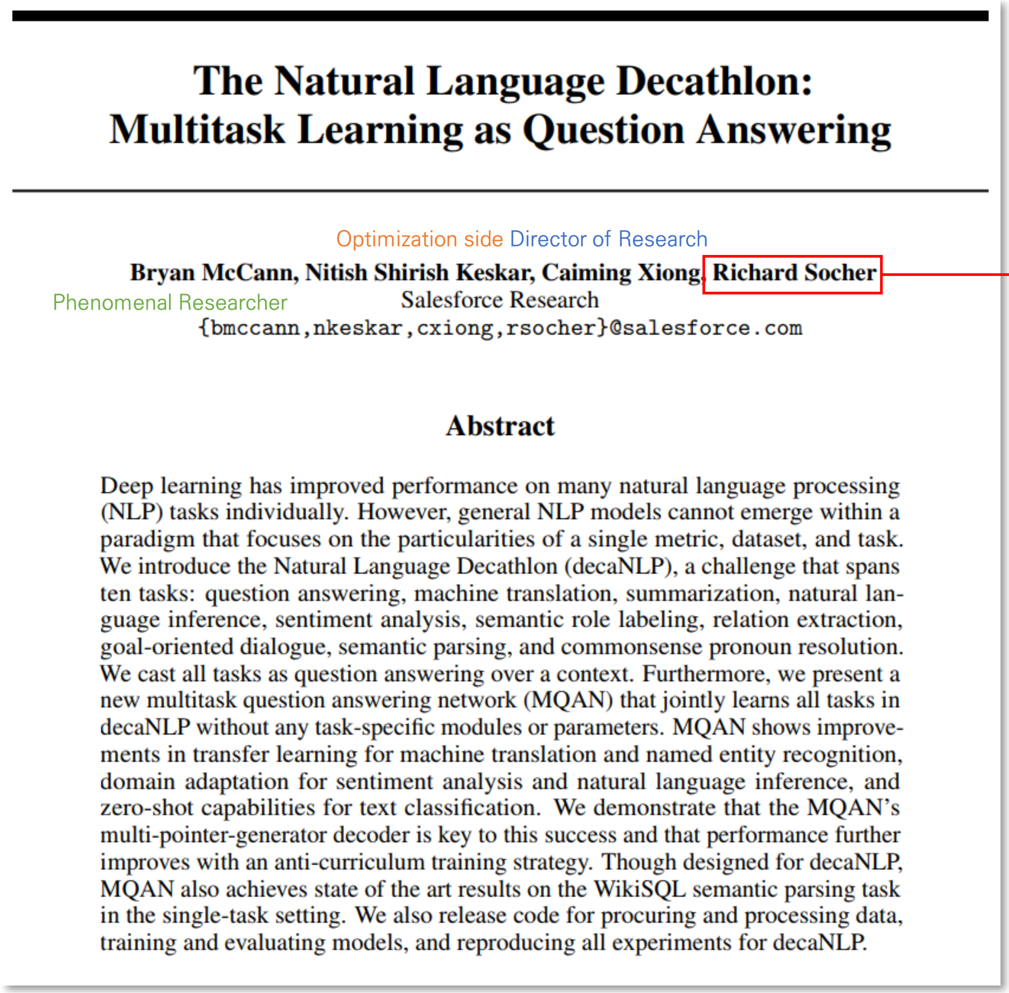
CS224n: NLP with Deep Learning | Winter 2019

Lecture 17. Multitask Learning

- Introduction
- Tasks and Metrics
- Multitask Question Answering Network (MQAN)
- Experiments and Analysis

DecaNLP



The Natural Language Decathlon: Multitask Learning as Question Answering

Optimization side Director of Research

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, Richard Socher

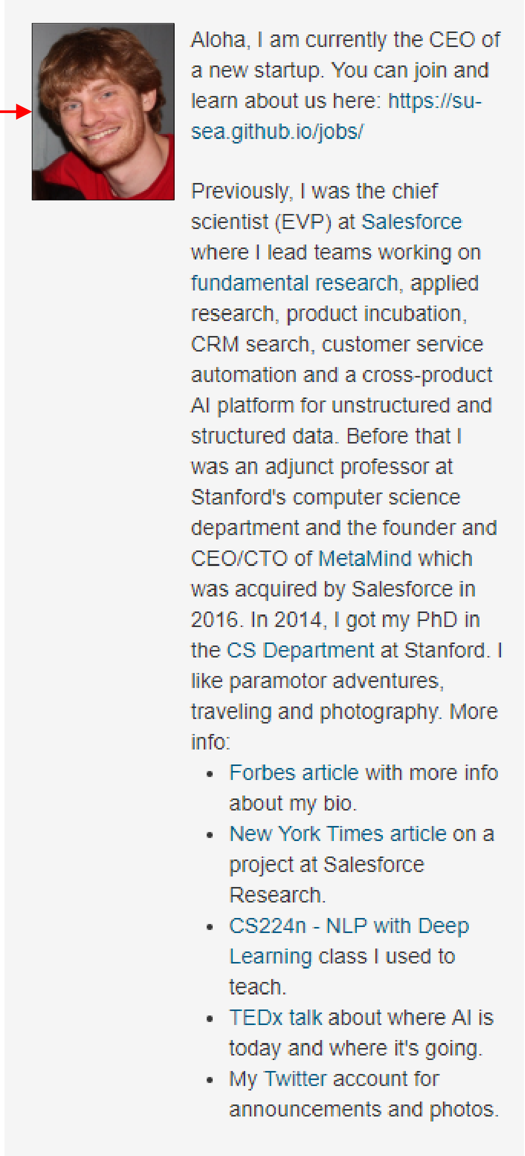
Phenomenal Researcher

Salesforce Research

{bmccann,nkeskar,cxiong,rsocher}@salesforce.com

Abstract

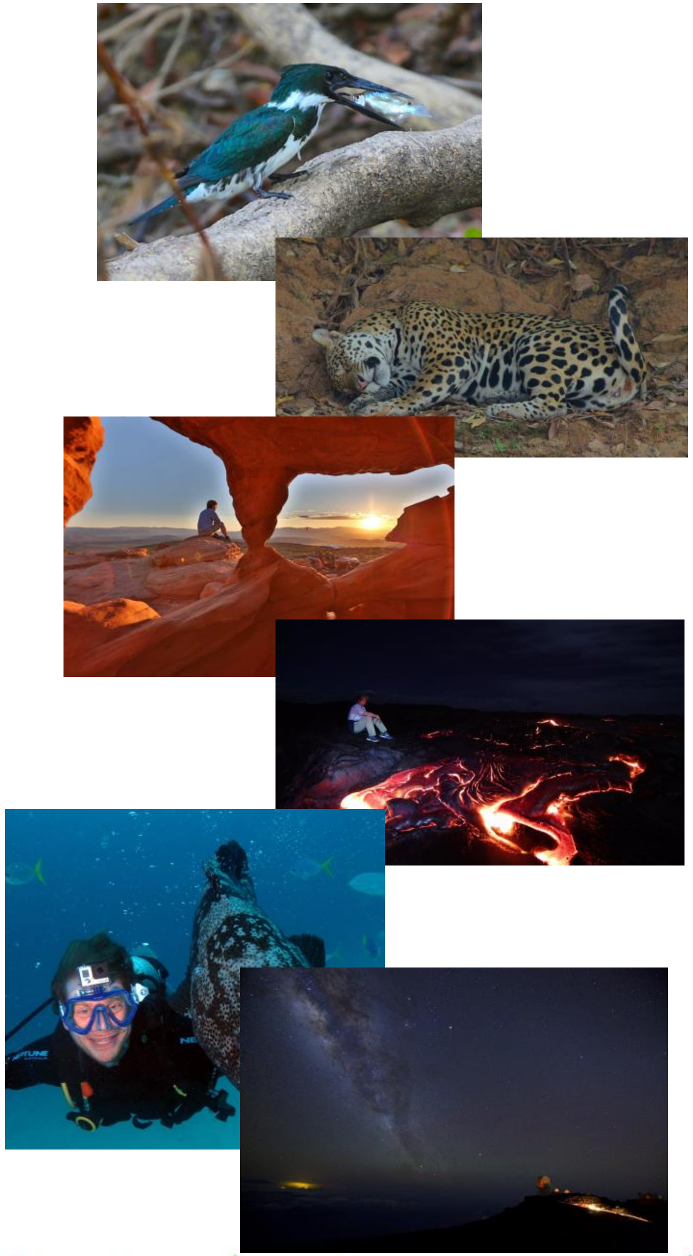
Deep learning has improved performance on many natural language processing (NLP) tasks individually. However, general NLP models cannot emerge within a paradigm that focuses on the particularities of a single metric, dataset, and task. We introduce the Natural Language Decathlon (decaNLP), a challenge that spans ten tasks: question answering, machine translation, summarization, natural language inference, sentiment analysis, semantic role labeling, relation extraction, goal-oriented dialogue, semantic parsing, and commonsense pronoun resolution. We cast all tasks as question answering over a context. Furthermore, we present a new multitask question answering network (MQAN) that jointly learns all tasks in decaNLP without any task-specific modules or parameters. MQAN shows improvements in transfer learning for machine translation and named entity recognition, domain adaptation for sentiment analysis and natural language inference, and zero-shot capabilities for text classification. We demonstrate that the MQAN's multi-pointer-generator decoder is key to this success and that performance further improves with an anti-curriculum training strategy. Though designed for decaNLP, MQAN also achieves state of the art results on the WikiSQL semantic parsing task in the single-task setting. We also release code for procuring and processing data, training and evaluating models, and reproducing all experiments for decaNLP.



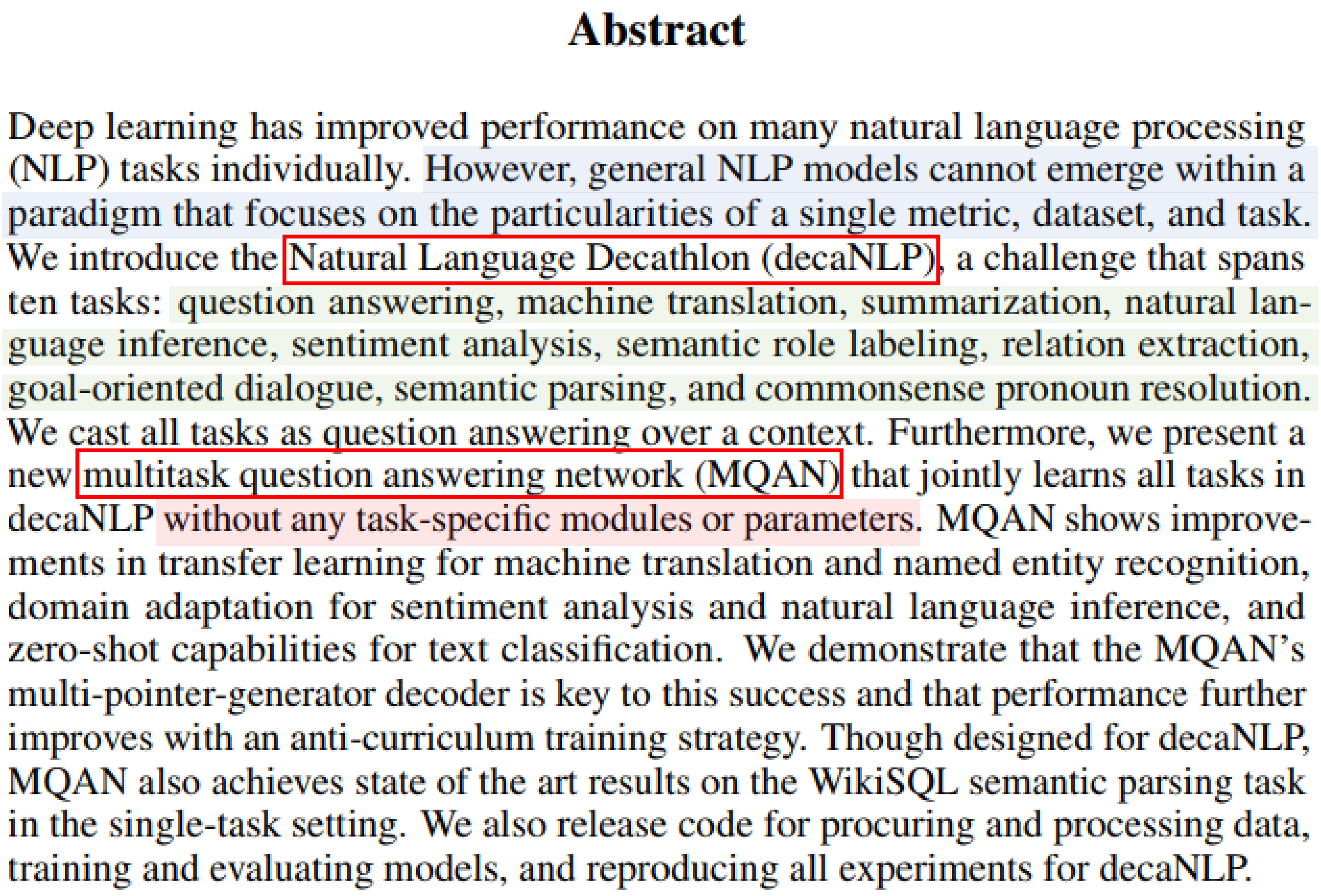
Aloha, I am currently the CEO of a new startup. You can join and learn about us here: https://su-sea.github.io/jobs/

Previously, I was the chief scientist (EVP) at Salesforce where I lead teams working on fundamental research, applied research, product incubation, CRM search, customer service automation and a cross-product AI platform for unstructured and structured data. Before that I was an adjunct professor at Stanford's computer science department and the founder and CEO/CTO of MetaMind which was acquired by Salesforce in 2016. In 2014, I got my PhD in the CS Department at Stanford. I like paramotor adventures, traveling and photography. More info:

- Forbes article with more info about my bio.
- New York Times article on a project at Salesforce Research.
- CS224n - NLP with Deep Learning class I used to teach.
- TEDx talk about where AI is today and where it's going.
- My Twitter account for announcements and photos.



DecaNLP



Abstract

Deep learning has improved performance on many natural language processing (NLP) tasks individually. However, general NLP models cannot emerge within a paradigm that focuses on the particularities of a single metric, dataset, and task. We introduce the Natural Language Decathlon (decaNLP), a challenge that spans ten tasks: question answering, machine translation, summarization, natural language inference, sentiment analysis, semantic role labeling, relation extraction, goal-oriented dialogue, semantic parsing, and commonsense pronoun resolution. We cast all tasks as question answering over a context. Furthermore, we present a new multitask question answering network (MQAN) that jointly learns all tasks in decaNLP without any task-specific modules or parameters. MQAN shows improvements in transfer learning for machine translation and named entity recognition, domain adaptation for sentiment analysis and natural language inference, and zero-shot capabilities for text classification. We demonstrate that the MQAN's multi-pointer-generator decoder is key to this success and that performance further improves with an anti-curriculum training strategy. Though designed for decaNLP, MQAN also achieves state of the art results on the WikiSQL semantic parsing task in the single-task setting. We also release code for procuring and processing data, training and evaluating models, and reproducing all experiments for decaNLP.

