신경망의 정보 흐름을 학습시키는 것을 특징으로 하는 정보 이전 장치.

청구항 10

제9항에 있어서,

상기 학습부는

상기 정보 흐름이 학습된 제2 딥 신경망의 웨이트(weight)를 초기 웨이트로 하는 분류 로스(classification loss)를 사용한 분류 태스크를 수행함으로써, 상기 제2 딥 신경망을 학습시키는 것을 특징으로 하는 정보 이전 장치.

발명의 설명

기 술 분 야

[0001] 본 발명은 정보 이전(knowledge transfer) 기술에 관한 것으로서, 보다 구체적으로 딥 신경망을 구성하는 레이어 간의 상관관계에 대응하는 정보 흐름을 이용하여 새로운 딥 신경망에 정보를 전달할 수 있는 정보 이전 방법 및 그 장치에 관한 것이다.

배 경 기 술

[0003] DNN(Deep Neural Network)은 매우 다양한 분야에서 높은 성능을 자랑하고 있다. 특정 태스크에 맞게 DNN를 설계하고 이를 학습하게 되는데, 대부분의 학습된 DNN은 단순히 추론을 하는 데에만 사용한다. 하지만 이는 이미 학습된 DNN의 능력을 모두 사용했다고 할 수 없다.

[0004] 이미 학습된 네트워크의 능력을 보다 충분히 사용하는 방법 중 하나는 네트워크의 유용한 정보를 추출하고 이를 새로운 네트워크에 전달하는 것이다. 이를 정보 이전(Knowledge transfer) 기술이라 하고, 기존의 많은 연구자들은 이러한 기술에 대해 연구를 진행하였다.

[0005] 종래의 정보 이전 기술에 대한 일 예는, 다크 정보(Dark knowledge)라는 기술을 이용하여 이미 학습된 네트워크

의 출력 특성(output feature)을 소프트하게(softened) 하여 이를 그대로 새로운 네트워크도 만들게끔 학습을 하였다.

[0006] 종래의 정보 이전 기술에 대한 다른 일 예는, 가장 마지막에 있는 출력 특성 뿐만이 아니라 중간에 있는 특성도 동일하게 만들게 하는 로스(loss)를 줌으로써 새로운 네트워크가 티쳐(teacher) 네트워크의 특성을 흉내 낼 수 있게 하였다.

[0007] 하지만, 이러한 기존 기법인 학생(student) DNN으로 하여금 티쳐(teacher) DNN이 생성해내는 특성을 정확히 같이 만들라고 하는 것은 매우 힘든 일이 될 수 있고, 이는 단순히 티쳐 DNN을 흉내 내는 일이라고 할 수 있다.

[0009] 따라서, 새로운 방식의 정보 이전 기술의 필요성이 대두된다.

발명의 내용

해결하려는 과제

[0011] 본 발명의 실시예들은, 딥 신경망을 구성하는 레이어 간의 상관관계에 대응하는 정보 흐름을 이용하여 새로운 딥 신경망에 정보를 이전(또는 전달)할 수 있는 정보 이전 방법 및 그 장치를 제공한다.

과제의 해결 수단

[0013] 본 발명의 일 실시예에 따른 정보 이전 방법은 이미 학습된 제1 딥 신경망의 레이어들 간 정보 흐름을 정의하는 단계; 및 상기 정의된 정보 흐름을 이용하여 학습시키고자 하는 제2 딥 신경망을 학습시키는 단계를 포함한다.

[0014] 상기 정의하는 단계는 두 레이어 사이의 특성들 간의 흐름을 상기 정보 흐름으로 정의하며, 상기 흐름은 두 벡터 간의 방향성을 나타내는 내적을 사용한 매트릭스로 정의될 수 있다.

[0015] 상기 학습시키는 단계는 상기 제2 딥 신경망의 정보 흐름이 상기 제1 딥 신경망의 정보 흐름과 똑같아지도록 상기 제2 딥 신경망의 정보 흐름을 학습시킬 수 있다.

[0016] 상기 학습시키는 단계는 상기 제2 딥 신경망의 정보 흐름이 상기 제1 딥 신경망의 정보 흐름과 똑같아지도록 유클리드 로스(Euclidean loss)를 이용하여 상기 제2 딥 신경망의 정보 흐름을 학습시킬 수 있다.

[0017] 상기 학습시키는 단계는 상기 정보 흐름이 학습된 제2 딥 신경망의 웨이트(weight)를 초기 웨이트로 하는 분류 로스(classification loss)를 사용한 분류 태스크를 수행함으로써, 상기 제2 딥 신경망을 학습시킬 수 있다.

[0019] 본 발명의 일 실시예에 따른 정보 이전 장치는 이미 학습된 제1 딥 신경망의 레이어들 간 정보 흐름을 정의하는 정의부; 및 상기 정의된 정보 흐름을 이용하여 학습시키고자 하는 제2 딥 신경망을 학습시키는 학습부를 포함한다.

[0020] 상기 정의부는 두 레이어 사이의 특성들 간의 흐름을 상기 정보 흐름으로 정의하며, 상기 흐름은 두 벡터 간의 방향성을 나타내는 내적을 사용한 매트릭스로 정의될 수 있다.

