Born-Again Multi-Task Networks for Natural Language

Hyungrak Kim

Content

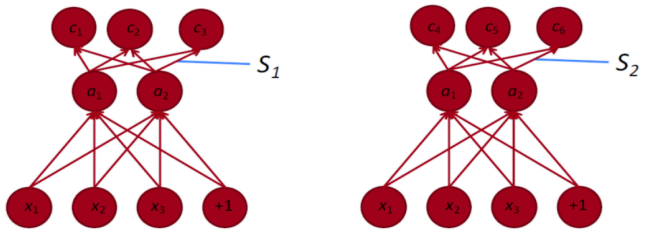
1. Introduction

- Knowledge Distillation 을 이용해서 Single Task model ->teach -> Multi Task Model
- Multi-task model BERT Fine Tuning해서 standard single task와 multi-task 보다 좋은 결과 가져옴
- Contribution :
  - 새로운 Teacher Annealing 제안

Multi-Task Learning

Multi-Task Learning

Multi-Task Learing이란 여러 학습 과제를 동시에 해결하는 기계학습의 한 종류입니다. 예컨대 같은 학습말뭉치로 **개체명 인식(Named Entity Recognition)**과 **품사분류(Part-Of-Speech Tagging)**를 동시에 수행하는 뉴럴네트워크를 만들 수 있습니다. 아래 그림을 먼저 볼까요?



위 그림의 두 네트워크는 마지막에 붙어있는 **소프트맥스 계층(Softmax layer)**만 제외하면 완전히 동일합니다. 다만 S_1의 소프트맥스 확률값은 NER, S_2는 포스태깅 과제를 수행하면서 나오는 스코어라는 점에 유의할 필요가 있습니다. 위와 같은 Multi-Task Learining 네트워크에서는 아래 수식처럼 역전파시 S_1의 그래디언트와 S_2의 그래디언트가 동일한 네트워크에 함께 전달되면서 학습이 이뤄지게 됩니다.

$$\delta^{total} = \delta^{NER} + \delta^{POS}$$

Knowledge Distillation

Softmax Output = Knowledge = Soft Label

ex) ensemble

dog

cow	dog	cat	car	
0	1	0	0	original hard targets
10^{-6}	.9	.1	10^{-9}	output of geometric ensemble
.05	.3	.2	.005	softened output of ensemble

Comparison with the 'hard label' and the 'soft label'

1. Introduction

- Teacher Annealing

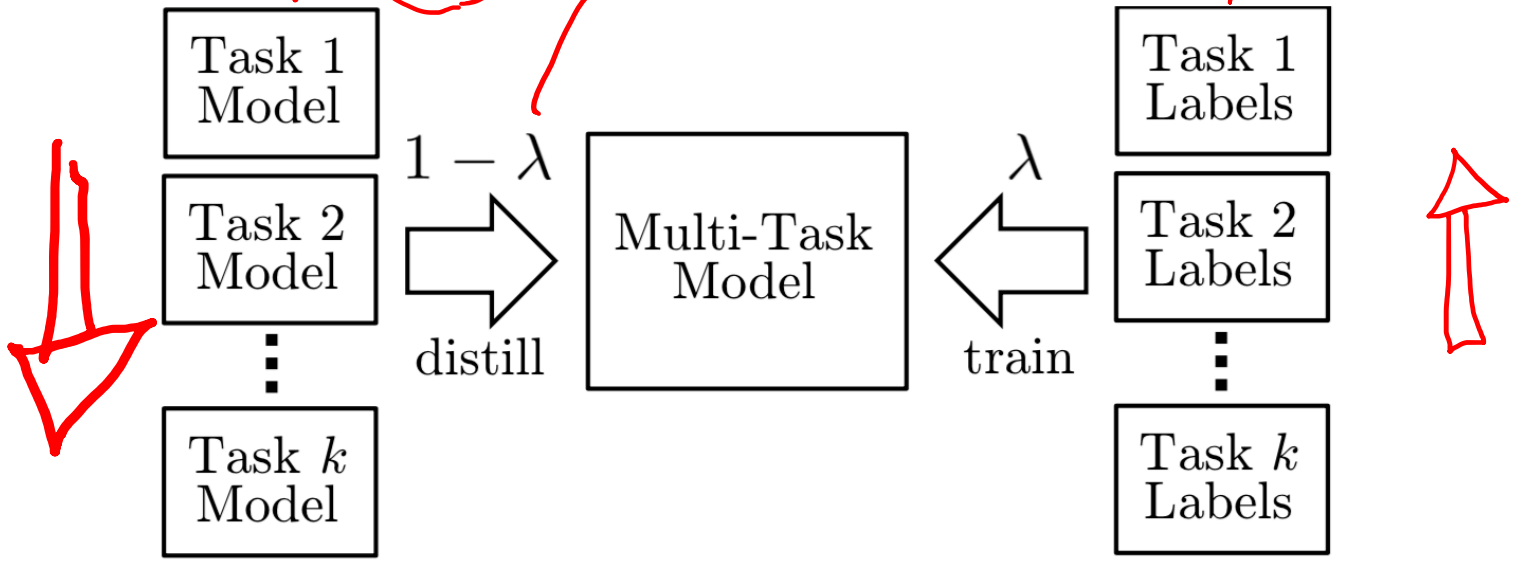


Figure 1: An overview of our method. λ is increased linearly from 0 to 1 over the course of training.