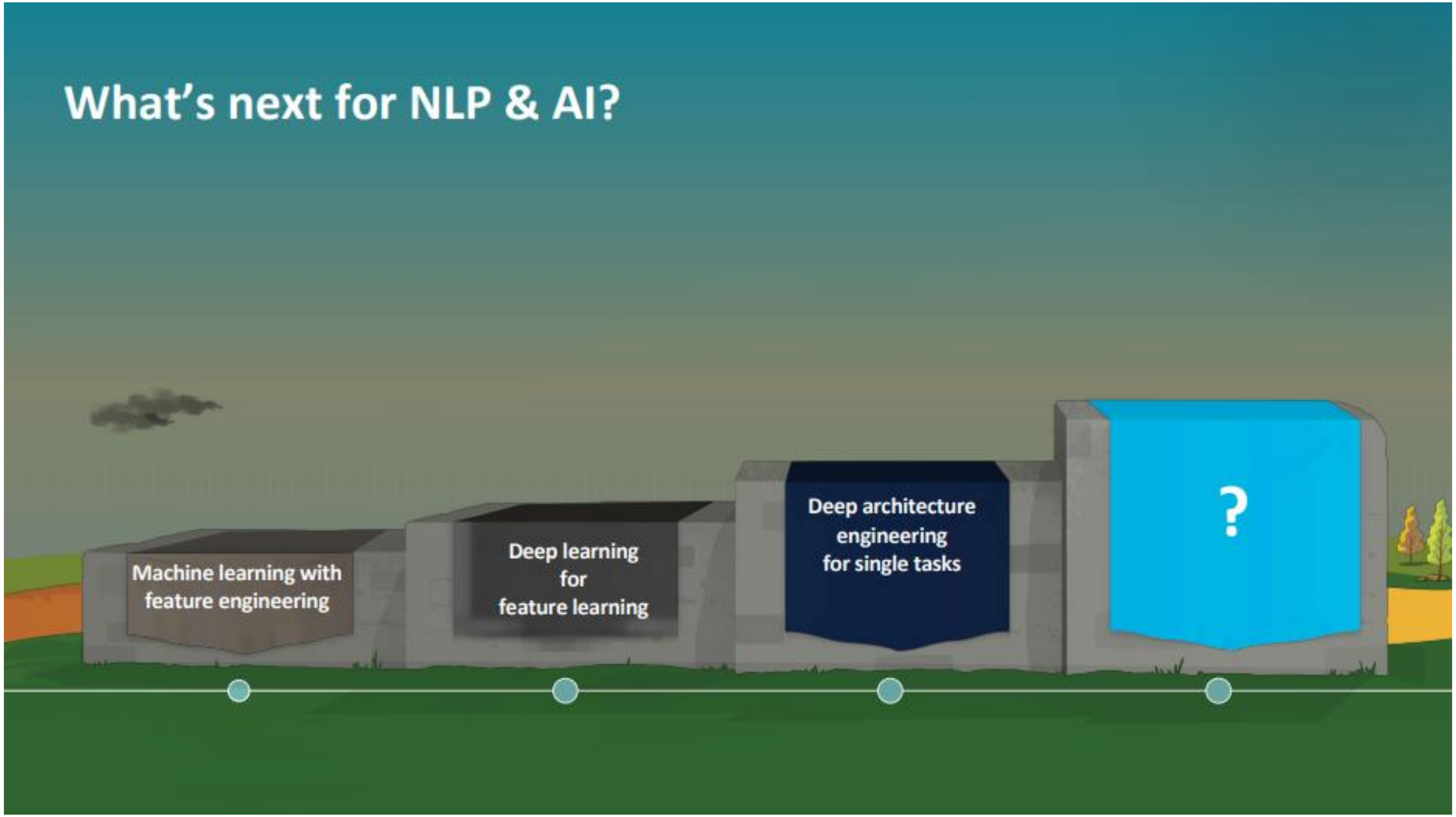
Single task limitation

이 10가지 task를 QA over a Context로 cast

decaNLP

단일 task에 대한 성과도 좋았음

Introduction



What's next for NLP & AI?

Machine learning with feature engineering
Deep learning for feature learning
Deep architecture engineering for single tasks
?

Introduction

- NLP-task들에 대한 일반화된 모델 제안
  - 단일 모델(Unified)이 상기 10가지 task를 동시에 수행할 수 있는 능력을 갖추게 함
- Meta-Learning Approach

$$(x, y); t \rightarrow (x, y, t)$$

- 기존의 단일 task에 대한 Deep Architecture Engineering은 project가 다 끝난 이후, 모든 걸 다시 시작해야 함
  - Initialization
  - Step size
  - Regularize, Model Architecture → Task-Specific…!
- BERT is non task-specific model? Fine-tuning이 가지는 Catastrophic Forgetting!

Introduction

- 우리가 제안하고자 하는 NLP 모델은 Unified Multitask Model!
- 기본적으로 Seq2Seq model
    + pointer network
    + Co-Attention (Advanced Attention Mechanism)
    + MultiHead Self-Attention
    + Question Answering
    + Curriculum Learning
- 우리가 제안한 모델이 들어갈 library decaNLP!
    - 뭐든 팝니다! 이 resource로
    - Multitask Learning, Transfer Learning, General Embeddings and Encoders,
    - Architecture Search, Zero-Shot Learning, General Purpose Question Answering,
    - Meat-Leaning, Other related areas of NLP

Introduction

A) We do general QA!

Q) Can I ask your model what the sentiment is this tweet?

A) No, that’s sentiment analysis. Go to that different workshop. It’s down the hall

Q) That is a question. Why can’t you answer it in the general QA workshop?

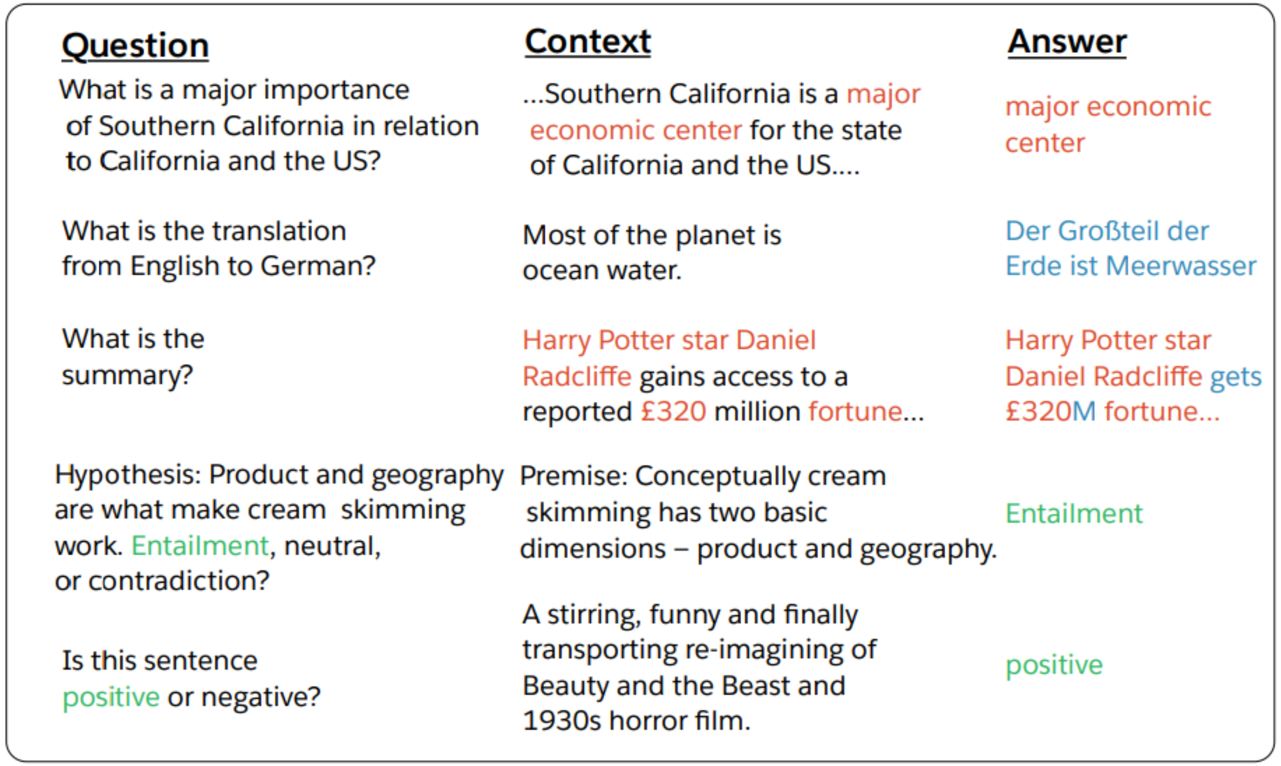
A) Well, if you want to work on more general stuff, it has to be an unsupervised. Kind of task and the feature will not be supervised.

Why a unified multi-task model for NLP ?

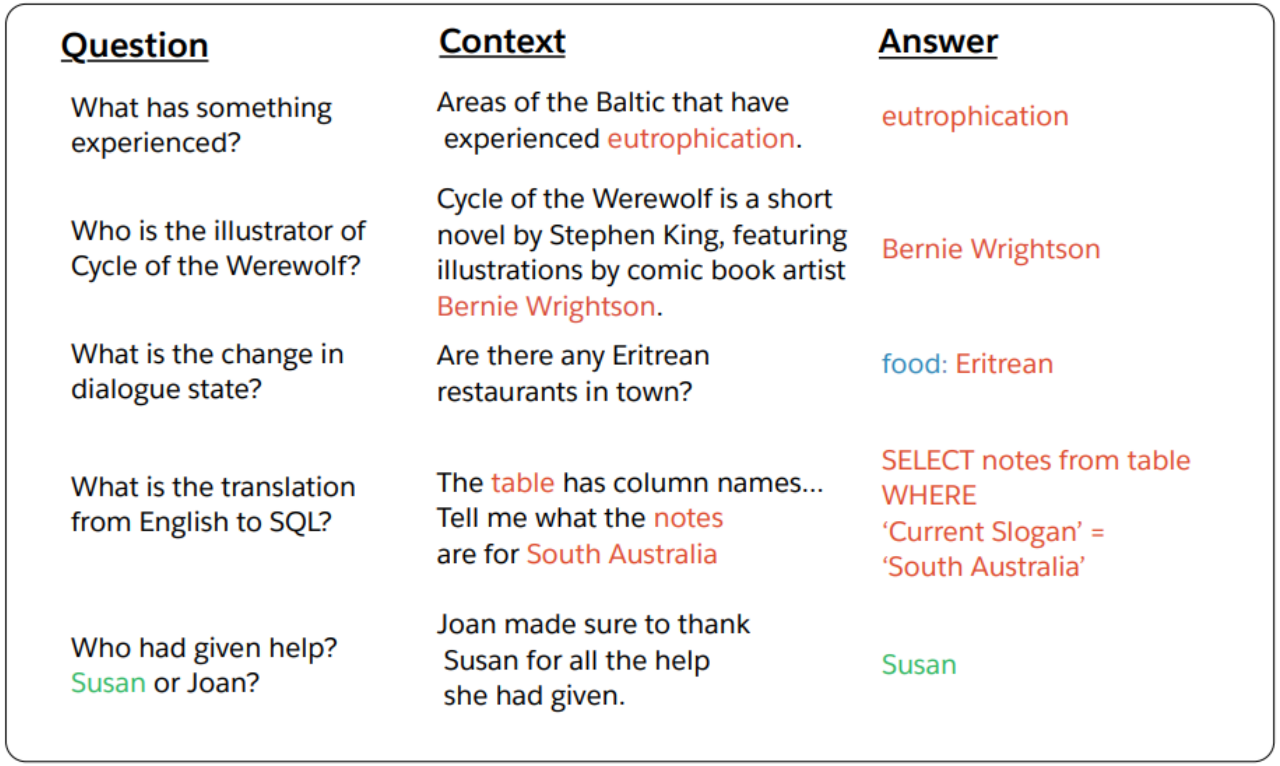
- Multi-task learning 은 General NLP System 이 넘어야할 큰 장벽
- 하나의 통합된 모델은 지식을 어떻게 전달할지 결정할 수 있음
  - Domain adaptation, weight sharing, transfer and zero shot learning

- 하나의 통합된 Multi-task 모델은 …
  - 새로운 task 가 주어졌을 때 쉽게 적응할 수 있음
  - 실제 production 을 위해 deploy 하는 것이 매우 간단해짐
  - 더 많은 사람들이 새로운 task 를 해결할 수 있도록 진입장벽을 낮춰 줌
  - 잠재적으로 Continual learning 으로 나아갈 수 있음