[0021] 상기 학습부는 상기 제2 딥 신경망의 정보 흐름이 상기 제1 딥 신경망의 정보 흐름과 똑같아지도록 상기 제2 딥 신경망의 정보 흐름을 학습시킬 수 있다.

[0022] 상기 학습부는 상기 제2 딥 신경망의 정보 흐름이 상기 제1 딥 신경망의 정보 흐름과 똑같아지도록 유클리드 로스(Euclidean loss)를 이용하여 상기 제2 딥 신경망의 정보 흐름을 학습시킬 수 있다.

[0023] 상기 학습부는 상기 정보 흐름이 학습된 제2 딥 신경망의 웨이트(weight)를 초기 웨이트로 하는 분류 로스(classification loss)를 사용한 분류 태스크를 수행함으로써, 상기 제2 딥 신경망을 학습시킬 수 있다.

발명의 효과

[0025] 본 발명의 실시예들에 따르면, 딥 신경망을 구성하는 레이어 간의 상관관계에 대응하는 정보 흐름을 이용하여 새로운 딥 신경망에 정보를 이전(또는 전달)할 수 있다.

[0026] 본 발명의 실시예들에 따르면, 기존의 방법들에 비해 덜 제한적이며 자유롭고, 압축 시 성능 면에서 기존 방법에 비해 뛰어난 효과가 있다.

[0027] 본 발명의 실시예들에 따르면, 딥 신경망을 이용한 영상 인식, 물체 인식, 판별 및 오류 검출 등 다양한 분야에

서 학습 속도를 향상시킬 수 있으며, 컴퓨터에 비해서 메모리와 연산 능력이 제한적인 스마트 디바이스 예를 들어, 스마트폰, 스마트 와치 등에서 기존의 방법들보다 높은 성능을 기대할 수 있다.

도면의 간단한 설명

[0028] 도 1은 본 발명의 일 실시예에 따른 정보 이전 방법에 대한 동작 흐름도를 나타낸 것이다.

도 2는 정보 흐름을 정의하는 과정을 설명하기 위한 일 예시도를 나타낸 것이다.

도 3은 정보 흐름을 이용하여 DNN을 학습하는 과정을 설명하기 위한 일 예시도를 나타낸 것이다.

도 4는 본 발명의 방법에 대한 결과의 일 예시도를 나타낸 것이다.

도 5는 본 발명의 일 실시예에 따른 방법을 설명하기 위한 일 예시도를 나타낸 것이다.

도 6은 본 발명의 일 실시예에 따른 정보 이전 장치에 대한 구성을 나타낸 것이다.

발명을 실시하기 위한 구체적인 내용

[0029] 이하, 본 발명에 따른 실시예들을 첨부된 도면을 참조하여 상세하게 설명한다. 그러나 본 발명이 실시예들에 의해 제한되거나 한정되는 것은 아니다. 또한, 각 도면에 제시된 동일한 참조 부호는 동일한 부재를 나타낸다.

[0031] 본 발명의 실시예들은, 딥 신경망을 구성하는 레이어 간의 상관관계를 정보 흐름으로 정의하고, 이렇게 정의된 정보 흐름을 이용하여 새로운 딥 신경망 예를 들어, 학생 DNN에 정보를 이전(또는 전달)하는 것을 그 요지로 한다.

[0032] 여기서, 본 발명은 미리 학습된 제1 DNN 예를 들어, 티쳐 DNN을 구성하는 레이어들 간의 상관관계 즉, 정보의 흐름을 학습시키고자 하는 제2 DNN을 구성하는 레이어들 간의 정보 흐름과 같도록 제2 DNN을 학습시킬 수 있다.

[0034] 도 1은 본 발명의 일 실시예에 따른 정보 이전 방법에 대한 동작 흐름도를 나타낸 것이다.

[0035] 도 1을 참조하면, 본 발명의 일 실시예에 따른 정보 이전 방법은 이미 학습된 제1 딥 신경망(DNN) 예를 들어, 티쳐 DNN을 구성하는 레이어들 간 정보 흐름을 정의한다(S110).

[0036] 여기서, 단계 S110은 이미 학습된 티쳐 DNN을 구성하는 레이어들에서 두 레이어들 사이의 특성들 간의 흐름을 정의할 수 있으며, 흐름은 두 레이어의 특성들 간의 내적을 계산함으로써, 정의될 수 있다. 또한, 흐름은 두 레이어의 특성들에 대한 두 벡터 간의 방향성을 나타내는 내적을 사용한 매트릭스로 정의될 수도 있다.

[0037] 예를 들어, 단계 S110은 도 2에 도시된 바와 같이, 티쳐 DNN을 구성하는 레이어의 입력과 출력 간의 특성들(feature map)을 이용하여 티쳐 DNN의 정보 흐름을 나타내는 매트릭스 예를 들어, Gram Matrix 또는 FSP(Flow of Solving a Problem) Matrix로 정보 흐름을 정의할 수 있다. 이렇게 정의된 정보 흐름 또는 매트릭스는 학습시키고자 하는 새로운 DNN 예를 들어, 학생 DNN을 학습시키는데 사용될 수 있다(transfer the distilled knowledge).

