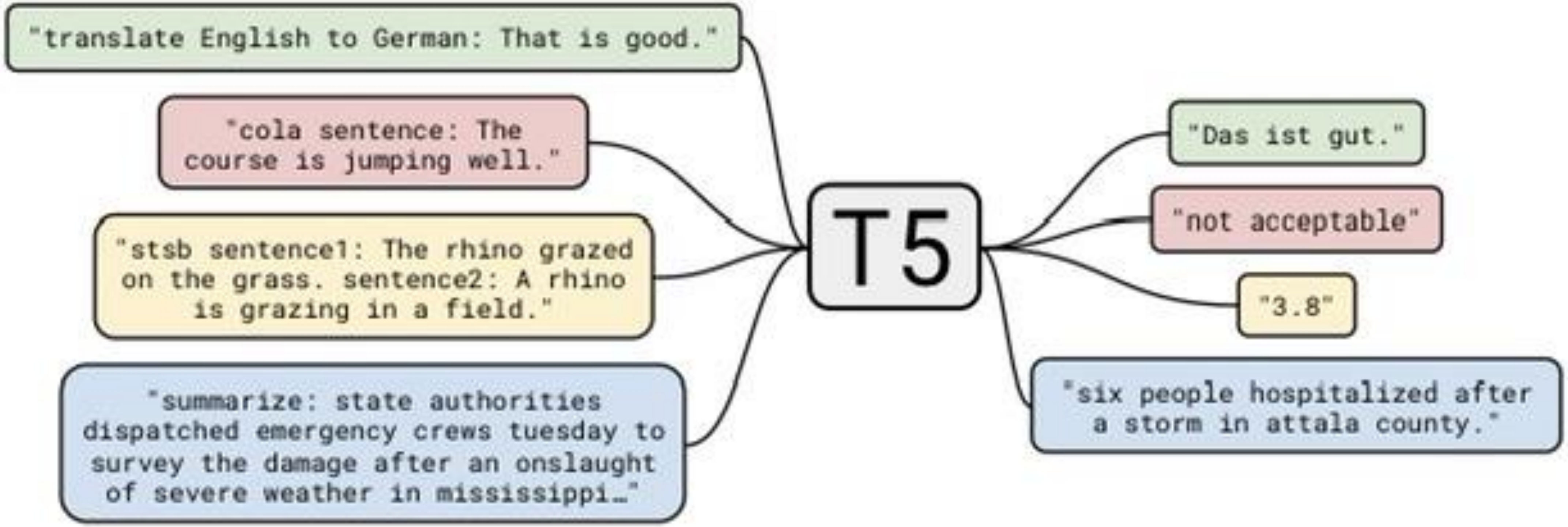
# Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

주세준

2021.07.21




# WHAT IS T5 (Text-to-Text Transfer Transformer)

# baseline

- Standard Transformer
- BERT base size
- Pretraining에 text-to-text denoising objective
- C4 +0.1독+0.1불+0.1로마
- Unsupervised objective ( missing or corrupted token )

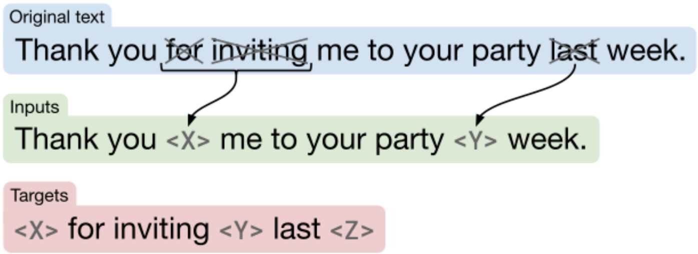
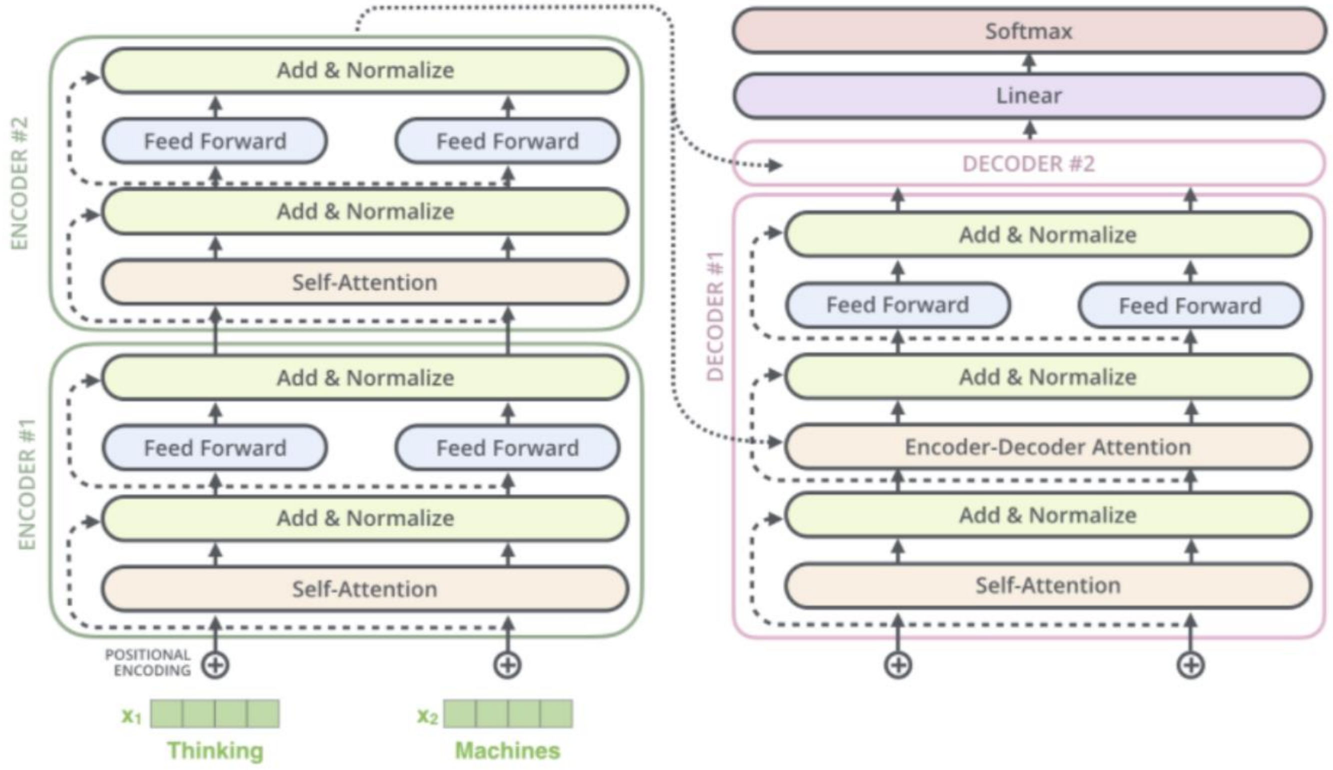