Tasks and Metrics

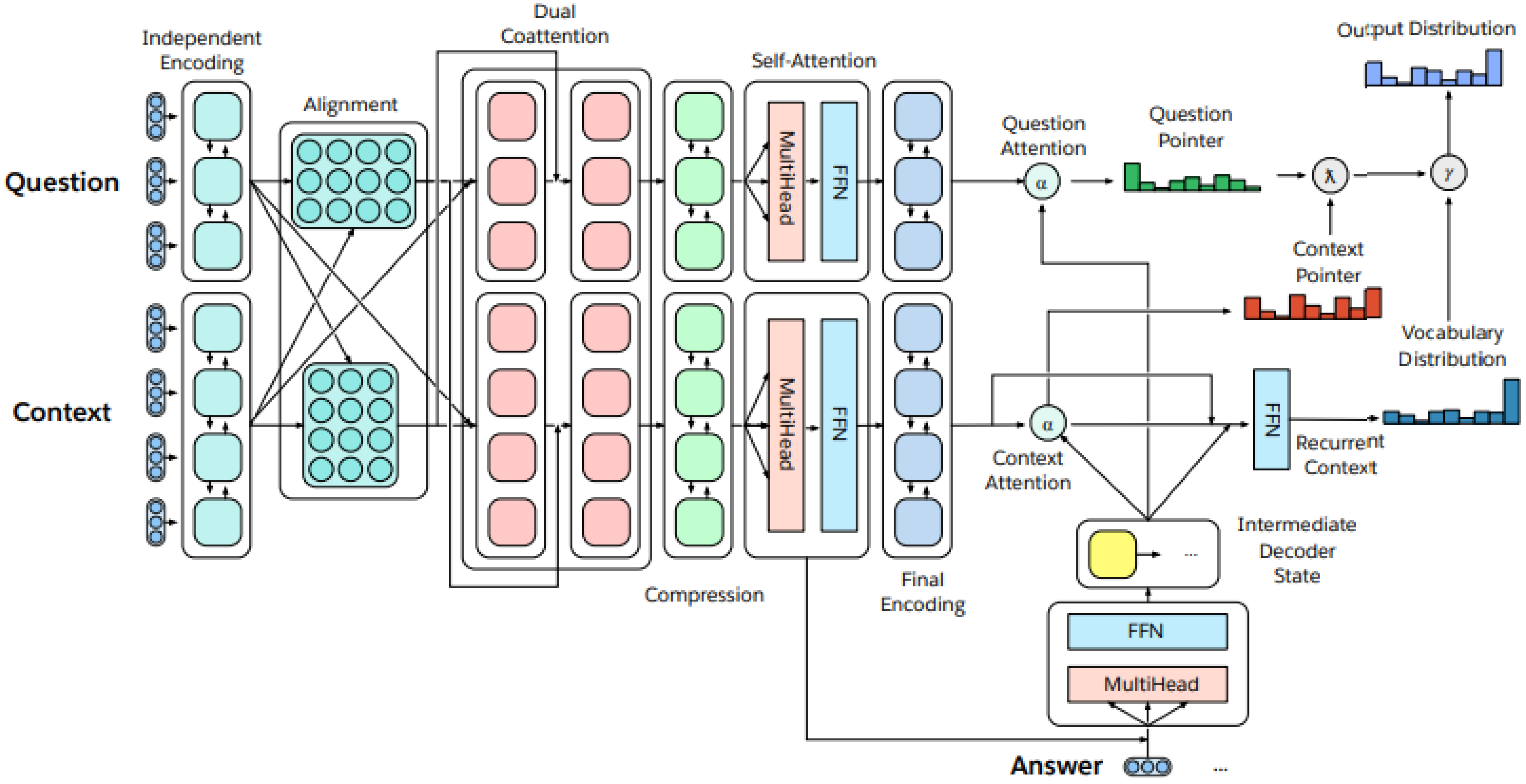


Question	Context	Answer
What is a major importance of Southern California in relation to California and the US?	...Southern California is a major economic center for the state of California and the US....	major economic center
What is the translation from English to German?	Most of the planet is ocean water.	Der Großteil der Erde ist Meerwasser
What is the summary?	Harry Potter star Daniel Radcliffe gains access to a reported £320 million fortune...	Harry Potter star Daniel Radcliffe gets £320M fortune...
Hypothesis: Product and geography are what make cream skimming work. Entailment, neutral, or contradiction?	Premise: Conceptually cream skimming has two basic dimensions – product and geography.	Entailment
Is this sentence positive or negative?	A stirring, funny and finally transporting re-imagining of Beauty and the Beast and 1930s horror film.	positive



Question	Context	Answer
What has something experienced?	Areas of the Baltic that have experienced eutrophication.	eutrophication
Who is the illustrator of Cycle of the Werewolf?	Cycle of the Werewolf is a short novel by Stephen King, featuring illustrations by comic book artist Bernie Wrightson.	Bernie Wrightson
What is the change in dialogue state?	Are there any Eritrean restaurants in town?	food: Eritrean
What is the translation from English to SQL?	The table has column names... Tell me what the notes are for South Australia	SELECT notes from table WHERE 'Current Slogan' = 'South Australia'
Who had given help? Susan or Joan?	Joan made sure to thank Susan for all the help she had given.	Susan

Multitask Question Answering Network (MQAN)



Multitask Question Answering Network (MQAN)

Encoder

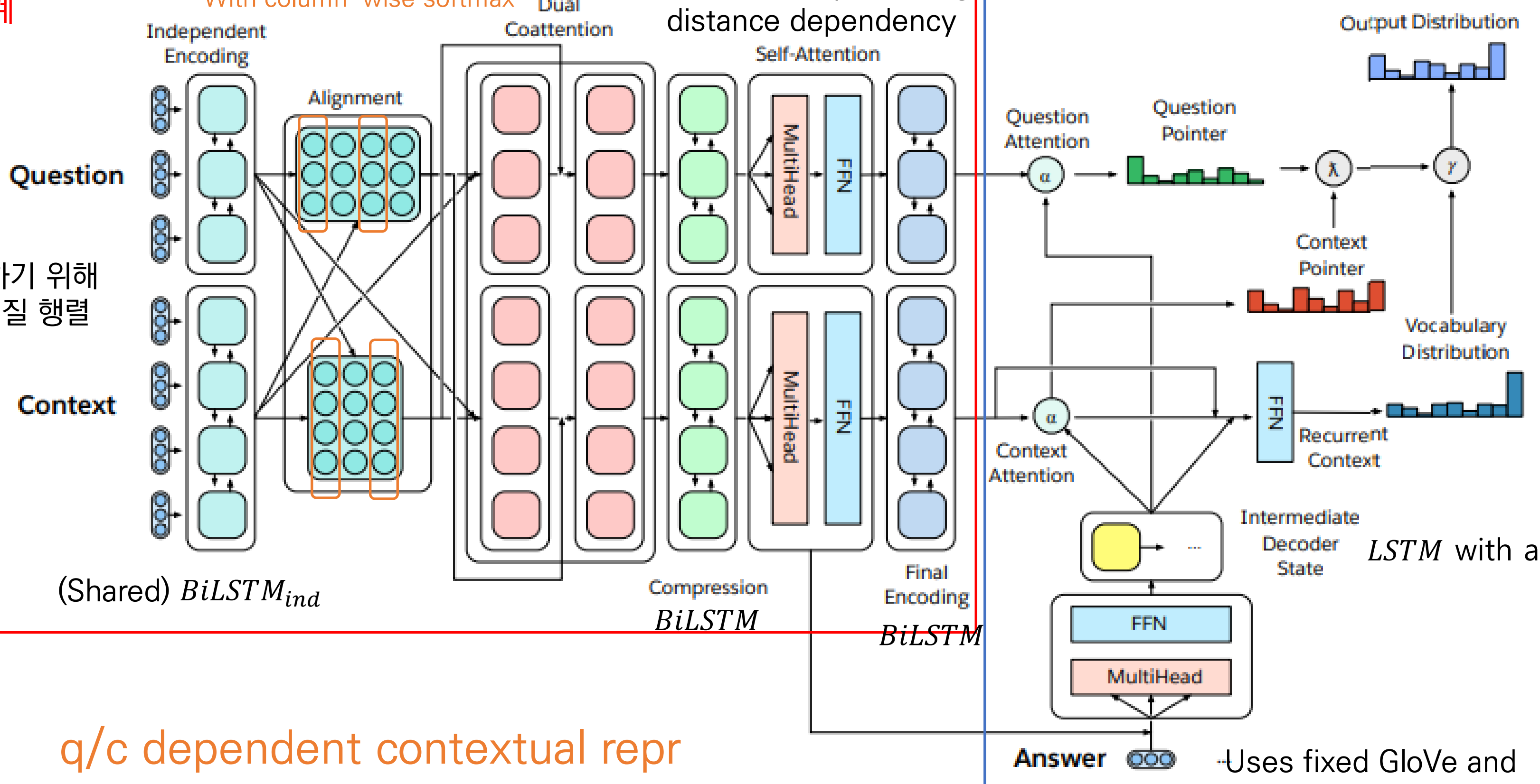
Research Director가 제안한
발전된 Attention 기법

Auto-Regressive Decoder

Local & Global interdependency
를 포착하기 위해 설계

With column-wise softmax

In order to capture long
distance dependency



최종 출력 분포에 영향을 줄
각 분포의 중요성을 조절
→ 어디서 단어를 고를지!

answer를 생성하기 위해
decoder에 주어질 행렬

(Shared) $BiLSTM_{ind}$

$BiLSTM$

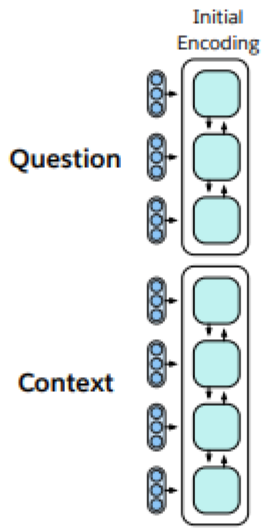
$BiLSTM$

$LSTM$ with attention

Uses fixed GloVe and
Character n-gram embedding

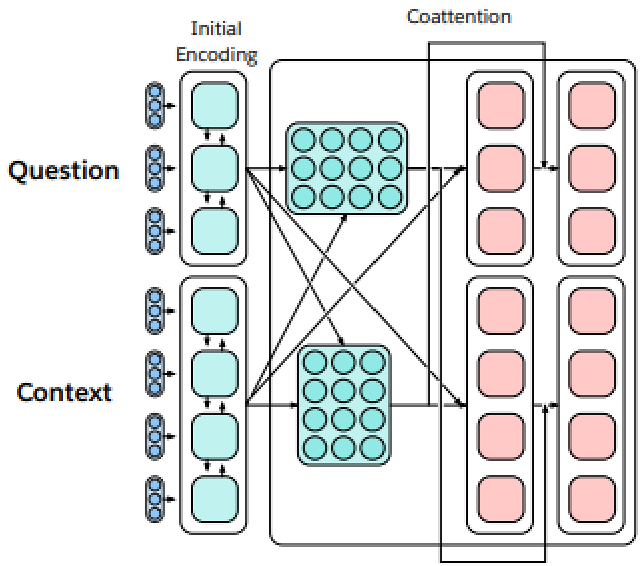
q/c dependent contextual repr

Multitask Question Answering Network (MQAN)



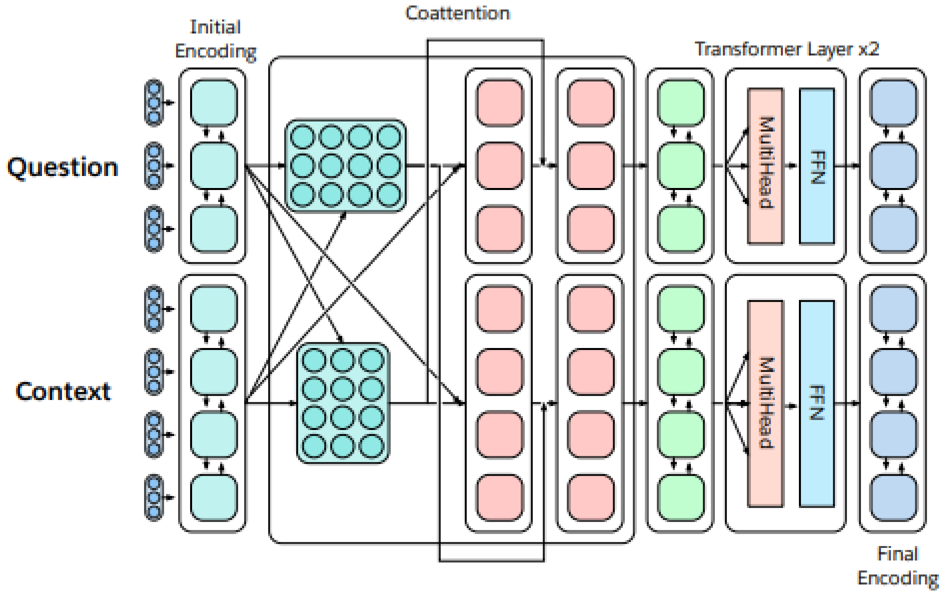
Fixed Glove+Character n-gram embeddings → Linear → Shared BiLSTM with skip connection

Multitask Question Answering Network (MQAN)



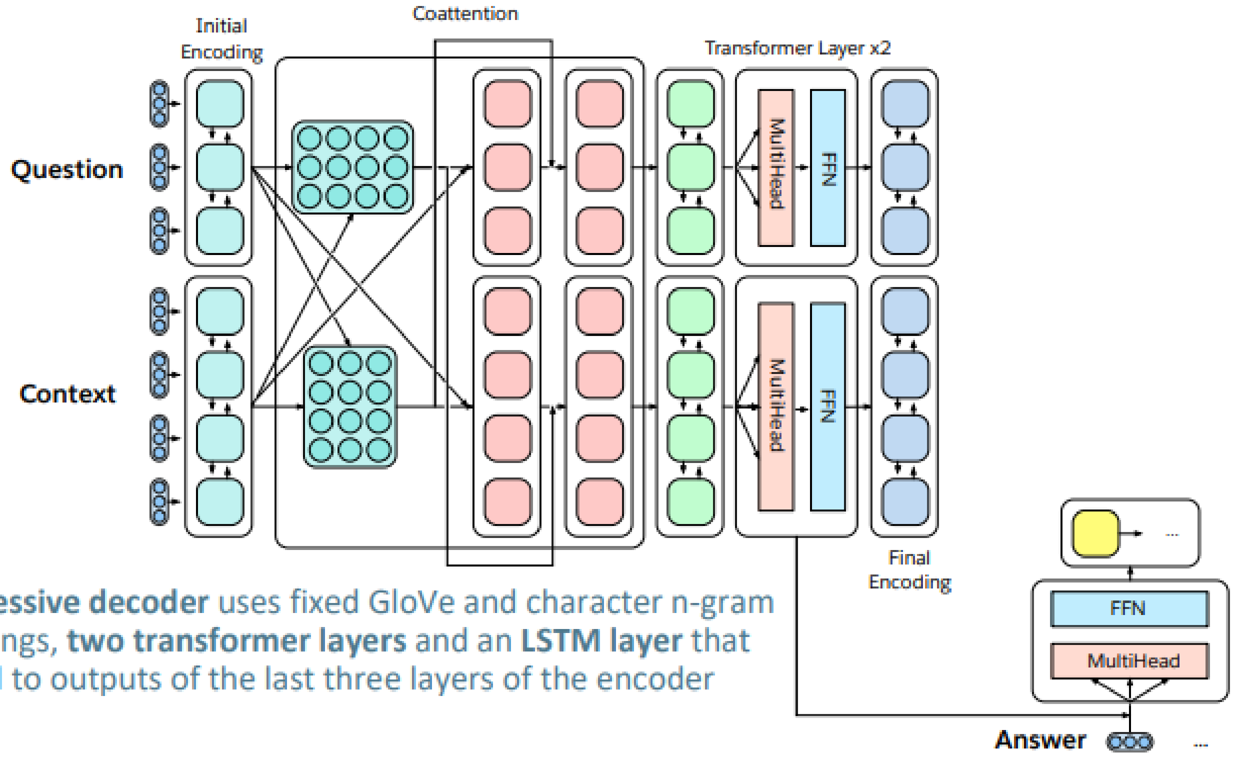
Attention summations from one sequence to the other and back again with skip connections

Multitask Question Answering Network (MQAN)