[0038] 예컨대, 두 레이어 간의 흐름을 나타내는 매트릭스가 G라고 가정하면, G는 $G(i,j) = f_i^1 \cdot f_j^2$ 로 나타낼 수 있으며, f_i^1는 한 레이어의 i번째 채널 특성을 의미하고, f_i^2 는 또 다른 레이어의 j번째 채널 특성을 의미할 수 있다.

[0039] 본 발명에서는 티쳐 DNN의 입력 공간(input space)에서 출력 공간(output space)로 만드는 정보의 흐름이 학생 DNN을 빠르게 학습시킬 수 있는 중요한 특성이 되며, 이러한 정보 흐름은 티쳐 DNN을 구성하는 레이어들 간의 상관 관계로 정의할 수도 있다.

[0040] 단계 S110에 의해 티쳐 DNN의 정보 흐름의 정의되면, 이렇게 정의된 정보 흐름을 이용하여 학습시키고자 하는 제2 DNN 예를 들어, 학생 DNN을 학습시킨다(S120).

[0041] 여기서, 단계 S120은 제2 DNN의 정보 흐름이 제1 DNN의 정보 흐름과 똑같아지도록 제1 DNN의 정보 흐름을 이용하여 제2 DNN의 정보 흐름을 학습시킬 수 있다.

[0042] 본 발명에서 단계 S120은 두 단계의 로스를 이용하여 제2 DNN을 학습시킬 수 있으며, 유클리드 로스(Euclidean loss)와 분류 로스(classification loss)의 두 로스를 이용하여 제2 DNN을 학습시킬 수 있다.

[0043] 즉, 도 3에 도시된 바와 같이, 단계 S120은 학생 DNN의 정보 흐름 예를 들어, 매트릭스(G^S)가 티쳐 DNN에서 정의된 정보 흐름 예를 들어, 매트릭스(G^T)와 똑같아지도록 유클리드 로스(Euclidean loss)를 이용하여 티쳐 DNN의 정보 흐름을 학습시킬 수 있다. 예컨대, 미리 학습된 티쳐 DNN의 매트릭스를 G1이라 가정하고, 학생 DNN의 매트릭스를 G2라 가정하면, 두 매트릭스를 똑같이 맞추도록 유클리드 로스(Euclidean loss) $L1 = \|G1 - G1\|^2$를 사용할 수 있다.

[0044] 그리고, 분류 로스를 이용하여 유클리드 로스에 의해 학습된 학생 DNN을 학습시킬 수 있다. 즉, 단계 S120은 유클리드 로스에 의해 학습된 학생 DNN의 웨이트를 초기 웨이트로 하여 분류 로스를 사용한 분류 태스크를 수행함으로써, 학생 DNN의 최종 학습시키고, 이를 통해 티쳐 DNN에서 학생 DNN으로 정보 전이를 빠르게 수행할 수 있다.

[0045] 이러한 본 발명은 빠른 최적화(fast optimization) 기법과 함께 정보 이전 기술을 이용하여 DNN의 성능을 향상시킬 수 있다.

[0047] 도 4는 본 발명의 방법에 대한 결과의 일 예시도를 나타낸 것으로, CIFAR-10의 데이터 셋을 사용한 경우의 결과를 나타낸 것이다.

[0048] 여기서, 데이터 셋은 10개의 클래스로 구성되어있으며, 총 5만장의 트레이닝 셋(training set)과 만장의 테스트 셋(test set)을 가지고 있으며, 티쳐 네트워크는 잔여 네트워크(Residual network) 중 26층짜리의 네트워크를 사용하고, 학생 네트워크는 티쳐 네트워크와 같은 구조를 사용할 수 있다.

[0049] 도 4에서의 매트릭스는 특성의 공간 사이즈(spatial size)가 줄어드는 총 3 부분에서 추출되었으며, 도 4에 도시된 바와 같이, 본 발명의 정보 전이 방법을 사용하였을 경우, 기존 기법보다 빠르게 학습되는 것을 볼 수 있으며, 약 3배 이상 빠르게 수렴하는 것을 알 수 있다.

[0051] 아래 표 1은 위의 열에서부터 각각 26층짜리 residual network를 64000 iteration으로 돌린 것, 21000 iteration으로 돌린 것, 본 발명을 이용하여(iteration 21000) 학습을 한 후, 21000 iteration을 돌린 것에 대한 성능 및 앙상블 성능을 보인 것이다.

표 1

[0052]

	Accuracy1	Accuracy2	Accuracy3	Ensemble
Original	91.61	91.56	92.09	93.48
Original with 1/3 iter	90.47	90.83	90.62	92.6
Proposed method	92.28	92.24	92.07	93.26

[0054] 표 1에서 보는 바와 같이 기존방식(original)의 경우, 64000 iteration에서 21000 iteration으로 1/3배가량 iteration수가 줄어들었을 때 약 1.5%가량 성능 하락을 보이는 것을 알 수 있는 반면, 본 발명은 1/3배의 iteration만 돌렸음에도 불구하고 기존의 성능을 그대로 보이는 것을 알고 있다. 이는 본 발명에서 정보 흐름의 매트릭스를 활용한 초기 웨이트가 좋은 초기 포인트가 되었음을 알 수 있다.