**Figure 2:** Schematic of the objective we use in our baseline model. In this example, we process the sentence "Thank you for inviting me to your party last week." The words "for", "inviting" and "last" (marked with an ×) are randomly chosen for corruption. Each consecutive span of corrupted tokens is replaced by a sentinel token (shown as <X> and <Y>) that is unique over the example. Since "for" and "inviting" occur consecutively, they are replaced by a single sentinel <X>. The output sequence then consists of the dropped-out spans, delimited by the sentinel tokens used to replace them in the input plus a final sentinel token <Z>.

# Dataset for Transfer Learning

- C4 dataset (**C**olossal **C**lean **C**rawled **C**orpus)
- 라벨이 없어야 한다.
- Wikipedia 텍스트의 품질은 우수하지만 스타일이 균일하고 목적에 따라 작다
- Common Crawl 다양하지만 품질이 낮다
- 여러 필터를 이용하여 데이터 품질 향상

=> 더 큰 데이터 셋을 통해 overfitting 방지



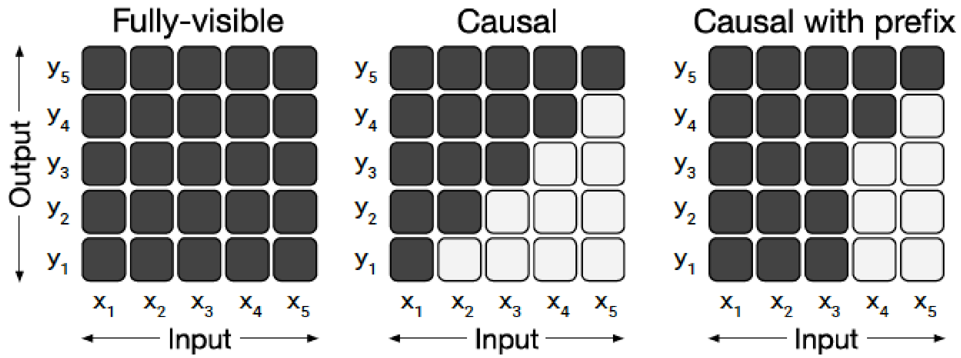

Model structure. From: Jay Alammar's blog

# T5 ARCHITECTURE

BASIC TRANSFORMER ENCODER DECODER

# T5 architecture

- Fully-visible : 출력단어(Query)가 모든 입력단어(Key)에 attention 할 수 있음Self-attention이라면 Transformer encoder 쪽
- Causal : 출력단어(Query)가 자신의 현재 포함 이전 타임 스텝의 입력단어(Key)에 attention 가능Self-attention이라면 Transformer decoder 쪽
- Causal with prefix : 출력단어(Query)가 자신의 현재 포함 이전 타임 스텝의 입력단어(Key)와 일정 길이의 prefix단어(key)에 attention 가능