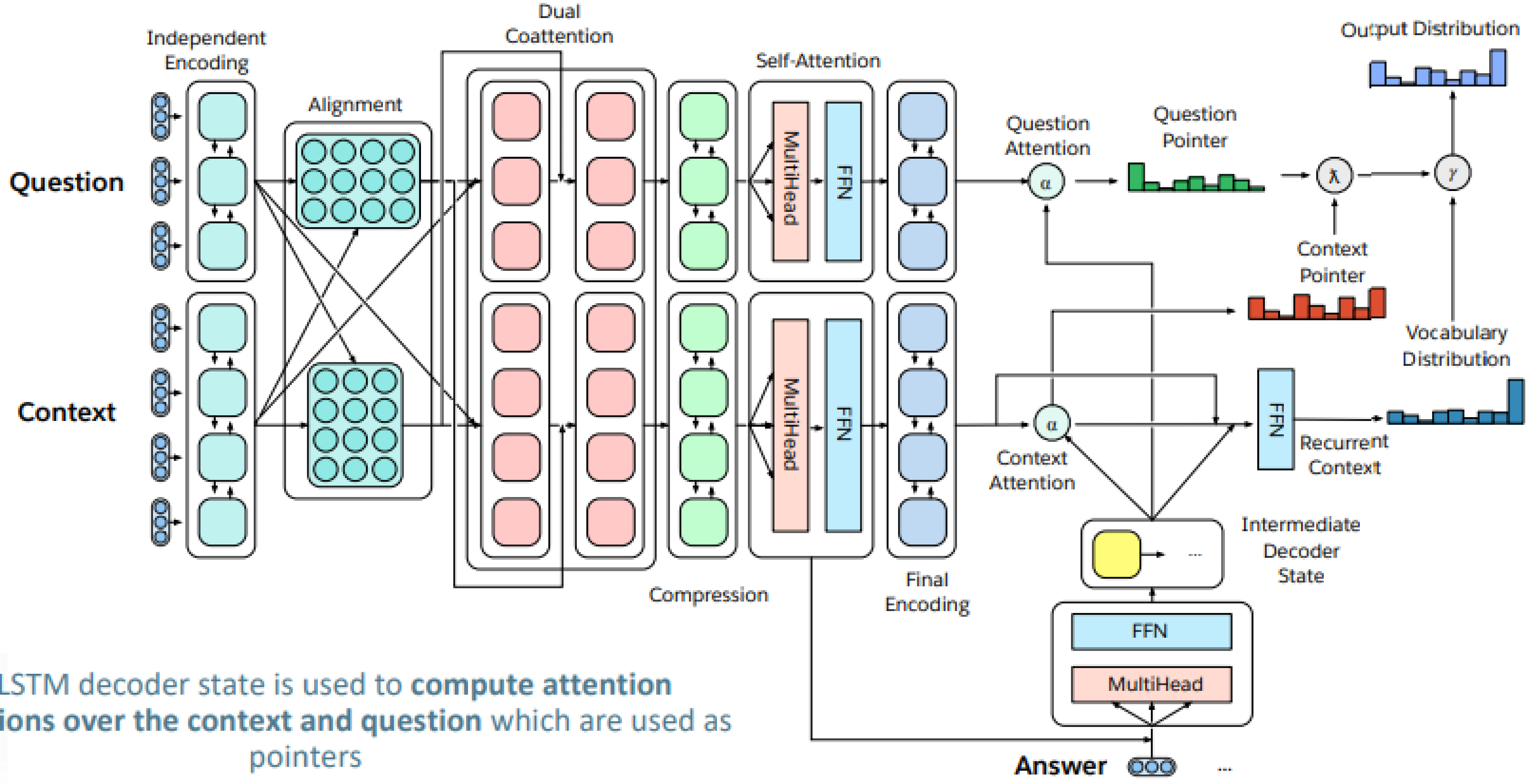
Separate BiLSTMs to reduce dimensionality, two transformer layers, another BiLSTM

Multitask Question Answering Network (MQAN)



Auto-regressive decoder uses fixed GloVe and character n-gram embeddings, **two transformer layers** and an **LSTM layer** that attend to outputs of the last three layers of the encoder

Multitask Question Answering Network (MQAN)



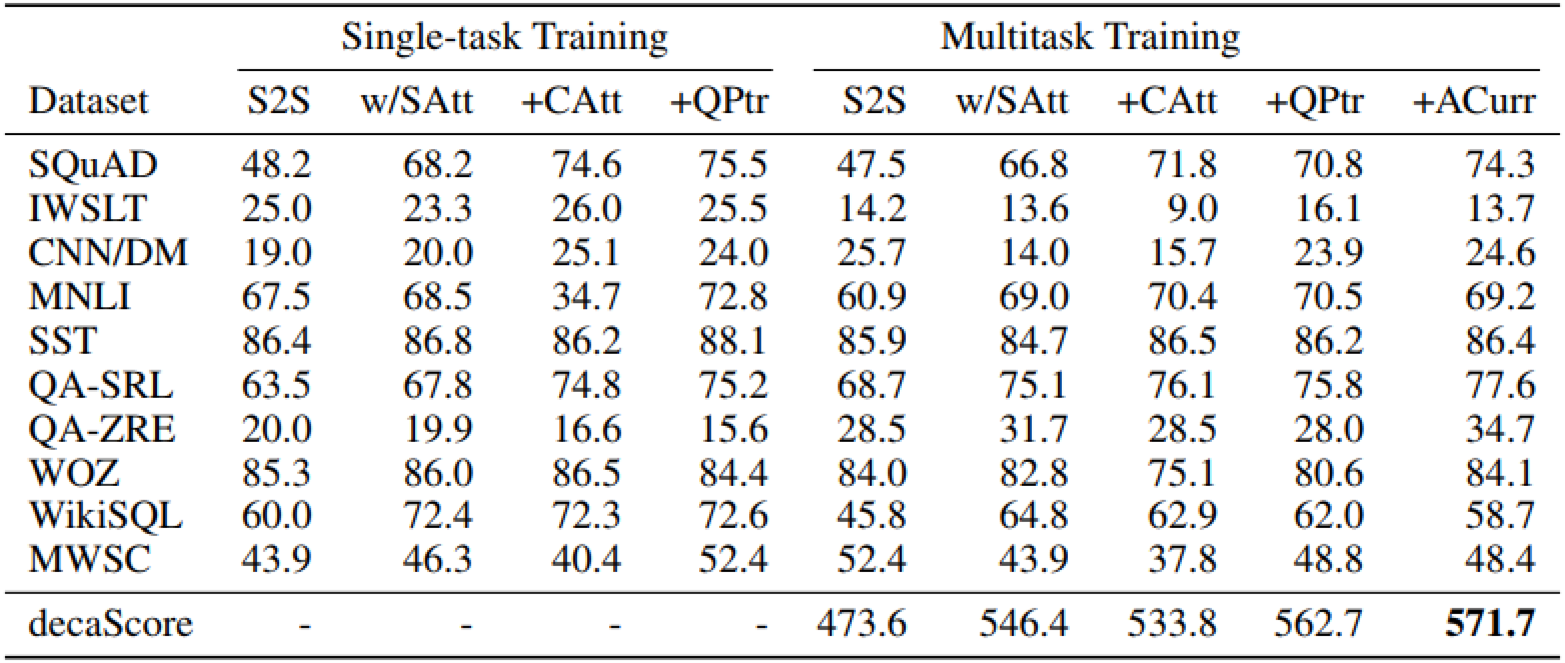
The LSTM decoder state is used to **compute attention distributions over the context and question** which are used as pointers

1. γ Switch
   External Vocabulary에서 복사/선택할 지를 결정
2. λ Switch
   Context/Question 중 어디서 copy할지를 결정

Experiments and Analysis

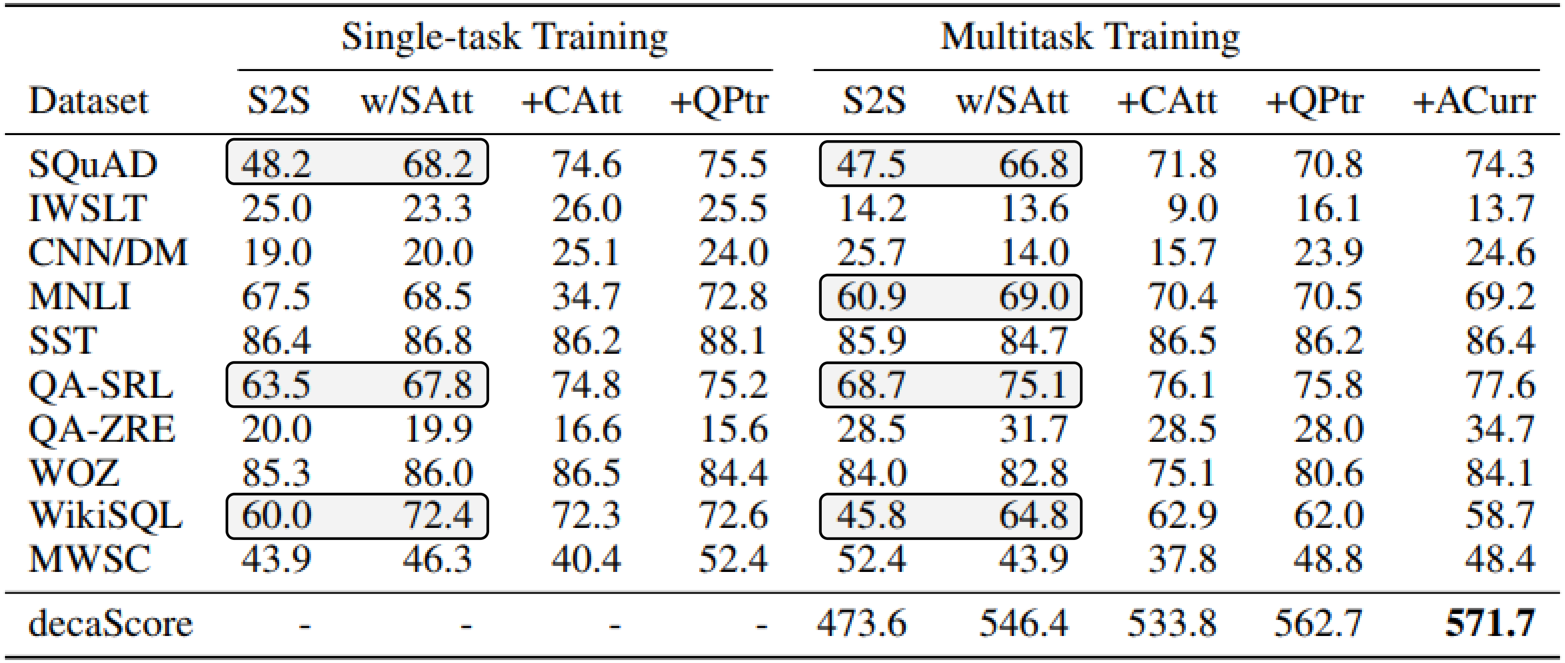
왼쪽은 single task, 각 task별 모델이 존재! 즉, 40개의 모델!

Task별 모델이 동일!! 즉, 4개의 모델



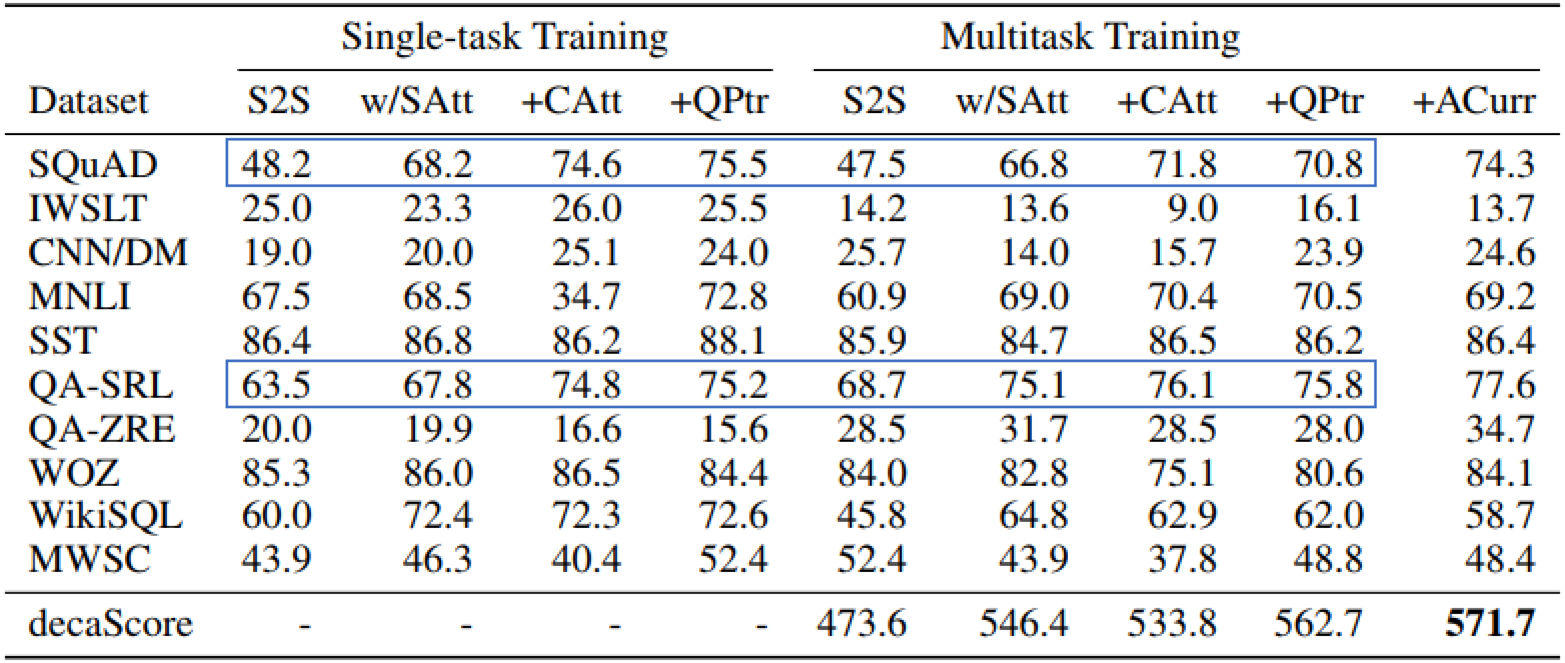
	Single-task Training				Multitask Training				
Dataset	S2S	w/SAtt	+CAtt	+QPtr	S2S	w/SAtt	+CAtt	+QPtr	+ACurr
SQuAD	48.2	68.2	74.6	75.5	47.5	66.8	71.8	70.8	74.3
IWSLT	25.0	23.3	26.0	25.5	14.2	13.6	9.0	16.1	13.7
CNN/DM	19.0	20.0	25.1	24.0	25.7	14.0	15.7	23.9	24.6
MNLI	67.5	68.5	34.7	72.8	60.9	69.0	70.4	70.5	69.2
SST	86.4	86.8	86.2	88.1	85.9	84.7	86.5	86.2	86.4
QA-SRL	63.5	67.8	74.8	75.2	68.7	75.1	76.1	75.8	77.6
QA-ZRE	20.0	19.9	16.6	15.6	28.5	31.7	28.5	28.0	34.7
WOZ	85.3	86.0	86.5	84.4	84.0	82.8	75.1	80.6	84.1
WikiSQL	60.0	72.4	72.3	72.6	45.8	64.8	62.9	62.0	58.7
MWSC	43.9	46.3	40.4	52.4	52.4	43.9	37.8	48.8	48.4
decaScore	-	-	-	-	473.6	546.4	533.8	562.7	**571.7**