[0056] 아래 표 2는 본 발명의 네트워크에서 추출된 정보를 얕은 네트워크로 전달했을 경우의 성능 향상을 기존 방식과 비교하여 나타낸 것으로, 티쳐 네트워크는 26개의 층을 사용한 잔여 네트워크이고, 학생 네트워크는 8개의 레이어를 사용한 결과를 나타낸 것이다. 여기서, 매트릭스는 특성의 공간 사이즈(spatial size)가 줄어드는 총 3 부분에서 추출될 수 있다.

표 2

[0058]

	Accuracy
Original	87.9
Romero	88.34
Proposed	88.62

[0060] 표 2를 통해 알 수 있듯이, 본 발명에 따른 방법(proposed)은 일반적인 기법(original)을 사용하여 학습 하였을

때보다 약 0.7%가량 성능이 증가된 것을 알 수 있다.

[0062] 이와 같이, 본 발명에 따른 방법은 기존 방식에 비해서 성능이 향상되고, 네트워크가 빠르게 학습될 수 있으며, 여러 네트워크의 앙상블이 필요한 태스크의 경우와 작은 네트워크를 필요로 하는 태스크의 경우 모두에 유용하게 사용될 수 있다.

[0064] 이러한 본 발명에 대해 조금 더 구체적으로 설명하면 다음과 같다.

[0065] 본 발명에 따른 정보 이전 방법은 제1 딥 신경망 예를 들어, 티쳐 DNN의 중요 정보를 정의하고 다른 DNN 예를 들어, 학생 DNN에 디스틸드 정보(distilled knowledge)를 전달 하는 것이다.

[0067] **본 발명에서의 디스틸드 정보**

[0068] DNN은 레이어간 특성들을 생성한다. 상위 레이어 특성들은 주요 태스크를 수행할 수 있는 유용한 특성들에 가깝다. DNN의 입력을 질문으로 하고 출력을 대답으로 한다면, 중간 결과로서 DNN의 중간에서 생성되는 특성들을 생각할 수 있으며, 이러한 관점에서 정보 이전 기술은 학생 DNN이 티쳐 DNN의 중간 결과를 단순히 모방하는 것으로 볼 수 있다. 그러나, DNN의 경우 입력으로부터 출력을 생성하는 문제를 해결할 수 있는 다양한 방법이 있으며, 이런 의미에서 티쳐 DNN의 생성된 특성들을 모방하는 것은 학생 DNN에 대한 제약이 될 수 있다.

[0069] 사람의 경우, 선생은 문제를 푸는 과정을 설명하고, 학생은 문제를 푸는 과정을 학습한다. 학생 DNN은 특정 질문이 입력되는 경우 중간 출력을 학습할 필요는 없지만 특정 형태의 질문을 마주하였을 때 그 문제를 해결할 수 있는 방법을 학습할 수 있다. 이와 같은 방법으로 본 발명은 중간 결과를 티칭하는 것보다 더 나은 방법을 제공할 수 있다.

[0071] **디스틸드 정보의 수학적인 표현**

[0072] 본 발명은 두 중간 결과 간의 관계를 정의할 수 있다. DNN의 경우 관계는 두 레이어들의 특성들 간 방향에 의해 수학적으로 고려될 수 있다. 본 발명에서는 솔루션 프로세서의 플로우(또는 흐름)를 FSP 매트릭스로 나타낼 수 있다. FSP 매트릭스 $G \in \mathbb{R}^{m \times n}$ 는 두 레이어들의 특성들에 의해 생성된다. 선택된 레이어들 중 하나가 특성 맵 $F^1 \in \mathbb{R}^{h \times w \times m}$ 을 생성한다고 하면 다른 선택된 레이어는 특성 맵 $F^2 \in \mathbb{R}^{h \times w \times n}$ 를 생성한다. 여기서, h, w, m은 각각 채널의 높이, 폭과 수를 의미할 수 있다.

[0073] 그러면, FSP 매트릭스 $G \in \mathbb{R}^{m \times n}$ 는 아래 <수학식 1>과 같이 계산될 수 있다.

[0075] [수학식 1]

$$G_{i,j}(x;W) = \sum_{s=1}^{h} \sum_{t=1}^{w} \frac{F^1_{s,t,i}(x;W) \times F^2_{s,t,j}(x;W)}{h \times w}$$

[0076]

[0078] 여기서, x와 W는 각각 DNN의 입력 이미지와 웨이트들을 의미할 수 있다. 본 발명에서는 CIFAR-10 데이터셋에 의해 트레이닝된 8, 26, 36 레이어들을 가지는 잔여 네트워크를 이용하여 설명할 수 있다. 공간적인 사이즈가 변화하는 CIFAR-10 데이터셋에 대해 잔여 네트워크는 세가지 포인트들을 가지고 있을 수 있으며, 본 발명은 도 5에 도시된 바와 같이 FSP 매트릭스을 생성하는 여러 포인트들을 선택할 수 있다.