# T5 architecture

- Text to Text 모델이라서 input에 진행하고 싶은 task에 대한 prefix를 붙여준다.
- encoder-decoder

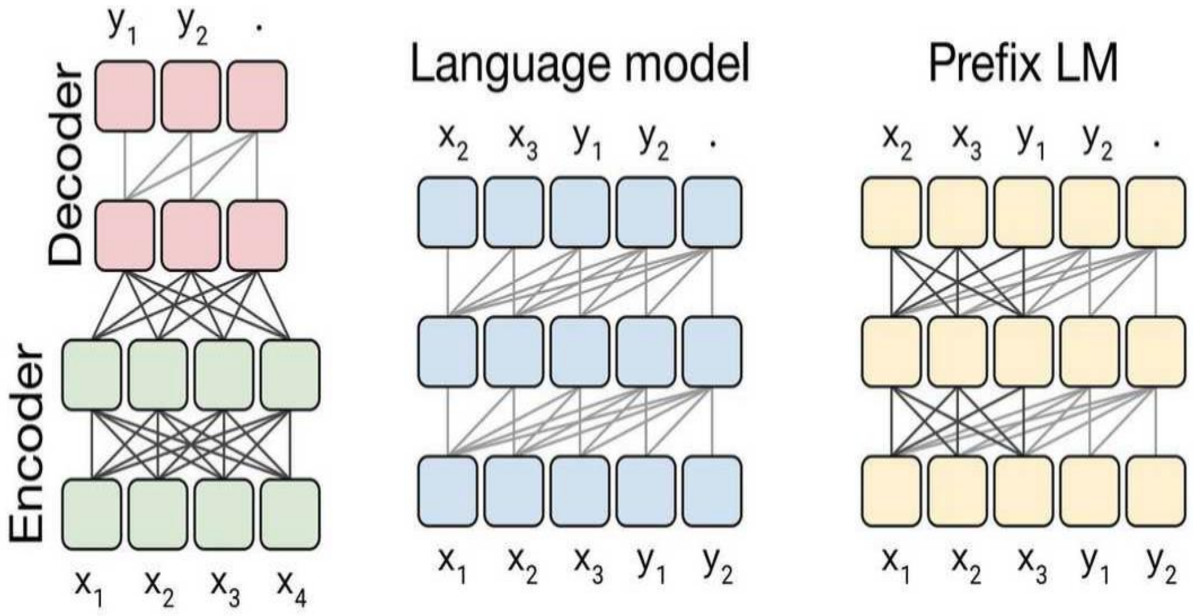
⇒Fully visisble, Casual

- LM

⇒디코더만 사용 casual

Prefix LM

⇒입력 text에선 AE, 출력 text에선 AR ( Fully visible)



# T5 architecture

- En-De 가 최고
- Layer를 줄이는 것은 큰 영향

| Architecture | Objective | Params | Cost | GLUE | CNNDM | SQuAD | SGLUE | EnDe | EnFr | EnRo |
|---|---|---|---|---|---|---|---|---|---|---|
| ★ Encoder-decoder | Denoising | $2P$ | $M$ | **83.28** | **19.24** | **80.88** | **71.36** | **26.98** | **39.82** | **27.65** |
| Enc-dec, shared | Denoising | $P$ | $M$ | 82.81 | 18.78 | **80.63** | **70.73** | 26.72 | 39.03 | **27.46** |
| Enc-dec, 6 layers | Denoising | $P$ | $M/2$ | 80.88 | 18.97 | 77.59 | 68.42 | 26.38 | 38.40 | 26.95 |
| Language model | Denoising | $P$ | $M$ | 74.70 | 17.93 | 61.14 | 55.02 | 25.09 | 35.28 | 25.86 |
| Prefix LM | Denoising | $P$ | $M$ | 81.82 | 18.61 | 78.94 | 68.11 | 26.43 | 37.98 | 27.39 |
| Encoder-decoder | LM | $2P$ | $M$ | 79.56 | 18.59 | 76.02 | 64.29 | 26.27 | 39.17 | 26.86 |
| Enc-dec, shared | LM | $P$ | $M$ | 79.60 | 18.13 | 76.35 | 63.50 | 26.62 | 39.17 | 27.05 |
| Enc-dec, 6 layers | LM | $P$ | $M/2$ | 78.67 | 18.26 | 75.32 | 64.06 | 26.13 | 38.42 | 26.89 |
| Language model | LM | $P$ | $M$ | 73.78 | 17.54 | 53.81 | 56.51 | 25.23 | 34.31 | 25.38 |
| Prefix LM | LM | $P$ | $M$ | 79.68 | 17.84 | 76.87 | 64.86 | 26.28 | 37.51 | 26.76 |

**Table 2:** Performance of the different architectural variants described in Section 3.2.2. We use $P$ to refer to the number of parameters in a 12-layer base Transformer layer stack and $M$ to refer to the FLOPs required to process a sequence using the encoder-decoder model. We evaluate each architectural variant using a denoising objective (described in Section 3.1.4) and an autoregressive objective (as is commonly used to train language models).

# Unsupervised objectives

| Objective | Inputs | Targets |
|---|---|---|
| Prefix language modeling | Thank you for inviting | me to your party last week . |
| BERT-style | Thank you <M> <M> me to your party apple week . | *(original text)* |
| Deshuffling | party me for your to . last fun you inviting week Thank | *(original text)* |
| I.i.d. noise, mask tokens | Thank you <M> <M> me to your party <M> week . | *(original text)* |
| I.i.d. noise, replace spans | Thank you <X> me to your party <Y> week . | <X> for inviting <Y> last <Z> |
| I.i.d. noise, drop tokens | Thank you me to your party week . | for inviting last |
| Random spans | Thank you <X> to <Y> week . | <X> for inviting me <Y> your party last <Z> |

**Table 3:** Examples of inputs and targets produced by some of the unsupervised objectives we consider applied to the input text "Thank you for inviting me to your party last week ." Note that all of our objectives process *tokenized* text. For this particular sentence, all words were mapped to a single token by our vocabulary. We write *(original text)* as a target to denote that the model is tasked with reconstructing the entire input text. <M> denotes a shared mask token and <X>, <Y>, and <Z> denote sentinel tokens that are assigned unique token IDs. The BERT-style objective (second row) includes a corruption where some tokens are replaced by a random token ID; we show this via the greyed-out word apple.