Experiments and Analysis



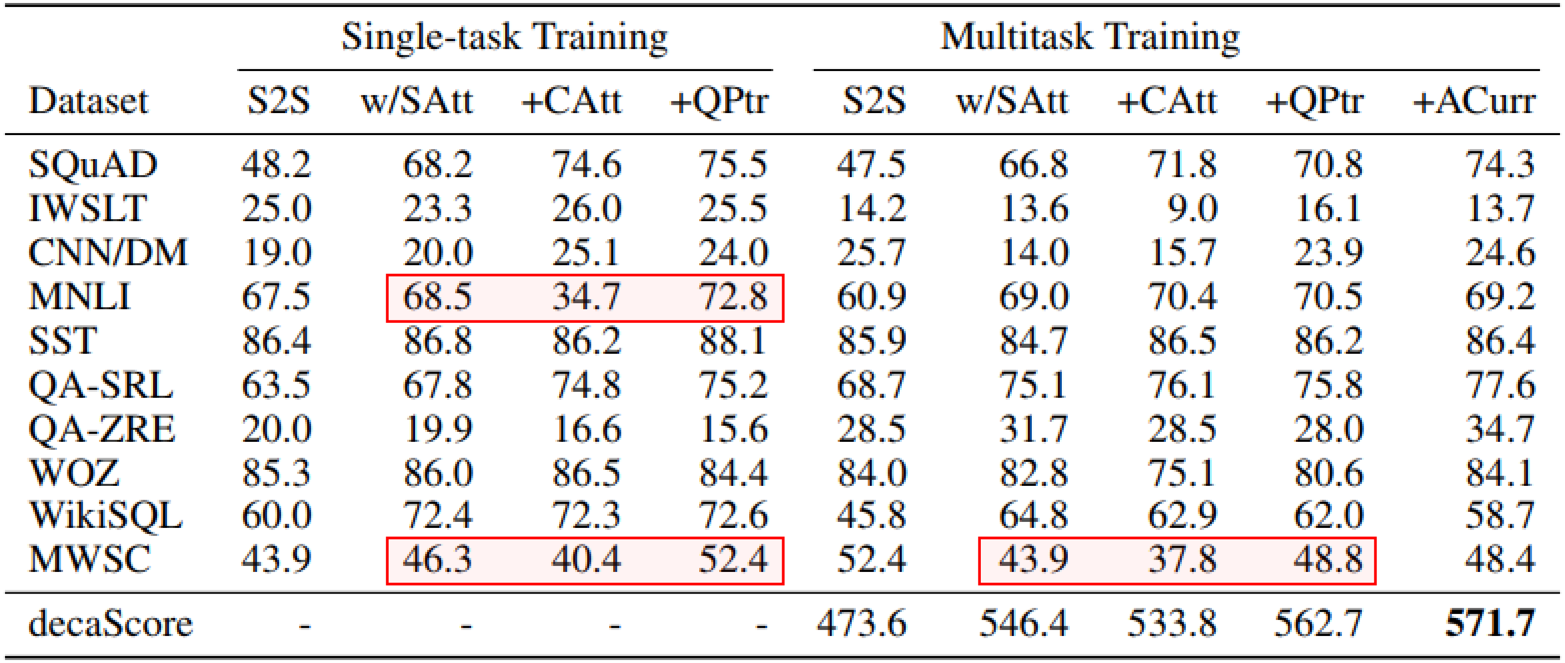
	Single-task Training				Multitask Training				
Dataset	S2S	w/SAtt	+CAtt	+QPtr	S2S	w/SAtt	+CAtt	+QPtr	+ACurr
SQuAD	48.2	68.2	74.6	75.5	47.5	66.8	71.8	70.8	74.3
IWSLT	25.0	23.3	26.0	25.5	14.2	13.6	9.0	16.1	13.7
CNN/DM	19.0	20.0	25.1	24.0	25.7	14.0	15.7	23.9	24.6
MNLI	67.5	68.5	34.7	72.8	60.9	69.0	70.4	70.5	69.2
SST	86.4	86.8	86.2	88.1	85.9	84.7	86.5	86.2	86.4
QA-SRL	63.5	67.8	74.8	75.2	68.7	75.1	76.1	75.8	77.6
QA-ZRE	20.0	19.9	16.6	15.6	28.5	31.7	28.5	28.0	34.7
WOZ	85.3	86.0	86.5	84.4	84.0	82.8	75.1	80.6	84.1
WikiSQL	60.0	72.4	72.3	72.6	45.8	64.8	62.9	62.0	58.7
MWSC	43.9	46.3	40.4	52.4	52.4	43.9	37.8	48.8	48.4
decaScore	-	-	-	-	473.6	546.4	533.8	562.7	**571.7**

Experiments and Analysis



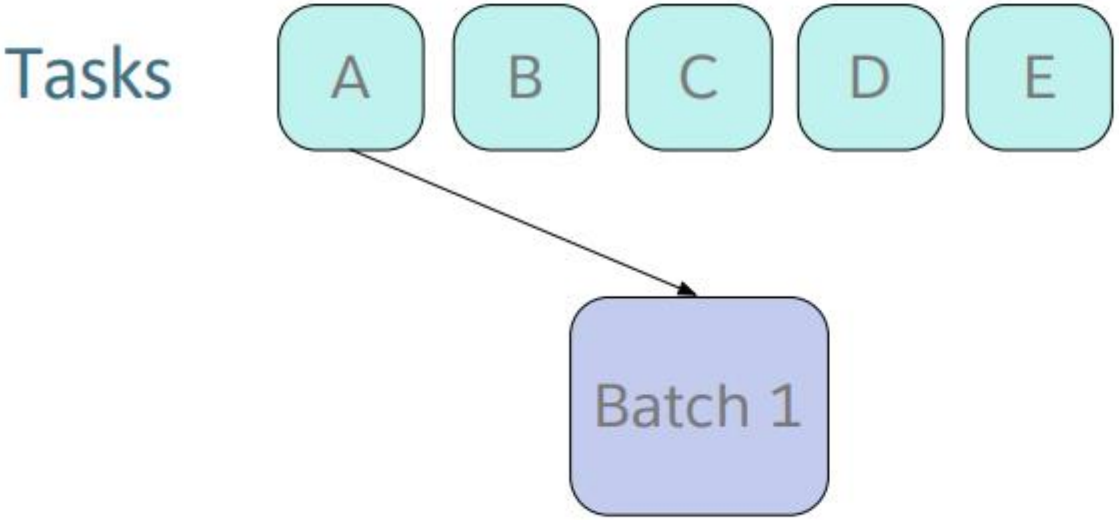
	Single-task Training				Multitask Training				
Dataset	S2S	w/SAtt	+CAtt	+QPtr	S2S	w/SAtt	+CAtt	+QPtr	+ACurr
SQuAD	48.2	68.2	74.6	75.5	47.5	66.8	71.8	70.8	74.3
IWSLT	25.0	23.3	26.0	25.5	14.2	13.6	9.0	16.1	13.7
CNN/DM	19.0	20.0	25.1	24.0	25.7	14.0	15.7	23.9	24.6
MNLI	67.5	68.5	34.7	72.8	60.9	69.0	70.4	70.5	69.2
SST	86.4	86.8	86.2	88.1	85.9	84.7	86.5	86.2	86.4
QA-SRL	63.5	67.8	74.8	75.2	68.7	75.1	76.1	75.8	77.6
QA-ZRE	20.0	19.9	16.6	15.6	28.5	31.7	28.5	28.0	34.7
WOZ	85.3	86.0	86.5	84.4	84.0	82.8	75.1	80.6	84.1
WikiSQL	60.0	72.4	72.3	72.6	45.8	64.8	62.9	62.0	58.7
MWSC	43.9	46.3	40.4	52.4	52.4	43.9	37.8	48.8	48.4
decaScore	-	-	-	-	473.6	546.4	533.8	562.7	**571.7**

Experiments and Analysis



	Single-task Training				Multitask Training				
Dataset	S2S	w/SAtt	+CAtt	+QPtr	S2S	w/SAtt	+CAtt	+QPtr	+ACurr
SQuAD	48.2	68.2	74.6	75.5	47.5	66.8	71.8	70.8	74.3
IWSLT	25.0	23.3	26.0	25.5	14.2	13.6	9.0	16.1	13.7
CNN/DM	19.0	20.0	25.1	24.0	25.7	14.0	15.7	23.9	24.6
MNLI	67.5	68.5	34.7	72.8	60.9	69.0	70.4	70.5	69.2
SST	86.4	86.8	86.2	88.1	85.9	84.7	86.5	86.2	86.4
QA-SRL	63.5	67.8	74.8	75.2	68.7	75.1	76.1	75.8	77.6
QA-ZRE	20.0	19.9	16.6	15.6	28.5	31.7	28.5	28.0	34.7
WOZ	85.3	86.0	86.5	84.4	84.0	82.8	75.1	80.6	84.1
WikiSQL	60.0	72.4	72.3	72.6	45.8	64.8	62.9	62.0	58.7
MWSC	43.9	46.3	40.4	52.4	52.4	43.9	37.8	48.8	48.4
decaScore	-	-	-	-	473.6	546.4	533.8	562.7	**571.7**

