

[논문 리뷰] End-to-End Multi-Task Learning with Attention

[통계](#) [수정](#) [삭제](#)

diddu · 방금 전

0

[논문](#)

논문 리뷰

▼ [목록 보기](#)

2/2



💡 End-to-End Multi-Task Learning with Attention_논문 리뷰

📌 Introduction

training multiple tasks whilst learning a shared representation two challenges

1) Network Architecture (how to share)

- Network Architecture는 task-shared와 task-specific feature 모두를 표현해야 한다.
- to avoid overfitting (learn generalisable representation)
- to avoid underfitting (learn features tailored to each task)

2) Loss function (how to balance task)

- automatically or design a network

위 두 가지 중 하나만 해결한 연구들은 존재하지만, 두 가지 모두를 해결한 연구는 없었다. 이 연구에서는 위의 두 가지 이슈를 모두 해결하기 위해

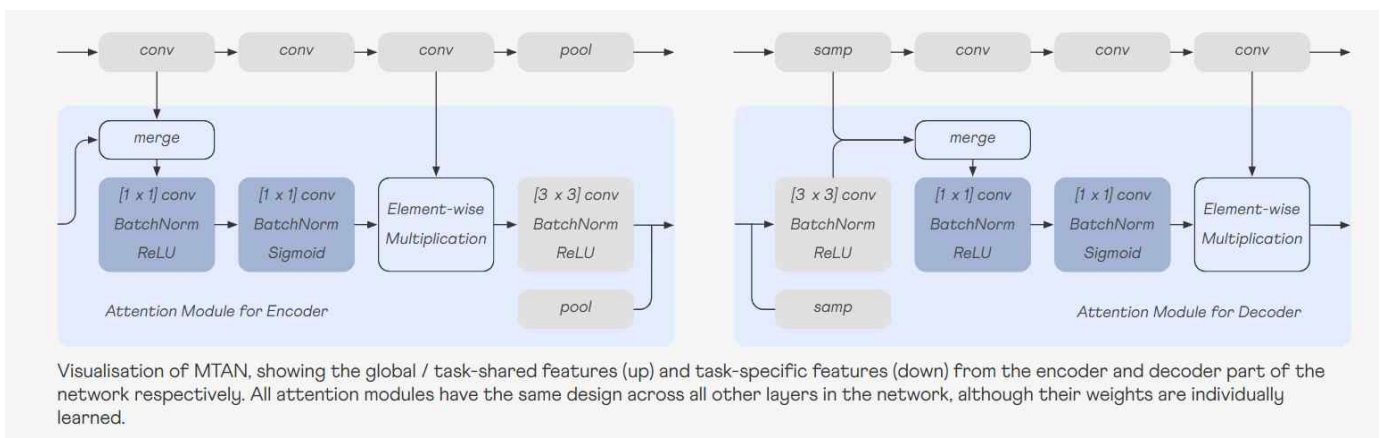
- (i) task-shared와 task-specific feature 모두를 자동으로 배울 수 있는 network design
- (ii) loss weighting scheme 선택에 대한 내재적인 robustness를 배우는 일체화된 접근법을 제안한다.

📌 Architecture Design

MTAN은 두 가지 주요 컴포넌트로 구성된다.

- **a single shared network** : 특정 task에 기반하여 설계될 수 있다. (learns a compact global feature pool across all tasks)
- **K task-specific attention networks** : 각각의 task-specific 네트워크는 attention 모듈들로 이루어져 있다. (each attention module applies a soft attention mask)

1. shared network에서 생성한 feature map
2. 이 위에 attention mask(=feature selector_end to end automatically learned)를 적용
3. 해당 task에 필요한 특정 정보를 추출



→ 각 태스크마다 매우 적은 양의 파라미터를 도입함으로써, 네트워크의 효율성을 높인다.

2) $[3 \times 3]$ convolutional layer는 공유 feature extractor를 나타내며, 다른 attention module로 전달하기 위해 사용된다.

3) pooling 또는 sampling layer가 해당 해상도에 맞추기 위해 따라온다.

- **Performance Expectation**

만약 어떤 attention mask가 1에 가깝다면 (즉, identity map이 되면), attended features는 global feature maps과 동일해진다.

즉, 해당 task들이 모든 features를 공유하게 된다.