# Unsupervised objectives

- BERT-style을 제일 잘함
- Deshuffling 성능 저하
- Greedy 하게 탐색

| Objective | GLUE | CNNDM | SQuAD | SGLUE | EnDe | EnFr | EnRo |
|---|---|---|---|---|---|---|---|
| Prefix language modeling | 80.69 | 18.94 | 77.99 | 65.27 | **26.86** | 39.73 | **27.49** |
| BERT-style [Devlin et al., 2018] | **82.96** | **19.17** | **80.65** | **69.85** | **26.78** | **40.03** | **27.41** |
| Deshuffling | 73.17 | 18.59 | 67.61 | 58.47 | 26.11 | 39.30 | 25.62 |

**Table 4:** Performance of the three disparate pre-training objectives described in Section 3.3.1.

# Unsupervised objectives

- BERT-style방법 중 최선의 방법을 찾자
- Replace corrupt span 이 제일 좋았다
- 그 중 15%일때 가장 성능이 좋다

| Objective | GLUE | CNNDM | SQuAD | SGLUE | EnDe | EnFr | EnRo |
|---|---|---|---|---|---|---|---|
| BERT-style [Devlin et al., 2018] | 82.96 | 19.17 | **80.65** | 69.85 | 26.78 | **40.03** | 27.41 |
| MASS-style [Song et al., 2019] | 82.32 | 19.16 | 80.10 | 69.28 | 26.79 | **39.89** | 27.55 |
| ★ Replace corrupted spans | 83.28 | **19.24** | **80.88** | **71.36** | **26.98** | 39.82 | **27.65** |
| Drop corrupted tokens | **84.44** | **19.31** | **80.52** | 68.67 | **27.07** | 39.76 | **27.82** |

**Table 5:** Comparison of variants of the BERT-style pre-training objective. In the first two variants, the model is trained to reconstruct the original uncorrupted text segment. In the latter two, the model only predicts the sequence of corrupted tokens.

| Corruption rate | GLUE | CNNDM | SQuAD | SGLUE | EnDe | EnFr | EnRo |
|---|---|---|---|---|---|---|---|
| 10% | **82.82** | 19.00 | **80.38** | 69.55 | **26.87** | 39.28 | **27.44** |
| ★ 15% | **83.28** | 19.24 | **80.88** | **71.36** | **26.98** | **39.82** | **27.65** |
| 25% | **83.00** | **19.54** | **80.96** | 70.48 | **27.04** | **39.83** | **27.47** |
| 50% | 81.27 | 19.32 | 79.80 | 70.33 | **27.01** | **39.90** | **27.49** |

**Table 6:** Performance of the i.i.d. corruption objective with different corruption rates.

| Span length | GLUE | CNNDM | SQuAD | SGLUE | EnDe | EnFr | EnRo |
|---|---|---|---|---|---|---|---|
| ★ Baseline (i.i.d.) | **83.28** | 19.24 | 80.88 | 71.36 | **26.98** | **39.82** | **27.65** |
| 2 | **83.54** | 19.39 | **82.09** | **72.20** | **26.76** | **39.99** | **27.63** |
| 3 | **83.49** | **19.62** | **81.84** | **72.53** | **26.86** | 39.65 | **27.62** |
| 5 | **83.40** | 19.24 | **82.05** | **72.23** | **26.88** | 39.40 | **27.53** |
| 10 | 82.85 | 19.33 | **81.84** | 70.44 | **26.79** | 39.49 | **27.69** |

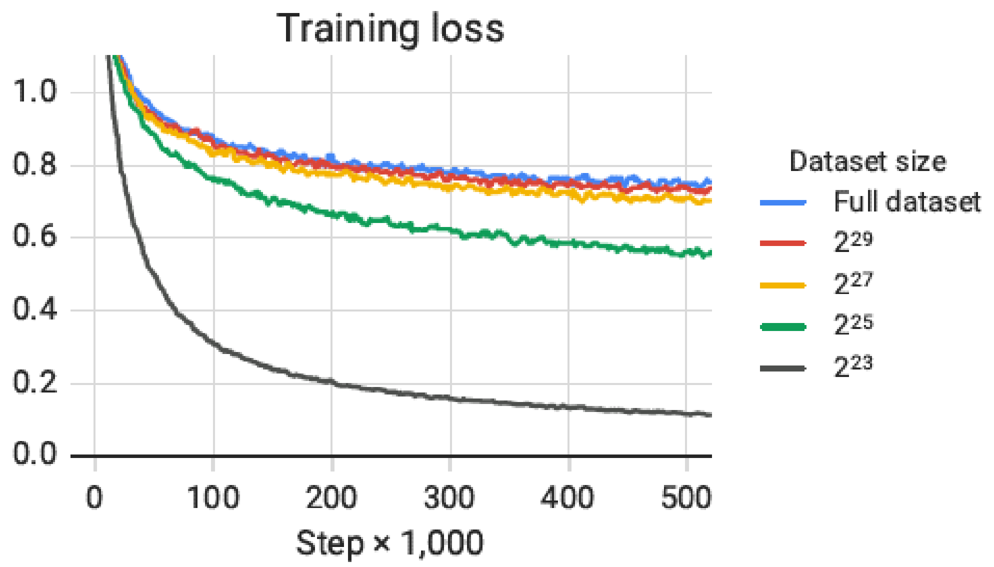
**Table 7:** Performance of the span-corruption objective (inspired by Joshi et al. [2019]) for different average span lengths. In all cases, we corrupt 15% of the original text sequence.