Training Strategies: Fully Joint

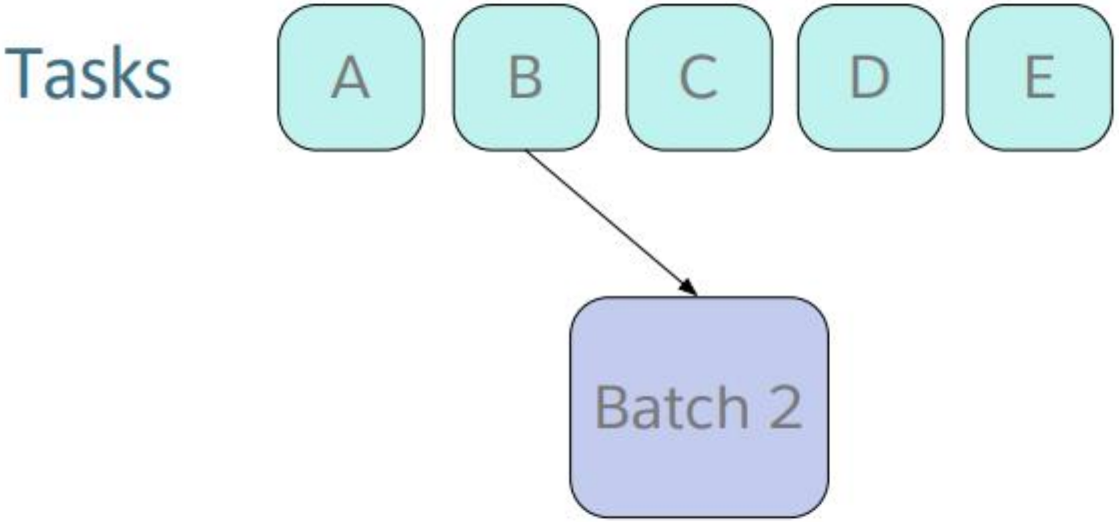


Tasks

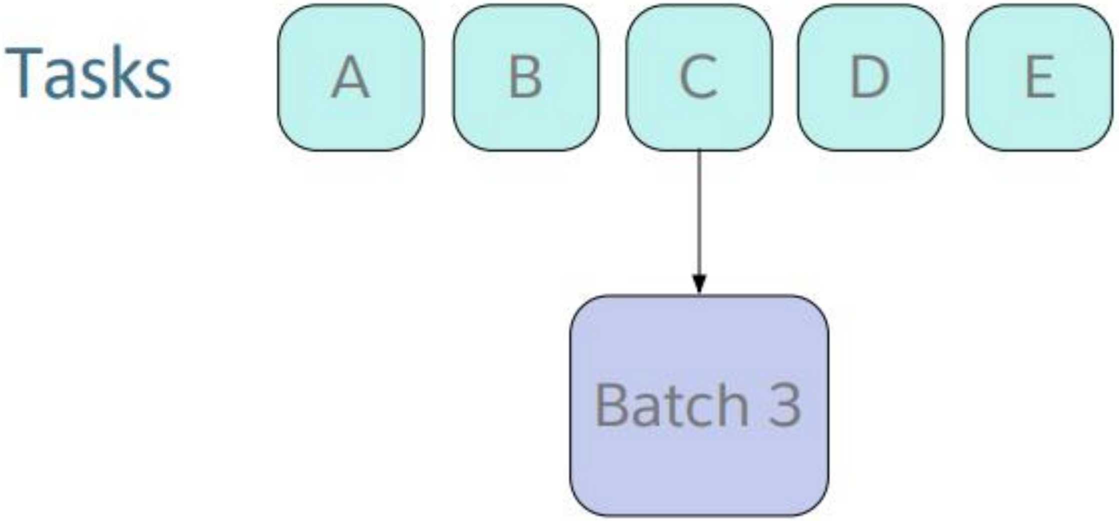
A B C D E

Batch 1

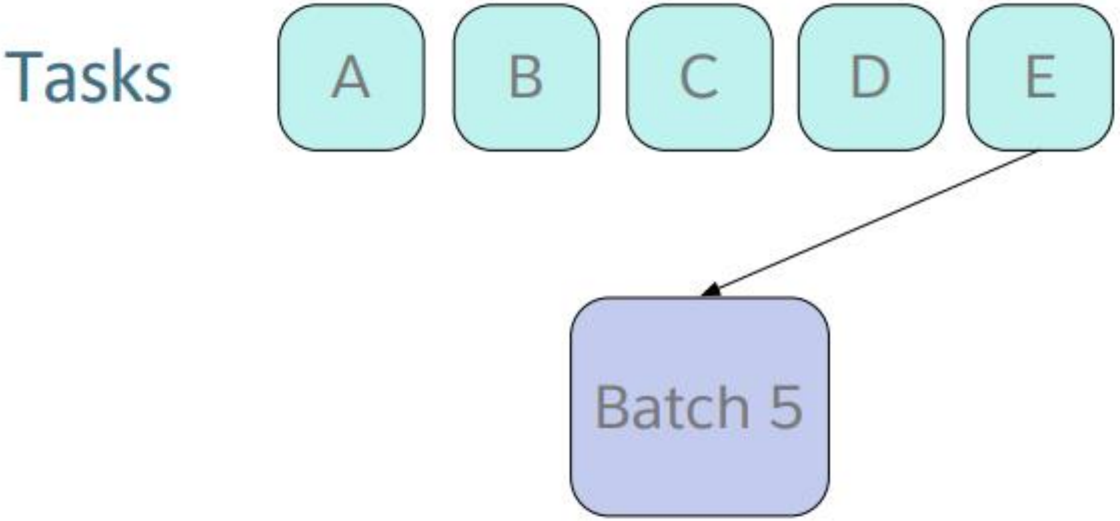
Training Strategies: Fully Joint



Training Strategies: Fully Joint



Training Strategies: Fully Joint



Tasks

A

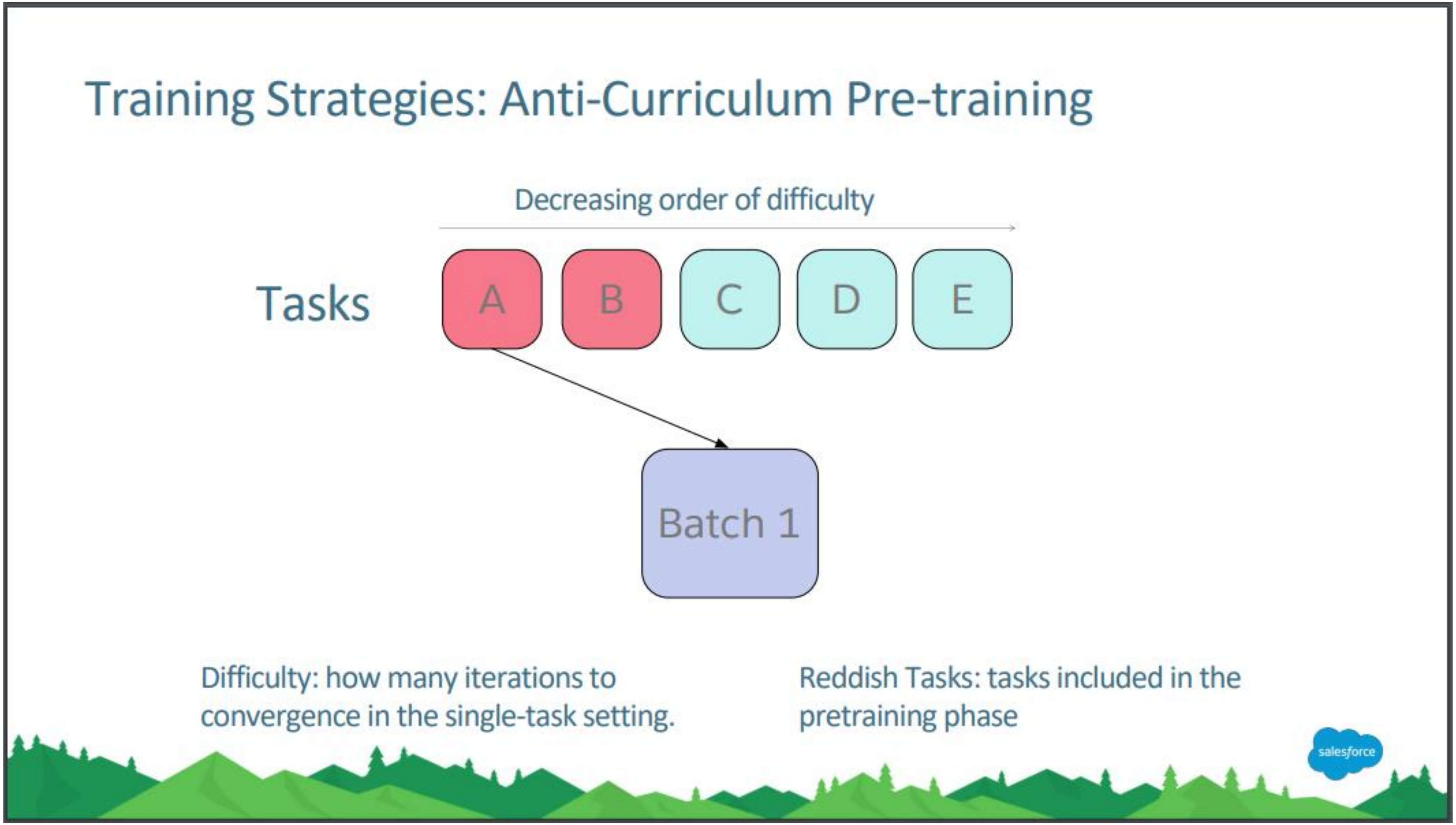
B

C

D

E

Batch 5



Training Strategies: Anti-Curriculum Pre-training

Decreasing order of difficulty

Tasks

A

B

C

D

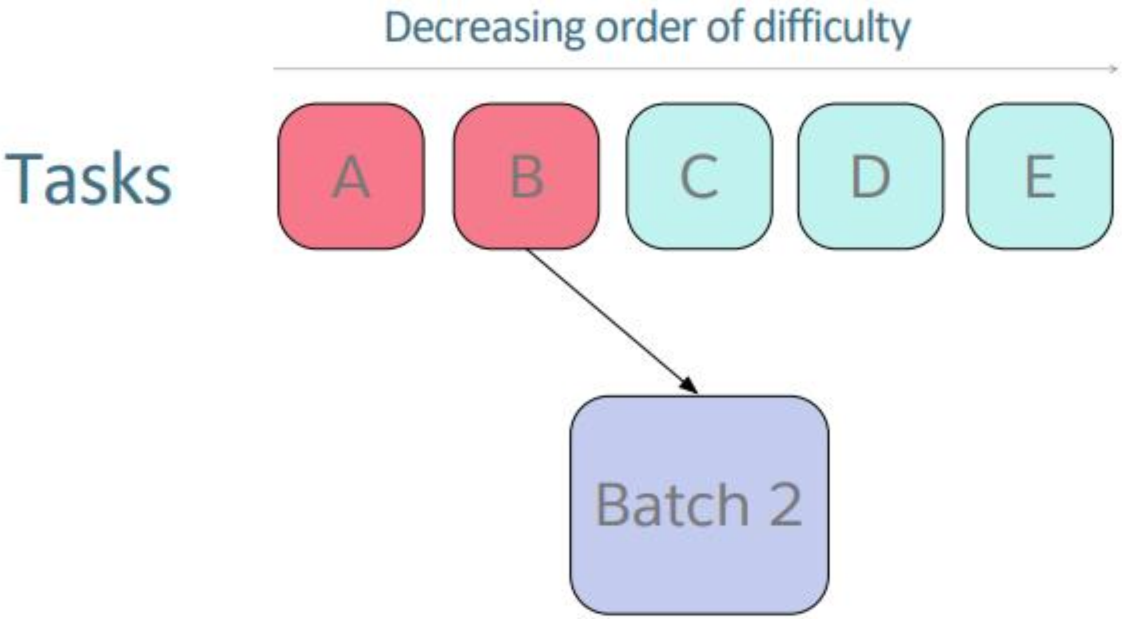
E

Batch 1

Difficulty: how many iterations to convergence in the single-task setting.

Reddish Tasks: tasks included in the pretraining phase

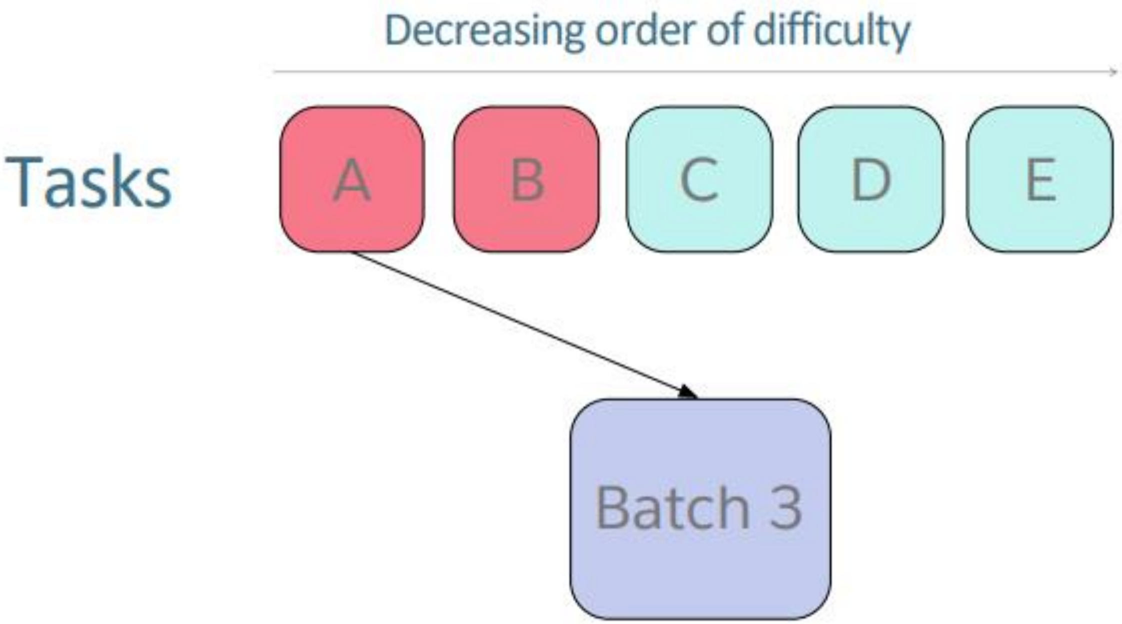
Training Strategies: Anti-Curriculum Pre-training



Difficulty: how many iterations to convergence in the single-task setting.

Reddish Tasks: tasks included in the pretraining phase

Training Strategies: Anti-Curriculum Pre-training



Decreasing order of difficulty

Tasks

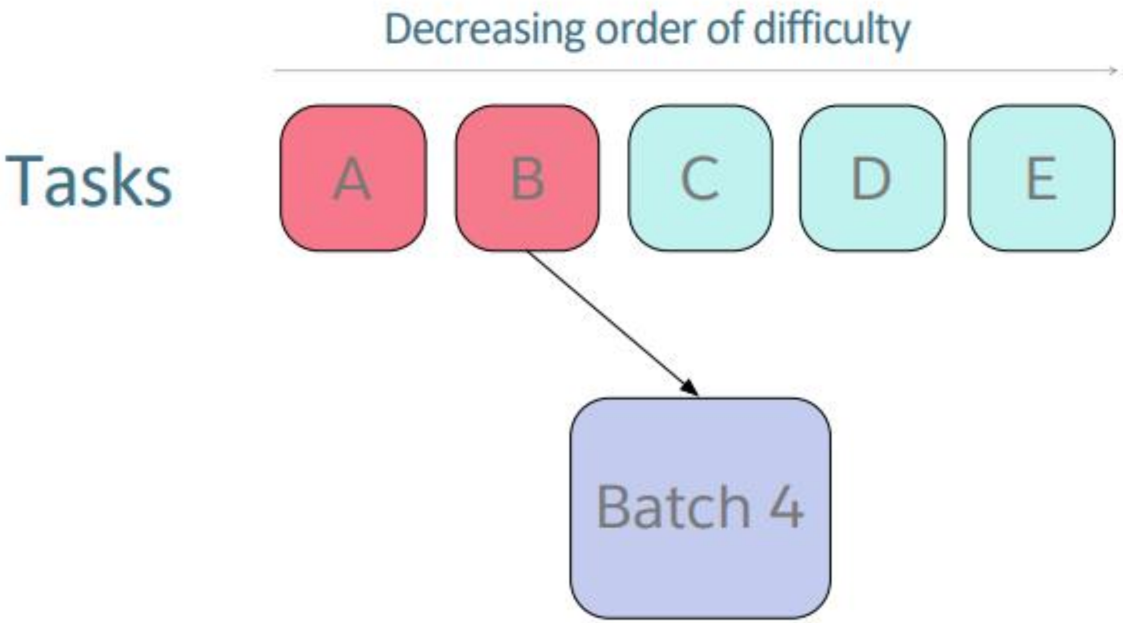
A B C D E

Batch 3

Difficulty: how many iterations to convergence in the single-task setting.

Reddish Tasks: tasks included in the pretraining phase

Training Strategies: Anti-Curriculum Pre-training



Decreasing order of difficulty

Tasks

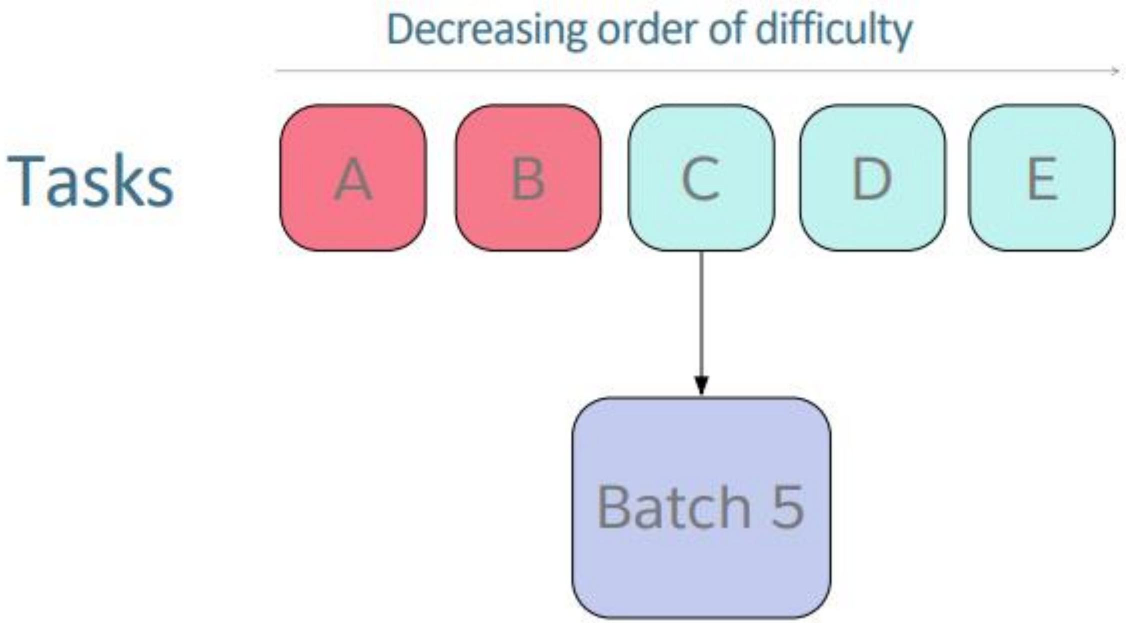
A B C D E

Batch 4

Difficulty: how many iterations to convergence in the single-task setting.

Reddish Tasks: tasks included in the pretraining phase

Training Strategies: Anti-Curriculum Pre-training



Decreasing order of difficulty

Tasks

A

B

C

D

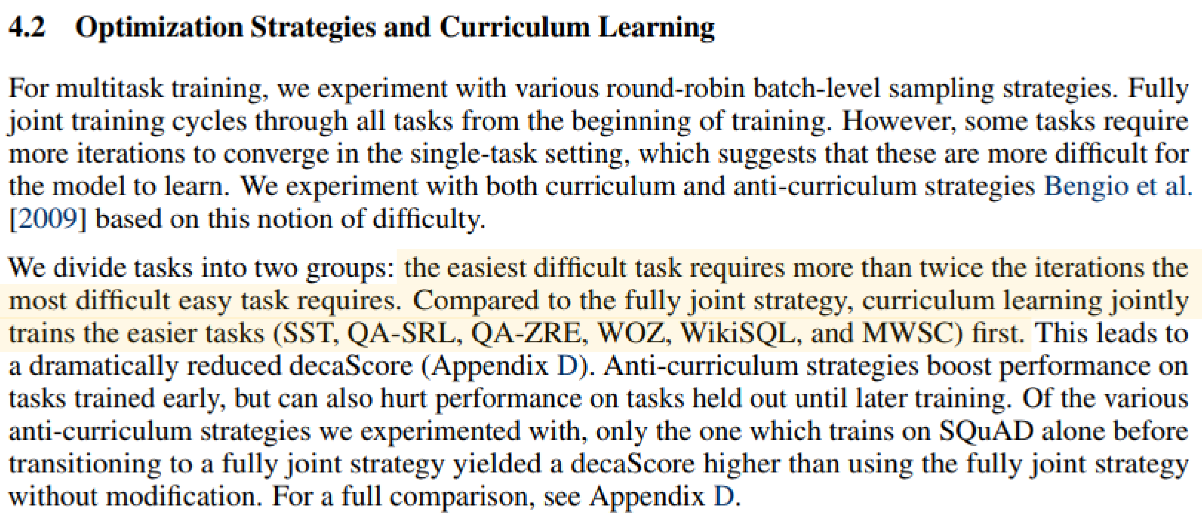
E

Batch 5

Difficulty: how many iterations to convergence in the single-task setting.