: 이러한 설계 방식은 MTAN이 다중 task 학습에서 파라미터 효율성과 성능 사이의 균형을 잘 유지하며 각각의 task에 필요한 고유한 feature들과 공통적인 feature들 모두를 잘 추출할 수 있음을 의미

Experiments

- **Baselines**

5가지 다른 네트워크 아키텍처(2개의 단일 태스크 + 3개의 다중 태스크)를 기반으로 MTAN을 비교한다. 모두 SegNet이라는 동일한 backbone을 기반으로 한다.

- Single-Task, One Task : 단일 태스크 학습을 위한 기본 SegNet
- Single-Task, STAN : 단일 태스크를 수행하는 동안 제안된 MTAN을 직접 적용하는 Single-Task Attention Network
- Multi-Task, Split (Wide, Deep) : 마지막 레이어에서 각 특정 task에 대한 최종 예측을 위해 분할하는 표준 다중 태스크 학습(hard-parameter sharing)
- Multi-task, Dense : 공유 네트워크와 task-specific 네트워크가 함께 있으며, 각 task-specific 네트워크는 shared 네트워크로부터 모든 feature를 받음

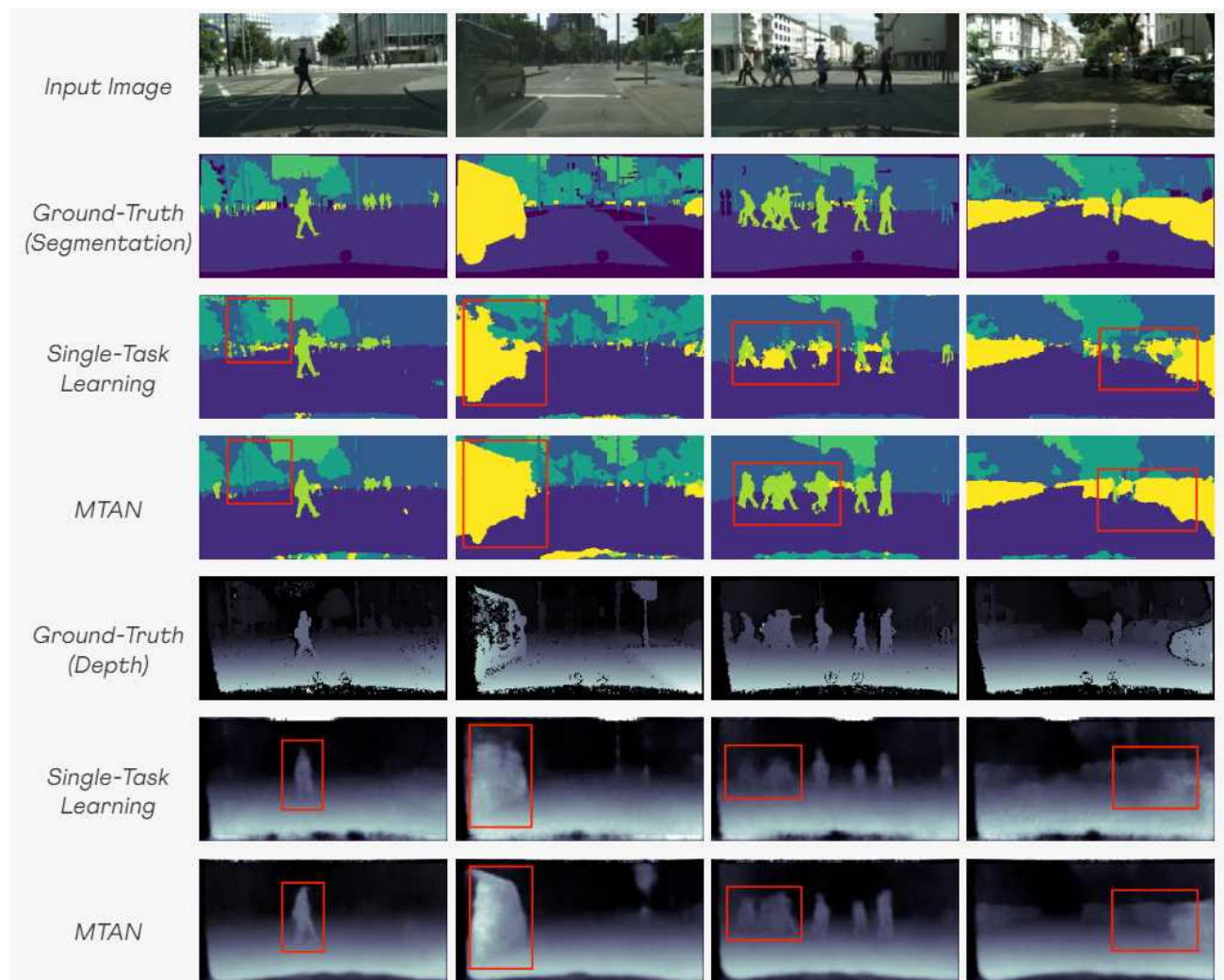
- **Dynamic Weight Average (DWA)**

$$\sum_i \exp(w_i(t-1)/T) \quad L_k(t-2)$$

Task Weighting Training Time Temperature (Softness of Task Weighting) Relative Loss Change