2. Related Work

- Big Ensemble모델에서 Knowledge 를 distillation 해서 Small Network에 적용시킬 수 없을까?(Resource가 너무 많이 들어감)
- Reinforcement Learning 사용 → meta learning
- Machine Translate 사용
- 일반모델 보다 Knowledge를 Distillation 해서 학습된 Model이 하는게 더 좋음
  - 모델 성능향상 기대

NLP → curriculum learning

3. Method

- 3.1 Multi-Task Setup
  - Basic BERT 사용
  - Classification Task Case
    - Softmax Function 사용
  - Regression Task Case
    - Sigmoid Activation 사용
  - Multi-task model
    - Example of different tasks are shuffled together, even within minibatche
  - Single-task model
    - Training. Single-task training is performed as in Devlin et al. (2019).
      - acob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language under- standing. In *NAACL-HLT*.

3. Method

- 3.2 Knowledge Distillation
  - Born-Again Network(teacher, student same Network Architecture)
    - Classification Task : Cross Entropy Function
    - Regression Task: L2 Distance Function

→ loss function

$$\mathcal{L}(\theta) = \sum_{x_\tau^i, y_\tau^i \in \mathcal{D}_\tau} \ell(y_\tau^i, f_\tau(x_\tau^i, \theta))$$

$$\mathcal{L}(\theta) = \sum_{x_\tau^i, y_\tau^i \in \mathcal{D}_\tau} \ell(f_\tau(x_\tau^i, \theta'), f_\tau(x_\tau^i, \theta))$$

3. Method

- 3.2 Knowledge Distillation
  - single-task models to teach a multi-task model with parameters θ:

$$\mathcal{L}(\theta) = \sum_{\tau \in \mathcal{T}} \sum_{x_\tau^i, y_\tau^i \in \mathcal{D}_\tau} \ell(f_\tau(x_\tau^i, \theta_\tau), f_\tau(x_\tau^i, \theta))$$

: teacher
: single task

: student
: multi-task

  - Teacher Annealing
    - Mixes the **Teacher Prediction** with **the Gold Label**
    - λ is linearly increased from 0 to 1 throughout training

$$\ell(\lambda y_\tau^i + (1-\lambda) f_\tau(x_\tau^i, \theta_\tau), f_\tau(x_\tau^i, \theta))$$

4. Experiment

- consists of 9 natural language understanding tasks on English data.
  - REE , MNLI, QNLI, MRPC, QQP, STS, SST-2, CoLA, WNLI
- Dataset
  - 단순 Shuffling (X)
  - Bowman et al. (2018)
    - Alex Wang, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. GLUE: A multi-task benchmark and analysis plat- form for natural language understanding. In *ICLR*.
  - task τ is proportional to $|D_\tau|^{0.75}$.
  - 아주 큰 데이터 셋의 분포가 많아지는 것을 방지

4. Experiment

- Layerwise-learning-rate trick
  - Input과 가까운 Layer수록 General feature 이므로 Learning Rate가 더 작도록)
- For single-task models, we use the same hyperparameters as in the original BERT experiments except we pick a layerwise- learning-rate decay α of 1.0 or 0.9 on the dev set for each task.

- For multi-task models, we train the model for longer (6 epochs instead of 3) and with a larger batch size (128 instead of 32), using α = 0.9 and a learning rate of 1e-4. All models use the BERT-Large pre-trained weights.

5. Result

Model	Avg.	CoLA[a] $\lvert\mathcal{D}\rvert$ = 8.5k	SST-2[b] 67k	MRPC[c] 3.7k	STS-B[d] 5.8k	QQP[e] 364k	MNLI[f] 393k	QNLI[g] 108k	RTE[h] 2.5k
Single	84.0	60.6	93.2	88.0	90.0	91.3	86.6	92.3	70.4
Multi	85.5	60.3	93.3	88.0	89.8	91.4	86.5	92.2	82.1
Single→Single	84.3	**61.7****	93.2	**88.7***	90.0	91.4	**86.8****	**92.5*****	70.0
Multi→Multi	85.6	60.9	93.5	88.1	89.8	**91.5***	86.7	92.3	82.0
Single→Multi	**86.0*****	**61.8****	**93.6***	**89.3****	89.7	**91.6***	**87.0*****	**92.5*****	**82.8***

Dataset references: [a]Warstadt et al. (2018) [b]Socher et al. (2013) [c]Dolan and Brockett (2005) [d]Cer et al. (2017) [e]Iyer et al. (2017) [f]Williams et al. (2018) [g]constructed from SQuAD (Rajpurkar et al., 2016) [h]Giampiccolo et al. (2007)

Table 1: Comparison of methods on the GLUE dev set. *, **, and *** indicate statistically significant ($p < .05$, $p < .01$, and $p < .001$) improvements over both Single and Multi according to bootstrap hypothesis tests.[3]

5. Result

- GLUE Reader Board

Model	GLUE score
BERT-Base (Devlin et al., 2019)	78.5
BERT-Large (Devlin et al., 2019)	80.5
BERT on STILTs (Phang et al., 2018)	82.0
MT-DNN (Liu et al., 2019b)	82.2
Span-Extractive BERT on STILTs (Keskar et al., 2019)	82.3
Snorkel MeTaL ensemble (Hancock et al., 2019)	83.2
MT-DNN$_{KD}$* (Liu et al., 2019a)	83.7
BERT-Large + BAM **(ours)**	82.3

Table 2: Comparison of test set results. *MT-DNN$_{KD}$ is distilled from a diverse ensemble of models.

	Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
	1	Facebook AI	RoBERTa		88.5	67.8	96.7	92.3/89.8	92.2/91.9	74.3/90.2	90.8	90.2	98.9	88.2	89.0	48.7
	2	XLNet Team	XLNet-Large (ensemble)		88.4	67.8	96.8	93.0/90.7	91.6/91.1	74.2/90.3	90.2	89.8	98.6	86.3	90.4	47.5
+	3	Microsoft D365 AI & MSR AI	MT-DNN-ensemble		87.6	68.4	96.5	92.7/90.3	91.1/90.7	73.7/89.9	87.9	87.4	96.0	86.3	89.0	42.8
	4	GLUE Human Baselines	GLUE Human Baselines		87.1	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0	92.8	91.2	93.6	95.9	
+	5	王玮	ALICE large ensemble (Alibaba DAMO N		86.3	68.6	95.2	92.6/90.2	91.1/90.6	74.4/90.7	88.2	87.9	95.7	83.5	80.8	43.9
	6	Stanford Hazy Research	Snorkel MeTaL		83.2	63.8	96.2	91.5/88.5	90.1/89.7	73.1/89.9	87.6	87.2	93.9	80.9	65.1	39.9
	7	XLM Systems	XLM (English only)		83.1	62.9	95.6	90.7/87.1	88.8/88.2	73.2/89.8	89.1	88.5	94.0	76.0	71.9	44.7
	8	张倬胜	SemBERT		82.9	62.3	94.6	91.2/88.3	87.8/86.7	72.8/89.8	87.6	86.3	94.6	84.5	65.1	42.4
	9	Kevin Clark	BERT + BAM		82.3	61.5	95.2	91.3/88.3	88.6/87.9	72.5/89.7	86.6	85.8	93.1	80.4	65.1	40.7
	10	Nitish Shirish Keskar	Span-Extractive BERT on STILTs		82.3	63.2	94.5	90.6/87.6	89.4/89.2	72.2/89.4	86.5	85.8	92.5	79.8	65.1	28.3
	11	Jason Phang	BERT on STILTs		82.0	62.1	94.3	90.2/86.6	88.7/88.3	71.9/89.4	86.4	85.6	92.7	80.1	65.1	28.3
+	12	Jacob Devlin	BERT: 24-layers, 16-heads, 1024-hidden		80.5	60.5	94.9	89.3/85.4	87.6/86.5	72.1/89.3	86.7	85.9	92.7	70.1	65.1	39.6
	13	Neil Houlsby	BERT + Single-task Adapters		80.2	59.2	94.3	88.7/84.3	87.3/86.1	71.5/89.4	85.4	85.0	92.4	71.6	65.1	9.2
	14	Zhuohan Li	Macaron Net-base		79.7	57.6	94.0	88.4/84.4	87.5/86.3	70.8/89.0	85.4	84.5	91.6	70.5	65.1	38.7
	15	Linyuan Gong	StackingBERT-Base		78.4	56.2	93.9	88.2/83.9	84.2/82.5	70.4/88.7	84.4	84.2	90.1	67.0	65.1	36.6

5. Result

- Single-Task Fine-Tuning

Model	Avg. Score
Multi	85.5
+Single-Task Fine-Tuning	+0.3
Single→Multi	86.0
+Single-Task Fine-Tuning	+0.1

Table 3: Combining multi-task training with single-task fine-tuning. Improvements are statistically significant ($p < .01$) according to Mann-Whitney U tests.[3]

5. Result

- Ablation Study

Model	Avg. Score
Single→Multi	86.0
No layer-wise LRs	−0.3
No task sampling	−0.4
No teacher annealing: $\lambda = 0$	−0.5
No teacher annealing: $\lambda = 0.5$	−0.3

Table 4: Ablation Study. Differences from Single→Multi are statistically significant ($p < .001$) according to Mann-Whitney U tests.[3]

6. Conclusion

- Multi-Task model 정확도 향상이 어렵다
- 비슷한 Task에 대한 정확도는 높음
- 계속적으로 knowledge Distillation연구 필요
  - DataSet이 작은 경우(Data에 Label이 많이 없다)에 활용도 높음
- Active Learning, Mento-net과 연관
  - Labeling된 데이터가 많이 없고, 작은 데이터 셋에서 어떻게하면 좀더 효과적으로 학습 할 수 있을지 고민
- RL Meta Learning과 잘 융합

Q&A

Thank you