Reddish Tasks: tasks included in the pretraining phase

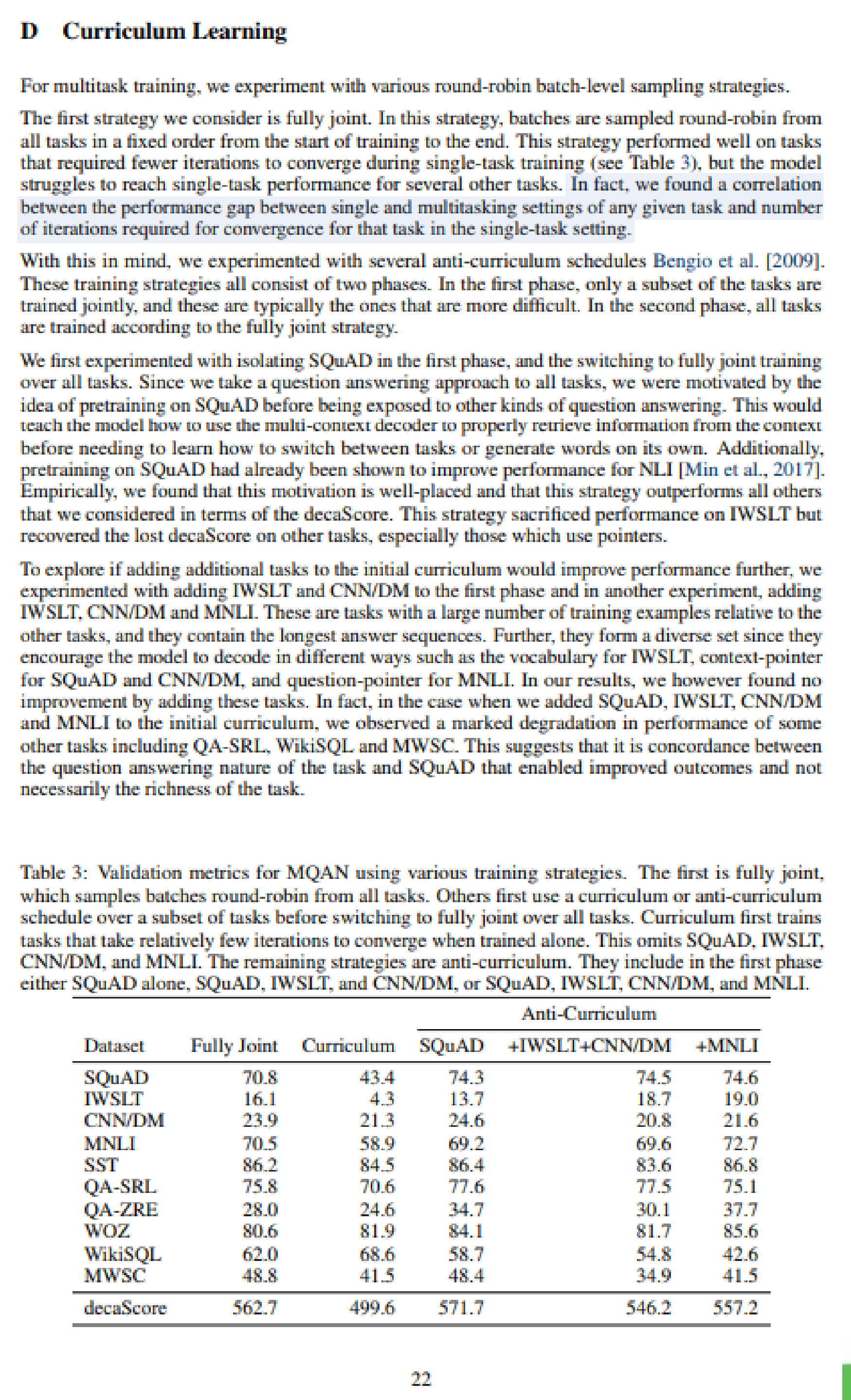
Experiments and Analysis



4.2 Optimization Strategies and Curriculum Learning

For multitask training, we experiment with various round-robin batch-level sampling strategies. Fully joint training cycles through all tasks from the beginning of training. However, some tasks require more iterations to converge in the single-task setting, which suggests that these are more difficult for the model to learn. We experiment with both curriculum and anti-curriculum strategies Bengio et al. [2009] based on this notion of difficulty.

We divide tasks into two groups: the easiest difficult task requires more than twice the iterations the most difficult easy task requires. Compared to the fully joint strategy, curriculum learning jointly trains the easier tasks (SST, QA-SRL, QA-ZRE, WOZ, WikiSQL, and MWSC) first. This leads to a dramatically reduced decaScore (Appendix D). Anti-curriculum strategies boost performance on tasks trained early, but can also hurt performance on tasks held out until later training. Of the various anti-curriculum strategies we experimented with, only the one which trains on SQuAD alone before transitioning to a fully joint strategy yielded a decaScore higher than using the fully joint strategy without modification. For a full comparison, see Appendix D.



D Curriculum Learning

For multitask training, we experiment with various round-robin batch-level sampling strategies.

The first strategy we consider is fully joint. In this strategy, batches are sampled round-robin from all tasks in a fixed order from the start of training to the end. This strategy performed well on tasks that required fewer iterations to converge during single-task training (see Table 3), but the model struggles to reach single-task performance for several other tasks. In fact, we found a correlation between the performance gap between single and multitasking settings of any given task and number of iterations required for convergence for that task in the single-task setting.

With this in mind, we experimented with several anti-curriculum schedules Bengio et al. [2009]. These training strategies all consist of two phases. In the first phase, only a subset of the tasks are trained jointly, and these are typically the ones that are more difficult. In the second phase, all tasks are trained according to the fully joint strategy.

We first experimented with isolating SQuAD in the first phase, and the switching to fully joint training over all tasks. Since we take a question answering approach to all tasks, we were motivated by the idea of pretraining on SQuAD before being exposed to other kinds of question answering. This would teach the model how to use the multi-context decoder to properly retrieve information from the context before needing to learn how to switch between tasks or generate words on its own. Additionally, pretraining on SQuAD had already been shown to improve performance for NLI [Min et al., 2017]. Empirically, we found that this motivation is well-placed and that this strategy outperforms all others that we considered in terms of the decaScore. This strategy sacrificed performance on IWSLT but recovered the lost decaScore on other tasks, especially those which use pointers.

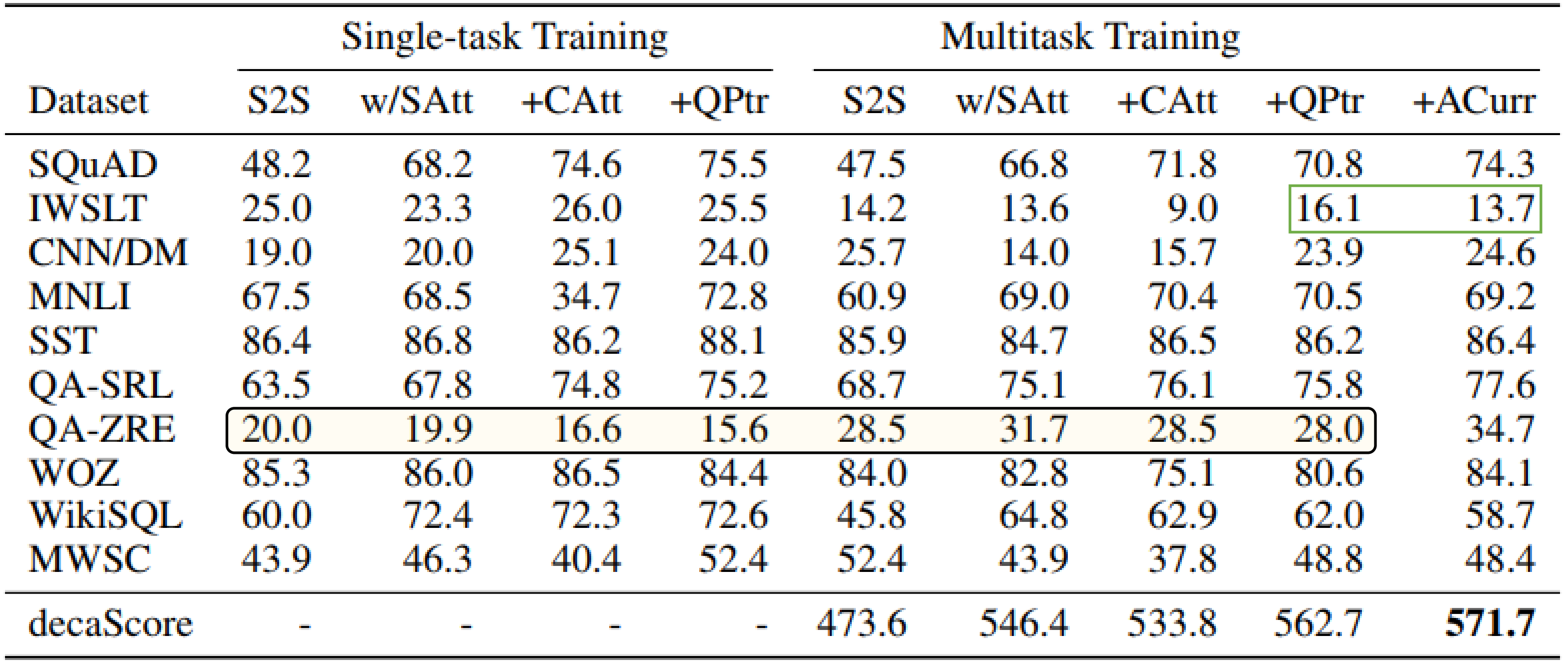
To explore if adding additional tasks to the initial curriculum would improve performance further, we experimented with adding IWSLT and CNN/DM to the first phase and in another experiment, adding IWSLT, CNN/DM and MNLI. These are tasks with a large number of training examples relative to the other tasks, and they contain the longest answer sequences. Further, they form a diverse set since they encourage the model to decode in different ways such as the vocabulary for IWSLT, context-pointer for SQuAD and CNN/DM, and question-pointer for MNLI. In our results, we however found no improvement by adding these tasks. In fact, in the case when we added SQuAD, IWSLT, CNN/DM and MNLI to the initial curriculum, we observed a marked degradation in performance of some other tasks including QA-SRL, WikiSQL and MWSC. This suggests that it is concordance between the question answering nature of the task and SQuAD that enabled improved outcomes and not necessarily the richness of the task.

Table 3: Validation metrics for MQAN using various training strategies. The first is fully joint, which samples batches round-robin from all tasks. Others first use a curriculum or anti-curriculum schedule over a subset of tasks before switching to fully joint over all tasks. Curriculum first trains tasks that take relatively few iterations to converge when trained alone. This omits SQuAD, IWSLT, CNN/DM, and MNLI. The remaining strategies are anti-curriculum. They include in the first phase either SQuAD alone, SQuAD, IWSLT, and CNN/DM, or SQuAD, IWSLT, CNN/DM, and MNLI.

			Anti-Curriculum		
Dataset	Fully Joint	Curriculum	SQuAD	+IWSLT+CNN/DM	+MNLI
SQuAD	70.8	43.4	74.3	74.5	74.6
IWSLT	16.1	4.3	13.7	18.7	19.0
CNN/DM	23.9	21.3	24.6	20.8	21.6
MNLI	70.5	58.9	69.2	69.6	72.7
SST	86.2	84.5	86.4	83.6	86.8
QA-SRL	75.8	70.6	77.6	77.5	75.1
QA-ZRE	28.0	24.6	34.7	30.1	37.7
WOZ	80.6	81.9	84.1	81.7	85.6
WikiSQL	62.0	68.6	58.7	54.8	42.6
MWSC	48.8	41.5	48.4	34.9	41.5
decaScore	562.7	499.6	571.7	546.2	557.2