# Pre-training data size

같은 데이터로 여러번 학습하는것보다
많은 데이터를 적은 횟수 학습하는게 낫다.

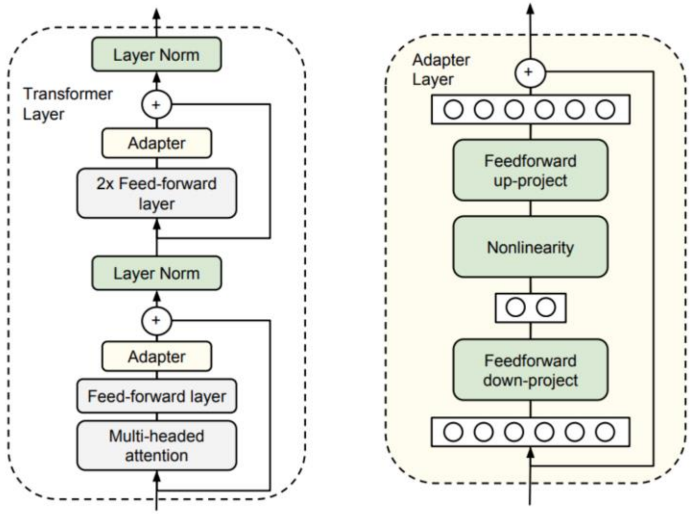
데이터 다다익선

| Number of tokens | Repeats | GLUE | CNNDM | SQuAD | SGLUE | EnDe | EnFr | EnRo |
|---|---|---|---|---|---|---|---|---|
| ★ Full data set | 0 | **83.28** | **19.24** | **80.88** | **71.36** | **26.98** | **39.82** | **27.65** |
| $2^{29}$ | 64 | **82.87** | **19.19** | **80.97** | **72.03** | **26.83** | **39.74** | **27.63** |
| $2^{27}$ | 256 | 82.62 | **19.20** | 79.78 | 69.97 | **27.02** | **39.71** | 27.33 |
| $2^{25}$ | 1,024 | 79.55 | 18.57 | 76.27 | 64.76 | 26.38 | 39.56 | 26.80 |
| $2^{23}$ | 4,096 | 76.34 | 18.33 | 70.92 | 59.29 | 26.37 | 38.84 | 25.81 |



# TRAINING STRATEGY

- 모든 파라미터를 fine-tuning하는 것은 별로
- -> Adaptive layer를 두어 그 부분만 튜닝
- -> Gradual unfreezing 마지막 레이어부터 천천히




- 리소스가 적게 필요한 테스크는 d 값 작아도 됨

| Fine-tuning method | GLUE | CNNDM | SQuAD | SGLUE | EnDe | EnFr | EnRo |
|---|---|---|---|---|---|---|---|
| ★ All parameters | **83.28** | **19.24** | **80.88** | **71.36** | **26.98** | **39.82** | **27.65** |
| Adapter layers, $d = 32$ | 80.52 | 15.08 | 79.32 | 60.40 | 13.84 | 17.88 | 15.54 |
| Adapter layers, $d = 128$ | 81.51 | 16.62 | 79.47 | 63.03 | 19.83 | 27.50 | 22.63 |
| Adapter layers, $d = 512$ | 81.54 | 17.78 | 79.18 | 64.30 | 23.45 | 33.98 | 25.81 |
| Adapter layers, $d = 2048$ | 81.51 | 16.62 | 79.47 | 63.03 | 19.83 | 27.50 | 22.63 |
| Gradual unfreezing | 82.50 | 18.95 | 79.17 | **70.79** | 26.71 | 39.02 | 26.93 |

# Multi-task learning

- Text-to-Text 에서는 단순히 데이터를 혼합하는 개념

-> 어떤 테스크의 데이터를 가져올지 **비율( 적정 비율을 찾자 )**

Example-Proportional mixing( 데이터셋 사이즈 비율 )

Temperature-scaled mixing(mBERT에서 사용한 방법, 덩어리로 나눔)

Equal mixing

-> 뭘해도 baseline보다 성능이 안나옴

| Mixing strategy | GLUE | CNNDM | SQuAD | SGLUE | EnDe | EnFr | EnRo |
|---|---|---|---|---|---|---|---|
| ★ Baseline (pre-train/fine-tune) | **83.28** | **19.24** | **80.88** | **71.36** | **26.98** | **39.82** | **27.65** |
| Equal | 76.13 | 19.02 | 76.51 | 63.37 | 23.89 | 34.31 | 26.78 |
| Examples-proportional, $K = 2^{16}$ | 80.45 | 19.04 | 77.25 | 69.95 | 24.35 | 34.99 | 27.10 |
| Examples-proportional, $K = 2^{17}$ | 81.56 | 19.12 | 77.00 | 67.91 | 24.36 | 35.00 | 27.25 |
| Examples-proportional, $K = 2^{18}$ | 81.67 | 19.07 | 78.17 | 67.94 | 24.57 | 35.19 | 27.39 |
| Examples-proportional, $K = 2^{19}$ | 81.42 | **19.24** | 79.78 | 67.30 | 25.21 | 36.30 | **27.76** |
| Examples-proportional, $K = 2^{20}$ | 80.80 | **19.24** | **80.36** | 67.38 | 25.66 | 36.93 | **27.68** |
| Examples-proportional, $K = 2^{21}$ | 79.83 | 18.79 | 79.50 | 65.10 | 25.82 | 37.22 | 27.13 |
| Temperature-scaled, $T = 2$ | 81.90 | **19.28** | 79.42 | 69.92 | 25.42 | 36.72 | 27.20 |
| Temperature-scaled, $T = 4$ | 80.56 | **19.22** | 77.99 | 69.54 | 25.04 | 35.82 | 27.45 |
| Temperature-scaled, $T = 8$ | 77.21 | 19.10 | 77.14 | 66.07 | 24.55 | 35.35 | 27.17 |