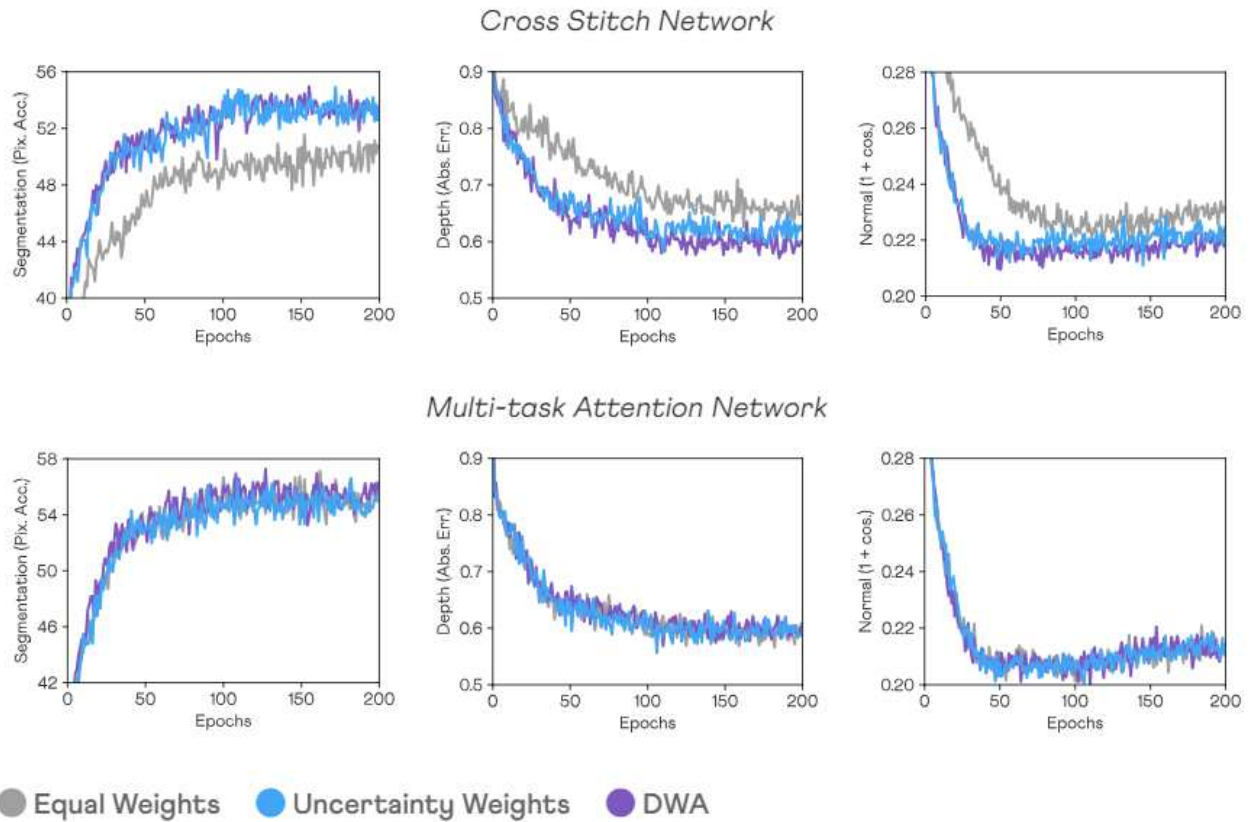
DWA는 GradNorm에서 영감을 받아 각 작업의 손실 변화율에 따라 작업 가중치를 평균화하는 방식으로 학습된다.