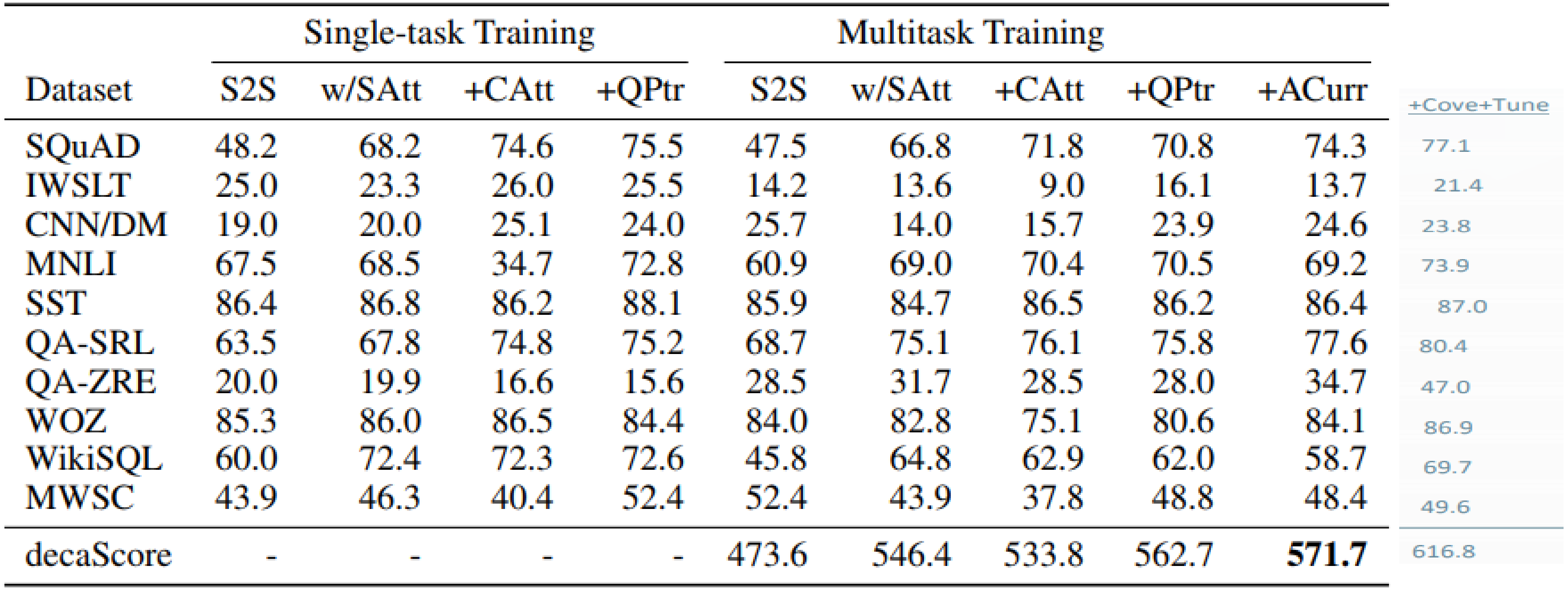
22

Experiments and Analysis



	Single-task Training				Multitask Training				
Dataset	S2S	w/SAtt	+CAtt	+QPtr	S2S	w/SAtt	+CAtt	+QPtr	+ACurr
SQuAD	48.2	68.2	74.6	75.5	47.5	66.8	71.8	70.8	74.3
IWSLT	25.0	23.3	26.0	25.5	14.2	13.6	9.0	16.1	13.7
CNN/DM	19.0	20.0	25.1	24.0	25.7	14.0	15.7	23.9	24.6
MNLI	67.5	68.5	34.7	72.8	60.9	69.0	70.4	70.5	69.2
SST	86.4	86.8	86.2	88.1	85.9	84.7	86.5	86.2	86.4
QA-SRL	63.5	67.8	74.8	75.2	68.7	75.1	76.1	75.8	77.6
QA-ZRE	20.0	19.9	16.6	15.6	28.5	31.7	28.5	28.0	34.7
WOZ	85.3	86.0	86.5	84.4	84.0	82.8	75.1	80.6	84.1
WikiSQL	60.0	72.4	72.3	72.6	45.8	64.8	62.9	62.0	58.7
MWSC	43.9	46.3	40.4	52.4	52.4	43.9	37.8	48.8	48.4
decaScore	-	-	-	-	473.6	546.4	533.8	562.7	**571.7**

Experiments and Analysis



	Single-task Training				Multitask Training					
Dataset	S2S	w/SAtt	+CAtt	+QPtr	S2S	w/SAtt	+CAtt	+QPtr	+ACurr	+Cove+Tune
SQuAD	48.2	68.2	74.6	75.5	47.5	66.8	71.8	70.8	74.3	77.1
IWSLT	25.0	23.3	26.0	25.5	14.2	13.6	9.0	16.1	13.7	21.4
CNN/DM	19.0	20.0	25.1	24.0	25.7	14.0	15.7	23.9	24.6	23.8
MNLI	67.5	68.5	34.7	72.8	60.9	69.0	70.4	70.5	69.2	73.9
SST	86.4	86.8	86.2	88.1	85.9	84.7	86.5	86.2	86.4	87.0
QA-SRL	63.5	67.8	74.8	75.2	68.7	75.1	76.1	75.8	77.6	80.4
QA-ZRE	20.0	19.9	16.6	15.6	28.5	31.7	28.5	28.0	34.7	47.0
WOZ	85.3	86.0	86.5	84.4	84.0	82.8	75.1	80.6	84.1	86.9
WikiSQL	60.0	72.4	72.3	72.6	45.8	64.8	62.9	62.0	58.7	69.7
MWSC	43.9	46.3	40.4	52.4	52.4	43.9	37.8	48.8	48.4	49.6
decaScore	-	-	-	-	473.6	546.4	533.8	562.7	**571.7**	616.8

Experiments and Analysis

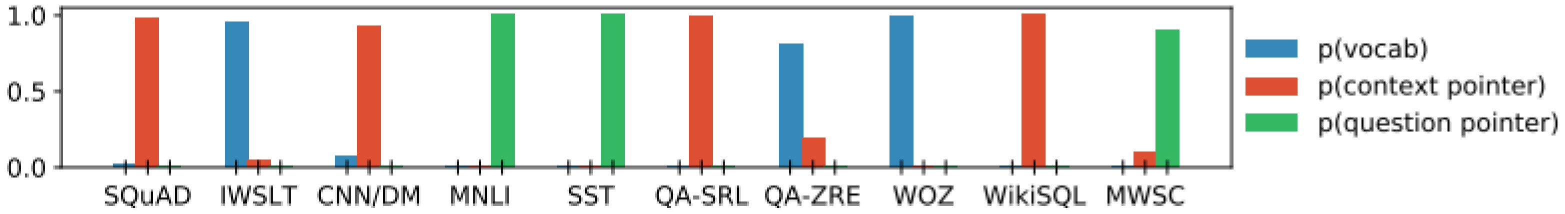


Figure 3: An analysis of how the MQAN chooses to output answer words. When p(generation) is highest, the MQAN places the most weight on the external vocab. When p(context) is highest, the MQAN places the most weight on the pointer distribution over the context. When p(question) is highest, the MQAN places the most weight on the pointer distribution over the question.

Task별 사용한 pointer가 달랐음!

Experiments and Analysis

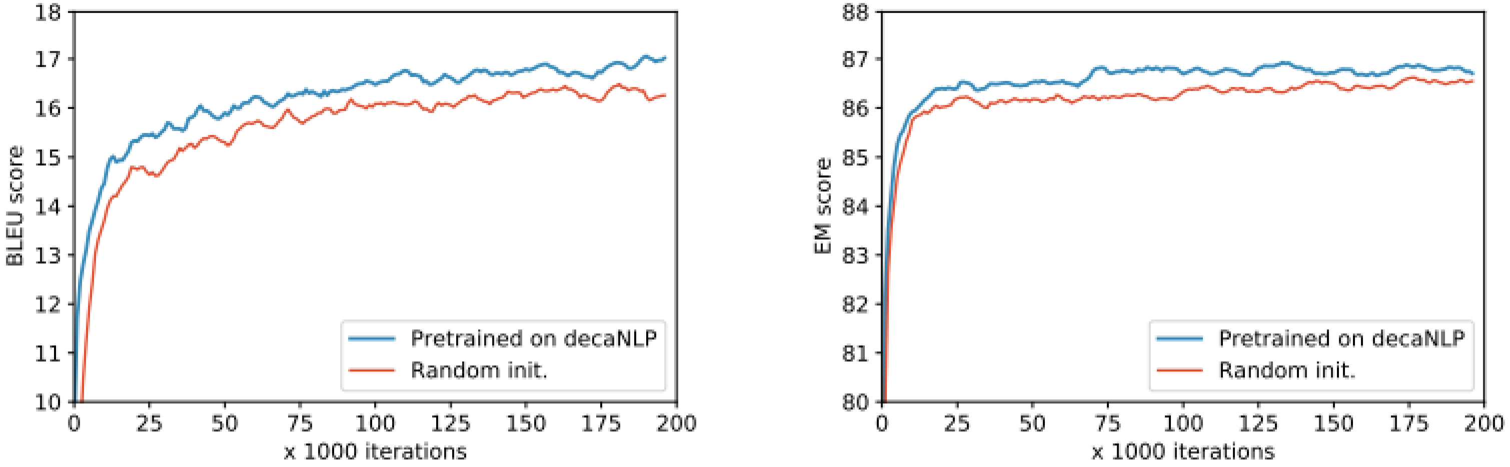


Figure 4: MQAN pretrained on decaNLP outperforms random initialization when adapting to new domains and learning new tasks. Left: training on a new language pair – English to Czech, right: training on a new task – Named Entity Recognition (NER).

MTL → 말 그대로의 transfer learning, domain adaptation, zero-shot 가능!

Experiments and Analysis

Basic Meta Learning

meta-learning: $\theta^\star = \arg\max_\theta \log p(\theta|\mathcal{D}_{\text{meta-train}})$

adaptation: $\phi^\star = \arg\max_\phi \log p(\phi|\mathcal{D}, \theta^\star)$

multi-task learning:

$$\theta^\star = \arg\max_\theta \sum_{i=1}^{n} \log p(\boxed{\theta}|\mathcal{D}_i)$$

learn θ such that $\phi_i = f_\theta(\mathcal{D}_i^{\text{tr}})$ is good for $\mathcal{D}_i^{\text{ts}}$

$$\theta^\star = \arg\max_\theta \sum_{i=1}^{n} \log p(\boxed{\phi_i}|\mathcal{D}_i^{\text{ts}})$$

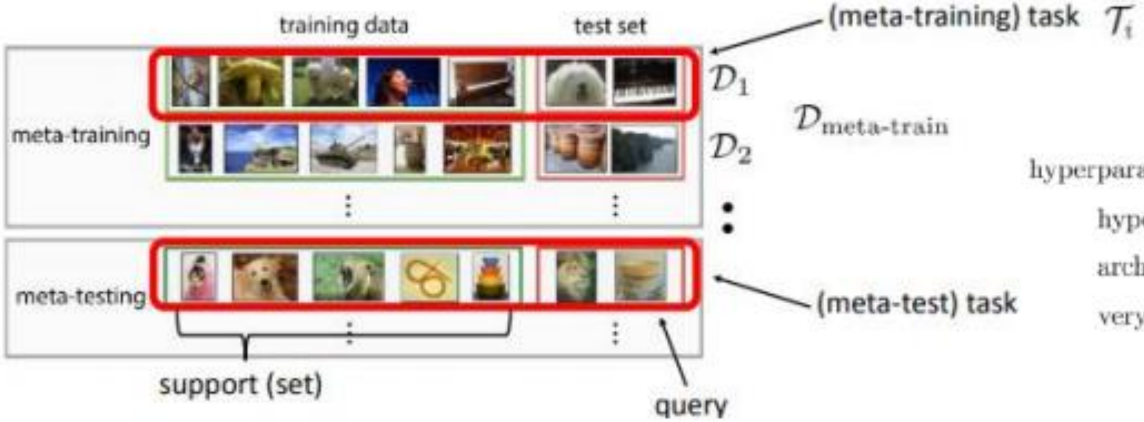
where $\phi_i = f_\theta(\mathcal{D}_i^{\text{tr}})$

$$\mathcal{D}_{\text{meta-train}} = \{(\mathcal{D}_1^{\text{tr}}, \mathcal{D}_1^{\text{ts}}), \ldots, (\mathcal{D}_n^{\text{tr}}, \mathcal{D}_n^{\text{ts}})\}$$

$$\mathcal{T}_i \begin{cases} \mathcal{D}_i^{\text{tr}} = \{(x_1^i, y_1^i), \ldots, (x_k^i, y_k^i)\} \\ \mathcal{D}_i^{\text{ts}} = \{(x_1^i, y_1^i), \ldots, (x_l^i, y_l^i)\} \end{cases}$$

shot

(i.e., k-shot, 5-shot)



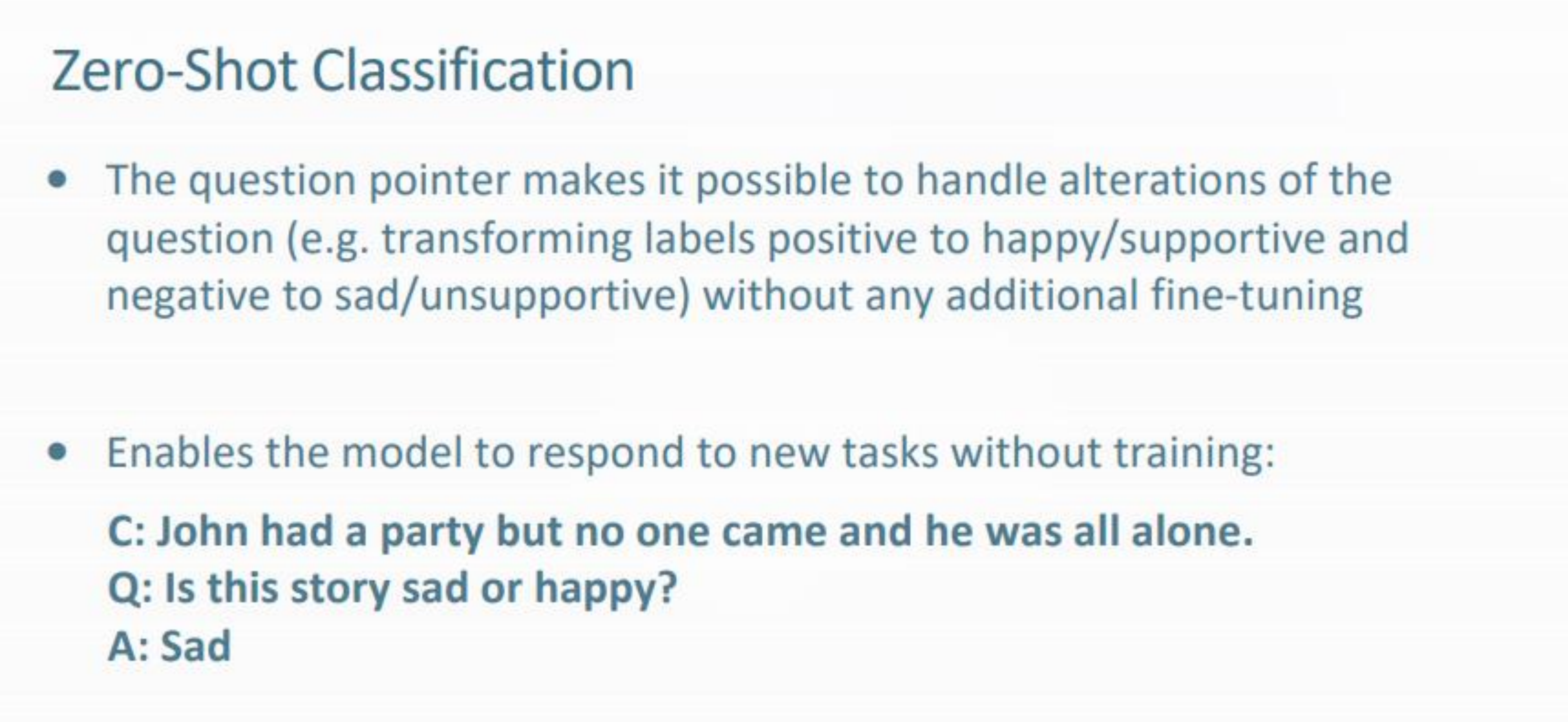
hyperparameter optimization & auto-ML: can be cast as meta-learning

hyperparameter optimization: θ = hyperparameters, ϕ = network weights

architecture search: θ = architecture, ϕ = network weights

very active area of research! but outside the scope of this course

Experiments and Analysis



Zero-Shot Classification

- The question pointer makes it possible to handle alterations of the question (e.g. transforming labels positive to happy/supportive and negative to sad/unsupportive) without any additional fine-tuning
- Enables the model to respond to new tasks without training:

  C: John had a party but no one came and he was all alone.
  Q: Is this story sad or happy?
  A: Sad