[0080] **FSP 매트릭스의 로스**

[0081] 학생 네트워크를 돕기 위하여, 본 발명은 티쳐 네트워크의 디스틸트 정보를 학생 네트워크로 이전 또는 전달하는 것으로, 상술한 바와 같이, 디스틸트 정보는 솔루션 과정의 플로우에 대한 정보를 포함하는 FSP 매트릭스 형태로 나타낼 수 있다. 티쳐 네트워크에 의해 생성되는 n FSP 행렬을 $G_i^T, i = 1, \ldots, n$, 학생 네트워크에 의해 생성되는 n FSP 행렬을 $G_i^S, i = 1, \ldots, n$ 라 가정하면, 본 발명은 같은 공간 사이즈를 가지는 티쳐 네트워크와 학생 네트워크 간 FSP 행렬의 쌍 $(G_i^T, G_i^S), i = 1, \ldots, n$ 을 고려한다. 각 쌍에 대한 비용 함수로 제곱된 L2 표준을 이용하며, 디스틸드 정보 이전에 대한 비용 함수는 아래 수학식 2와 같이 정

의될 수 있다.

[0083] [수학식 2]

$$L_{FSP}(W_t, W_s) = \frac{1}{N}\sum_{x}\sum_{i=1}^{n}\lambda_i \times \|(G_i^T(x;W_t) - G_i^S(x;W_s)\|_2^2$$

[0084]

[0086] 여기서. λ_i는 각 로스 텀에 대한 웨이트를 의미하고, N은 데이터 포인트들의 수를 의미할 수 있다.

[0087] 본 발명은 전체 로스 텀이 똑같이 중요하다 가정한다. 즉, 똑같은 λ_i를 사용한다.

[0089] **학습 프로시져**

[0090] 본 발명에 따른 정보 이전 방법은 티쳐 네트워크에 의해 생성된 디스틸드 정보를 사용한다. 본 발명에서 티쳐 네트워크가 무엇인지 명확하게 설명하기 위하여 다음과 같은 두 조건을 정의한다.

[0091] 첫 째, 티쳐 네트워크는 미리 설정된 데이터셋에 의해 미리 트레이닝되어 있어야 한다. 이 때, 티쳐 네트워크를 트레이닝하는 데이터셋은 학생 네트워크가 학습한 데이터셋과 같거나 다를 수 있다. 티쳐 네트워크는 전이 학습 태스크의 경우 학생 네트워크의 데이터셋과 다른 데이터셋을 사용할 수 잇다.

[0092] 둘 째, 티쳐 네트워크는 학생 네트워크보다 더 깊거나 더 얇을 수 있는데, 본 발명의 상세한 설명에서는 티쳐 네트워크가 학생 네트워크에 비해 그 깊이가 같거나 더 깊은 경우로 설명한 것이다.

[0093] 학습 프로시져는 트레이닝의 두 단계를 포함한다. 첫 째, 본 발명은 학생 네트워크의 FSP 매트릭스를 티쳐 네트워크의 FSP 매트릭스와 같게 만들기 위하여 로스 함수 L_{FSP}를 최소화한다. 첫 번째 단계를 거친 학생 네트워크는 두 번째 단계에서 주요 태스크 로스에 의해 트레이닝 된다. 본 발명에서는 주요 태스크 로스로, 소프트맥스 크로스 엔트로피 로스 Lori를 사용할 수 있다.

[0094] 즉, 본 발명의 학습 프로시져는 아래 <수학식 3>을 이용하여 첫 번째 단계인 FSP 매트릭스를 학습하고, 첫 번째 단계에서 FSP 매트릭스가 학습된 학생 네트워크에 대하여, 아래 <수학식 4>를 이용하여 두 번째 단계인 오리지널 태스크를 트레이닝한다.

[0096] <수학식 3>

$$W_s = \arg\min_{W_s} L_{FSP}(W_t, W_s)$$

[0097]

[0099] 여기서, Ws와 Wt는 학생 네트워크의 웨이트들과 티쳐 네트워크의 웨이트들을 의미할 수 있다.

[0101] <수학식 4>

$$W_s = \arg\min_{W_s} L_{ori}(W_s)$$

[0102]

[0104] 딥 잔여 네트워크(deep residual network)는 바로가기 연결(shortcut connection)을 통해 앙상블 구조를 만들 수 있으며, 바로가기 연결은 더 깊은 네트워크의 트레이닝을 수행할 수 있다.

[0105] 도 5는 본 발명의 일 실시예에 따른 방법을 설명하기 위한 일 예시도를 나타낸 것으로, 32 레이어들의 티쳐 네트워크와 14 레이어들의 학생 네트워크를 예로 설명하였지만, 티쳐 네트워크의 레이어 수와 학생 네트워크의 레이어 수는 변화될 수 있다.