• Results on Image-to-Image Predictions



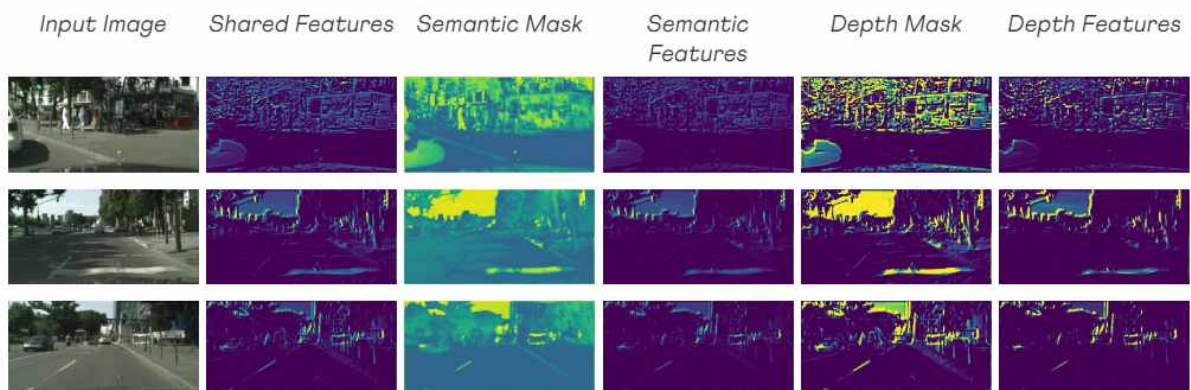
MTAN이 모든 baseline들을 능가한다. CityScapes 데이터셋에서 객체의 경계가 확실히 더

- Robustness to Task Weighting Schemes



MTAN은 다양한 손실 함수 가중치 체계에 걸쳐 높은 성능과 robustness를 유지

- Visualisation of Attended Features



Visualisation of the first layer of 7-class semantic and depth attention features of our proposed network. The colours for each image are rescaled to fit the data.

시각과아예 각 task에 유용한 부분에 집중하고 정보가 없는 부분을 마스킹하는 feature selector

- Results on Many-to-Many Predictions (Visual Decathlon Challenge)

Method	#P.	ImNet.	Airc.	C100	DPed	DTD	GTSR	Flwr	Oglt	SVHN	UCF	Mean	Score
Scratch [23]	10	59.87	57.10	75.73	91.20	37.77	96.55	56.3	88.74	96.63	43.27	70.32	1625
Finetune [23]	10	59.87	60.34	82.12	92.82	55.53	97.53	81.41	87.69	96.55	51.20	76.51	2500
Feature [23]	1	59.67	23.31	63.11	80.33	45.37	68.16	73.69	58.79	43.54	26.8	54.28	544
Res. Adapt.[23]	2	59.67	56.68	81.20	93.88	50.85	97.05	66.24	89.62	96.13	47.45	73.88	2118
DAN [25]	2.17	57.74	64.12	80.07	91.30	56.54	98.46	86.05	89.67	96.77	49.38	77.01	2851
Piggyback [19]	1.28	57.69	65.29	79.87	96.99	57.45	97.27	79.09	87.63	97.24	47.48	76.60	2838
Parallel SVD [24]	1.5	60.32	66.04	81.86	94.23	57.82	99.24	85.74	89.25	96.62	52.50	78.36	3398
MTAN (Ours)	1.74	63.90	61.81	81.59	91.63	56.44	98.80	81.04	89.83	96.88	50.63	77.25	2941

MTAN은 복잡한 정규화 전략 없이도 대부분의 baseline들을 능가한다.

📌 Conclusion

- Multi-Task Attention Network(MTAN)라는 새로운 다중 태스크 학습 방법
global feature pool과 각 태스크에 대한 task-specific attention 모듈로 구성되어 있고,
이를 통해 task-shared와 task-specific feature 모두를 end-to-end 방식으로 자동으로
학습할 수 있다.
- NYUv2 및 CityScapes 데이터셋에서의 여러 dense-prediction 작업과 Visual Decathlon
Challenge에서의 여러 이미지 분류 작업에 대한 실험 결과, MTAN 방법은 다른 방법들을
능가하거나 경쟁력있는 성능을 보였으며, 손실 함수에서 사용된 특정 task 가중치 체계에
대해서도 안정성을 보였다.
- MTAN의 attention mask를 통한 가중치 공유 능력 덕분에, MTAN은 파라미터에
효율적이면서 좋은 성능을 보인다.

Diddu



이전 포스트

[논문 리뷰] What Changes Can Large-scale Language Models Bring? Intensiv...

0개의 댓글

댓글을 작성하세요

댓글 작성

관심 있을 만한 포스트