# Multi-task + fine-tuning

| Training strategy | GLUE | CNNDM | SQuAD | SGLUE | EnDe | EnFr | EnRo |
|---|---|---|---|---|---|---|---|
| ★ Unsupervised pre-training + fine-tuning | **83.28** | **19.24** | **80.88** | **71.36** | **26.98** | 39.82 | 27.65 |
| Multi-task training | 81.42 | **19.24** | 79.78 | 67.30 | 25.21 | 36.30 | 27.76 |
| Multi-task pre-training + fine-tuning | **83.11** | **19.12** | **80.26** | **71.03** | **27.08** | 39.80 | **28.07** |
| Leave-one-out multi-task training | 81.98 | 19.05 | 79.97 | **71.68** | **26.93** | 39.79 | **27.87** |
| Supervised multi-task pre-training | 79.93 | 18.96 | 77.38 | 65.36 | 26.81 | **40.13** | **28.04** |

- Multi Task data ( not multilabel )
- Multi-task pre-training + fine-tuning
  - Pre-Training : Example-proportional mixtrure(K = 2^19)
  - Fine-Tuning : 테스크 별로 독립적으로.
- Leave-one-out Multi-task pre-training
  - Pre-Training시에 하나의 테스크 데이터를 빼고,
  - Fine-Tuning시에 그 테스크에 대해서 튜닝 진행
- Supervised multi-task pre-training
  - Pre-Training : Supervised with K = 2^19
  - Fine-Tuning 독립적으로.

결과적으로 그냥 Unsupervised와 비슷한 성능을 보인다

훈련 전반에 걸쳐 downstream task에 대한 성능을 볼 수 있다

무슨 데이터 이용..?
크기가 작아지나?

# Additional pre-train

- T5 base 가장 훌륭
- 1T 1trillion tokens same as T5 model
- Additional pre training 굳

| Model | GLUE | CNNDM | SQuAD | SGLUE | EnDe | EnFr | EnRo |
|---|---|---|---|---|---|---|---|
| ★ Baseline | 83.28 | 19.24 | 80.88 | 71.36 | 26.98 | 39.82 | 27.65 |
| Baseline-1T | 84.80 | 19.62 | 83.01 | 73.90 | 27.46 | 40.30 | 28.34 |
| T5-Base | **85.97** | **20.90** | **85.44** | **75.64** | **28.37** | **41.37** | **28.98** |

# ALL together

Objective:  denosing objective 성능 좋음 Span length 3 & corrupt 15% 진행 span 2,3,5,10

longer training:

데이터의 반복이 필요없는 C4의 충분한 데이터가 있음

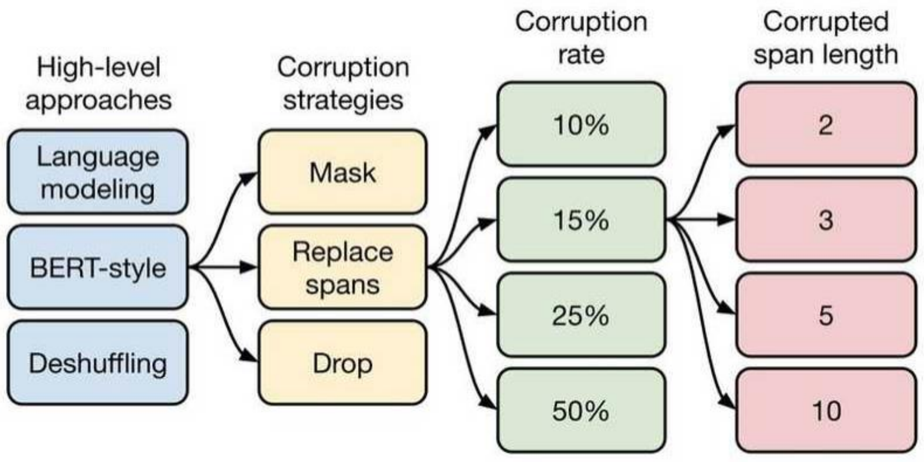
pre-training을 추가하는 것이 도움이 되었고, 배치 사이즈의 증가와 training의 증가가 역시 도움이 되었음. 그래서 C4가 좋고, 그걸 사용함.

Model size: 모델 사이즈의 크기가 성능을 향상시키는 것을 보여주었음

Multi-task pre-training: 지도/비지도 학습 테스크를 함께 Pre-training 하는 것이 비지도 학습만으로 Pre-training 하는 것 보다 좋다는 걸 보였다. 그러나 잘못해서 unlabeled data를 잘못 선택하면 fine-tuning 과정에서 문제가 생길 수 있음.

Example-proportional mixing 방법으로 Muti-task Pre-training을 진행하면 성능 뿐만 아니라, 속도 개선에서 도움이 된다.