[0106] 도 5에 도시된 바와 같이, FSP 행렬은 같은 공간 사이즈를 유지하는 세 개의 섹션들에서 추출될 수 있으며, 첫 번째 단계에서 학생 네트워크와 티쳐 네트워크이 FSP 행렬 간 거리가 최소가되도록 학생 네트워크를 트레이닝하고, 학생 DNN의 미리 트레이닝된 웨이트들은 두 번째 단계에서 초기 웨이트로 사용된다. 두 번째 단계는 노말 트레이닝 프로시져를 나타낸다.

[0107] 즉, 딥 잔여 네트워크는 세 개의 섹션을 포함하고, FSP 매트릭스를 만들기 위하여 두 개의 레이어들을 선택하는데, 두 개의 레이어들을 선택하는 방법에는 제한이 없다. 예를 들어, 첫 번째 레이어와 마지막 레이어를 FSP 매

트릭스를 만들기 위하여 선택할 수도 있다. FSP 매트릭스는 같은 공간 사이즈를 가지는 두 레이어 특성들에 의해 생성되기 때문에 두 레이어 특성들의 사이즈가 다를 경우 동일 공간 사이즈를 만들기 위하여 맥스 풀링 레이어를 사용할 수 있다.

[0108] 티쳐 DNN의 사이즈는 학생 DNN의 사이즈보다 더 크며, 학생 DNN은 티쳐 DNN에서 잔여 모듈들의 수를 간단하게 줄임으로써, 구성될 수 있다. 따라서, 학생 DNN은 티쳐 DNN보다 더 작은 파라미터들을 사용할 수 있다.

[0110] 도 5는 본 발명의 일 실시예에 따른 정보 이전 장치에 대한 구성을 나타낸 것으로, 상술한 도 1 내지 도 4의 방법을 수행하는 장치에 대한 구성을 나타낸 것이다.

[0111] 도 5를 참조하면, 본 발명의 일 실시예에 따른 정보 이전 장치(500)는 정의부(510)와 학습부(520)를 포함한다.

[0112] 정의부(510)는 이미 학습된 제1 딥 신경망 예를 들어, 티쳐 DNN의 레이어들 간 정보 흐름을 정의한다.

[0113] 여기서, 정의부(510)는 두 레이어 사이의 특성들 간의 흐름을 정보 흐름으로 정의하며, 흐름은 두 벡터 간의 방향성을 나타내는 내적을 사용한 매트릭스로 정의될 수 있다.

[0114] 학습부(520)는 정의부(510)에 의해 정의된 티쳐 DNN의 정보 흐름을 이용하여 학습시키고자 하는 제2 딥 신경망 예를 들어, 학생 DNN을 학습시킨다.

[0115] 여기서, 학습부(520)는 제2 딥 신경망의 정보 흐름이 제1 딥 신경망의 정보 흐름과 똑같아지도록 제2 딥 신경망의 정보 흐름을 학습시킬 수 있다.

[0116] 예컨대, 학습부(520)는 제2 딥 신경망의 정보 흐름이 제1 딥 신경망의 정보 흐름과 똑같아지도록 유클리드 로스(Euclidean loss)를 이용하여 제2 딥 신경망의 정보 흐름을 학습시킬 수 있으며, 추가적으로, 정보 흐름이 학습된 제2 딥 신경망의 웨이트(weight)를 초기 웨이트로 하는 분류 로스(classification loss)를 사용한 분류 태스크를 수행함으로써, 제2 딥 신경망을 학습시킬 수 있다.

[0118] 비록, 도 5의 장치에서 그 설명이 생략되었더라도 도 5의 장치는 도 1 내지 도 4에서 설명한 내용을 모두 포함할 수 있으며, 정보 이전과 관련된 모든 기능 또는 내용을 포함할 수 있다는 것은 이 기술 분야에 종사하는 당업자에게 있어서 자명하다.

[0120] 이상에서 설명된 시스템 또는 장치는 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 시스템, 장치 및 구성요소는, 예를 들어, 프로세서, 콘트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPA(field programmable array), PLU(programmable logic unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(processing element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 콘트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.

[0121] 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 가상 장치(virtual equipment), 컴퓨터 저장 매체 또는 장치, 또는 전송되는 신호 파(signal wave)에 영구적으로, 또는 일시적으로 구체화(embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

[0122] 실시예들에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판

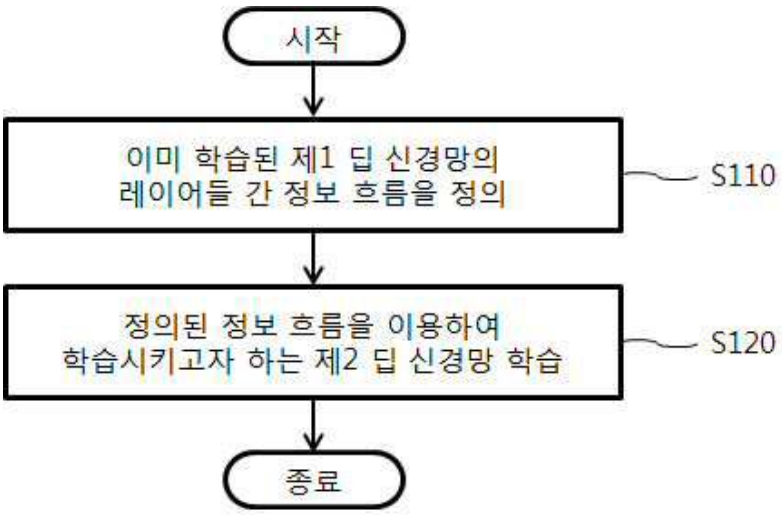
독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다. 상기된 하드웨어 장치는 실시예의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

[0123] 이상과 같이 실시예들이 비록 한정된 실시예와 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기의 기재로부터 다양한 수정 및 변형이 가능하다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.

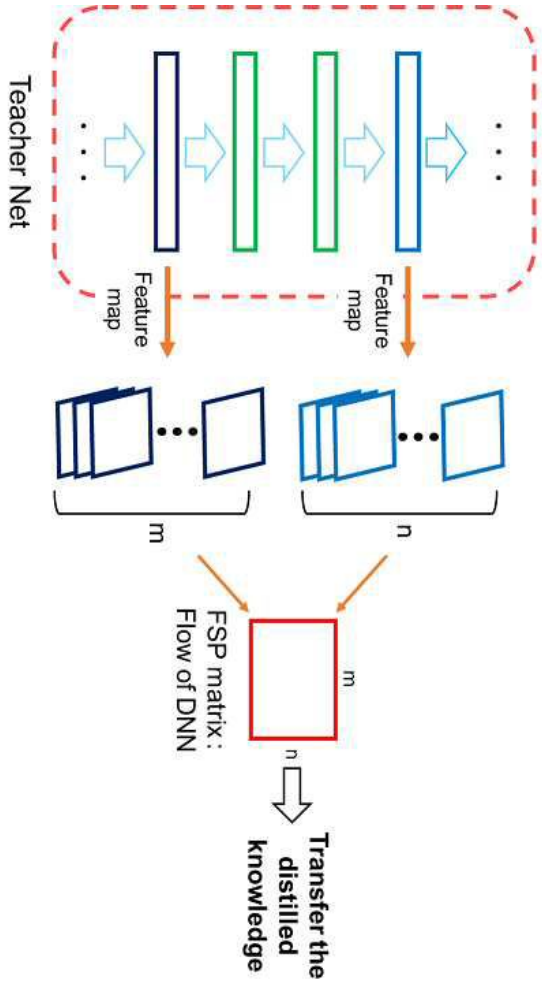
[0124] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