이러한 아이디어를 바탕으로 Unsupervied용 데이터의 크기를 제한



| Model | GLUE Average | CoLA Matthew's | SST-2 Accuracy | MRPC F1 | MRPC Accuracy | STS-B Pearson | STS-B Spearman |
|---|---|---|---|---|---|---|---|
| Previous best | 89.4[a] | 69.2[b] | 97.1[a] | **93.6**[b] | **91.5**[b] | 92.7[b] | 92.3[b] |
| T5-Small | 77.4 | 41.0 | 91.8 | 89.7 | 86.6 | 85.6 | 85.0 |
| T5-Base | 82.7 | 51.1 | 95.2 | 90.7 | 87.5 | 89.4 | 88.6 |
| T5-Large | 86.4 | 61.2 | 96.3 | 92.4 | 89.9 | 89.9 | 89.2 |
| T5-3B | 88.5 | 67.1 | 97.4 | 92.5 | 90.0 | 90.6 | 89.8 |
| T5-11B | **90.3** | **71.6** | **97.5** | 92.8 | 90.4 | **93.1** | **92.8** |

| Model | QQP F1 | QQP Accuracy | MNLI-m Accuracy | MNLI-mm Accuracy | QNLI Accuracy | RTE Accuracy | WNLI Accuracy |
|---|---|---|---|---|---|---|---|
| Previous best | 74.8[c] | **90.7**[b] | 91.3[a] | 91.0[a] | **99.2**[a] | 89.2[a] | 91.8[a] |
| T5-Small | 70.0 | 88.0 | 82.4 | 82.3 | 90.3 | 69.9 | 69.2 |
| T5-Base | 72.6 | 89.4 | 87.1 | 86.2 | 93.7 | 80.1 | 78.8 |
| T5-Large | 73.9 | 89.9 | 89.9 | 89.6 | 94.8 | 87.2 | 85.6 |
| T5-3B | 74.4 | 89.7 | 91.4 | 91.2 | 96.3 | 91.1 | 89.7 |
| T5-11B | **75.1** | 90.6 | **92.2** | **91.9** | 96.9 | **92.8** | **94.5** |

| Model | SQuAD EM | SQuAD F1 | SuperGLUE Average | BoolQ Accuracy | CB F1 | CB Accuracy | COPA Accuracy |
|---|---|---|---|---|---|---|---|
| Previous best | 90.1[a] | 95.5[a] | 84.6[d] | 87.1[d] | 90.5[d] | 95.2[d] | 90.6[d] |
| T5-Small | 79.10 | 87.24 | 63.3 | 76.4 | 56.9 | 81.6 | 46.0 |
| T5-Base | 85.44 | 92.08 | 76.2 | 81.4 | 86.2 | 94.0 | 71.2 |
| T5-Large | 86.66 | 93.79 | 82.3 | 85.4 | 91.6 | 94.8 | 83.4 |
| T5-3B | 88.53 | 94.95 | 86.4 | 89.9 | 90.3 | 94.4 | 92.0 |
| T5-11B | **91.26** | **96.22** | **88.9** | **91.2** | **93.9** | **96.8** | **94.8** |

| Model | MultiRC F1a | MultiRC EM | ReCoRD F1 | ReCoRD Accuracy | RTE Accuracy | WiC Accuracy | WSC Accuracy |
|---|---|---|---|---|---|---|---|
| Previous best | 84.4[d] | 52.5[d] | 90.6[d] | 90.0[d] | 88.2[d] | 69.9[d] | 89.0[d] |
| T5-Small | 69.3 | 26.3 | 56.3 | 55.4 | 73.3 | 66.9 | 70.5 |
| T5-Base | 79.7 | 43.1 | 75.0 | 74.2 | 81.5 | 68.3 | 80.8 |
| T5-Large | 83.3 | 50.7 | 86.8 | 85.9 | 87.8 | 69.3 | 86.3 |
| T5-3B | 86.8 | 58.3 | 91.2 | 90.4 | 90.7 | 72.1 | 90.4 |
| T5-11B | **88.1** | **63.3** | **94.1** | **93.4** | **92.5** | **76.9** | **93.8** |

| Model | WMT EnDe BLEU | WMT EnFr BLEU | WMT EnRo BLEU | CNN/DM ROUGE-1 | CNN/DM ROUGE-2 | CNN/DM ROUGE-L |
|---|---|---|---|---|---|---|
| Previous best | **33.8**[e] | **43.8**[e] | **38.5**[f] | 43.47[g] | 20.30[g] | 40.63[g] |
| T5-Small | 26.7 | 36.0 | 26.8 | 41.12 | 19.56 | 38.35 |
| T5-Base | 30.9 | 41.2 | 28.0 | 42.05 | 20.34 | 39.40 |
| T5-Large | 32.0 | 41.5 | 28.1 | 42.50 | 20.68 | 39.75 |
| T5-3B | 31.8 | 42.6 | 28.2 | 42.72 | 21.02 | 39.94 |
| T5-11B | 32.1 | 43.4 | 28.1 | **43.52** | **21.55** | **40.69** |

# PERFORMANCE ON FINETUNING TASKS

T5-11B best performance
모델은 클수록 좋다

# GLUE

- 90.3 SOTA 달성
- 11B model GLUE에 등록된 모델 중 가장 크기가 크다
- 하지만 smaller size의 다른 모델들( ALBERT 3B)는 결국 computational cost 가 더 크다

| Model | GLUE Average | CoLA Matthew's | SST-2 Accuracy | MRPC F1 | MRPC Accuracy | STS-B Pearson | STS-B Spearman |
|---|---|---|---|---|---|---|---|
| Previous best | 89.4[a] | 69.2[b] | 97.1[a] | **93.6**[b] | **91.5**[b] | 92.7[b] | 92.3[b] |
| T5-Small | 77.4 | 41.0 | 91.8 | 89.7 | 86.6 | 85.6 | 85.0 |
| T5-Base | 82.7 | 51.1 | 95.2 | 90.7 | 87.5 | 89.4 | 88.6 |
| T5-Large | 86.4 | 61.2 | 96.3 | 92.4 | 89.9 | 89.9 | 89.2 |
| T5-3B | 88.5 | 67.1 | 97.4 | 92.5 | 90.0 | 90.6 | 89.8 |
| T5-11B | **90.3** | **71.6** | **97.5** | 92.8 | 90.4 | **93.1** | **92.8** |

| Model | QQP F1 | QQP Accuracy | MNLI-m Accuracy | MNLI-mm Accuracy | QNLI Accuracy | RTE Accuracy | WNLI Accuracy |
|---|---|---|---|---|---|---|---|
| Previous best | 74.8[c] | **90.7**[b] | 91.3[a] | 91.0[a] | **99.2**[a] | 89.2[a] | 91.8[a] |
| T5-Small | 70.0 | 88.0 | 82.4 | 82.3 | 90.3 | 69.9 | 69.2 |
| T5-Base | 72.6 | 89.4 | 87.1 | 86.2 | 93.7 | 80.1 | 78.8 |
| T5-Large | 73.9 | 89.9 | 89.9 | 89.6 | 94.8 | 87.2 | 85.6 |
| T5-3B | 74.4 | 89.7 | 91.4 | 91.2 | 96.3 | 91.1 | 89.7 |
| T5-11B | **75.1** | 90.6 | **92.2** | **91.9** | 96.9 | **92.8** | **94.5** |

# SQuAD, SuperGlue

- SOTA 달성 ( ALBERT )
- SuperGlue에선  human performance 의 영역에 진입

| Model | SQuAD EM | SQuAD F1 | SuperGLUE Average | BoolQ Accuracy | CB F1 | CB Accuracy | COPA Accuracy |
|---|---|---|---|---|---|---|---|
| Previous best | 90.1[a] | 95.5[a] | 84.6[d] | 87.1[d] | 90.5[d] | 95.2[d] | 90.6[d] |
| T5-Small | 79.10 | 87.24 | 63.3 | 76.4 | 56.9 | 81.6 | 46.0 |
| T5-Base | 85.44 | 92.08 | 76.2 | 81.4 | 86.2 | 94.0 | 71.2 |
| T5-Large | 86.66 | 93.79 | 82.3 | 85.4 | 91.6 | 94.8 | 83.4 |
| T5-3B | 88.53 | 94.95 | 86.4 | 89.9 | 90.3 | 94.4 | 92.0 |
| T5-11B | **91.26** | **96.22** | **88.9** | **91.2** | **93.9** | **96.8** | **94.8** |

| Model | MultiRC F1a | MultiRC EM | ReCoRD F1 | ReCoRD Accuracy | RTE Accuracy | WiC Accuracy | WSC Accuracy |
|---|---|---|---|---|---|---|---|
| Previous best | 84.4[d] | 52.5[d] | 90.6[d] | 90.0[d] | 88.2[d] | 69.9[d] | 89.0[d] |
| T5-Small | 69.3 | 26.3 | 56.3 | 55.4 | 73.3 | 66.9 | 70.5 |
| T5-Base | 79.7 | 43.1 | 75.0 | 74.2 | 81.5 | 68.3 | 80.8 |
| T5-Large | 83.3 | 50.7 | 86.8 | 85.9 | 87.8 | 69.3 | 86.3 |
| T5-3B | 86.8 | 58.3 | 91.2 | 90.4 | 90.7 | 72.1 | 90.4 |
| T5-11B | **88.1** | **63.3** | **94.1** | **93.4** | **92.5** | **76.9** | **93.8** |

# WMT

| Model | WMT EnDe BLEU | WMT EnFr BLEU | WMT EnRo BLEU | CNN/DM ROUGE-1 | CNN/DM ROUGE-2 | CNN/DM ROUGE-L |
|---|---|---|---|---|---|---|
| Previous best | **33.8**[e] | **43.8**[e] | **38.5**[f] | 43.47[g] | 20.30[g] | 40.63[g] |
| T5-Small | 26.7 | 36.0 | 26.8 | 41.12 | 19.56 | 38.35 |
| T5-Base | 30.9 | 41.2 | 28.0 | 42.05 | 20.34 | 39.40 |
| T5-Large | 32.0 | 41.5 | 28.1 | 42.50 | 20.68 | 39.75 |
| T5-3B | 31.8 | 42.6 | 28.2 | 42.72 | 21.02 | 39.94 |
| T5-11B | 32.1 | 43.4 | 28.1 | **43.52** | **21.55** | **40.69** |

SOTA 달성 하지 못한

C4가 only English 라 발생한 상황 (mC4)

Language agnostic model

CNN/Daily Mail 에서도 SOTA 달성