도면

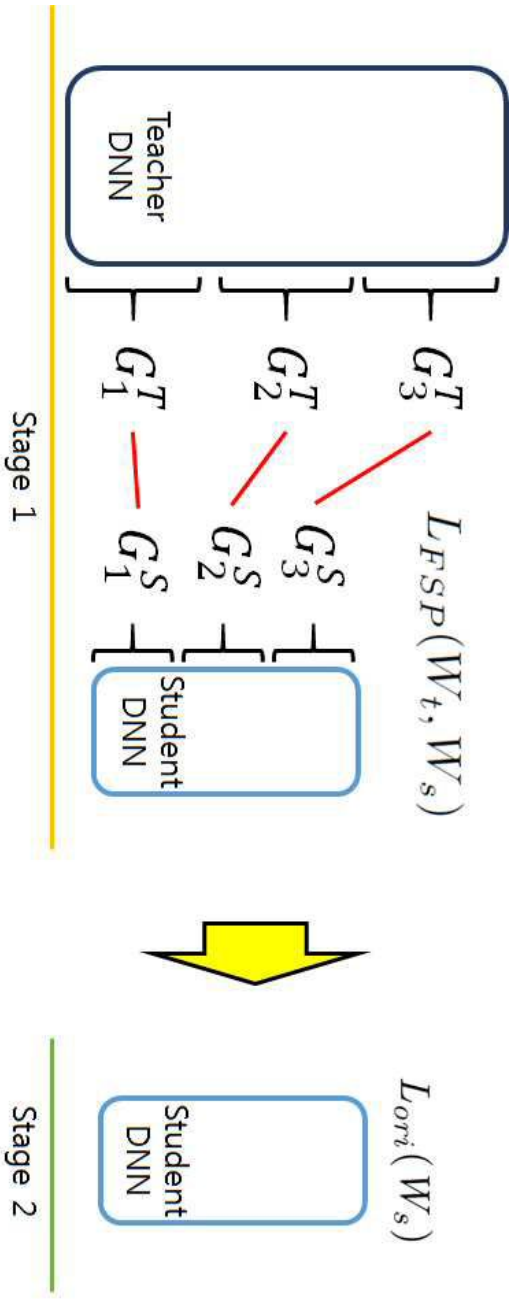
도면1



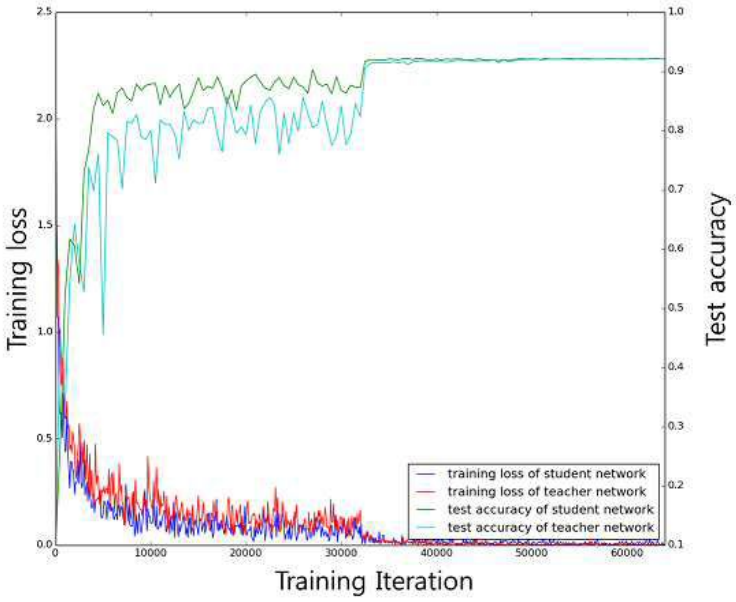
도면2


Teacher Net
Feature map
Feature map
m
n
FSP matrix : Flow of DNN
m
n
Transfer the distilled knowledge

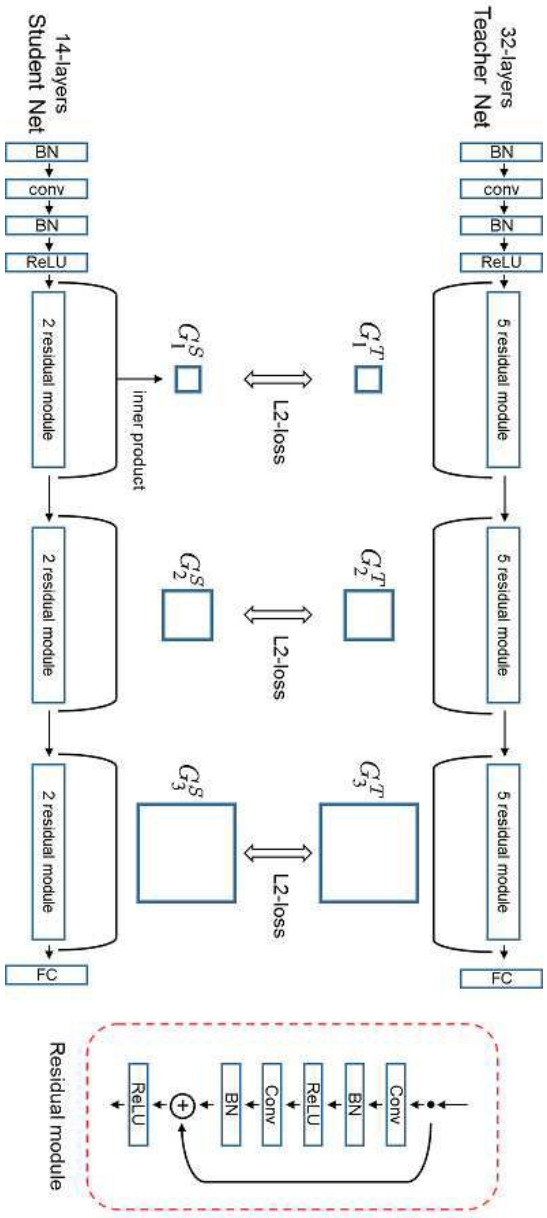
도면3


Stage 1
Teacher DNN
G_1^T
G_2^T
G_3^T
G_1^S
G_2^S
G_3^S
Student DNN
$L_{FSP}(W_t, W_s)$
Stage 2
Student DNN
$L_{ori}(W_s)$

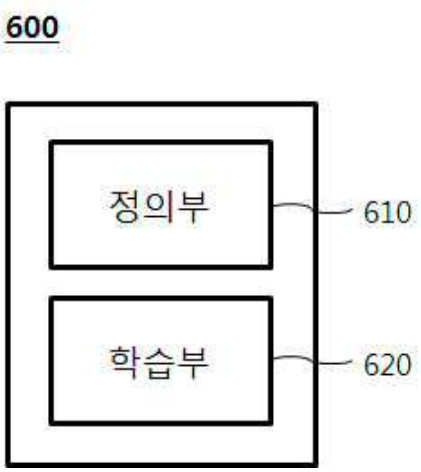
도면4


Training loss
Test accuracy
Training Iteration
training loss of student network
training loss of teacher network
test accuracy of student network
test accuracy of teacher network
0
10000
20000
30000
40000
50000
60000
0.0
0.5
1.0
1.5
2.0
2.5
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

도면5


14-layers
Student Net
32-layers
Teacher Net
BN
conv
BN
ReLU
2 residual module
5 residual module
inner product
L2-loss
FC
Residual module
ReLU
BN
Conv

도면6


600
정의부
610
학습부
620