The Natural Language Decathlon: Multitask Learning as Question Answering

Richard Socher
Chief Scientist at Salesforce

Joint work with
Bryan McCann, Nitish Keskar and Caiming Xiong
Salesforce Research

What’s next for NLP & AI?

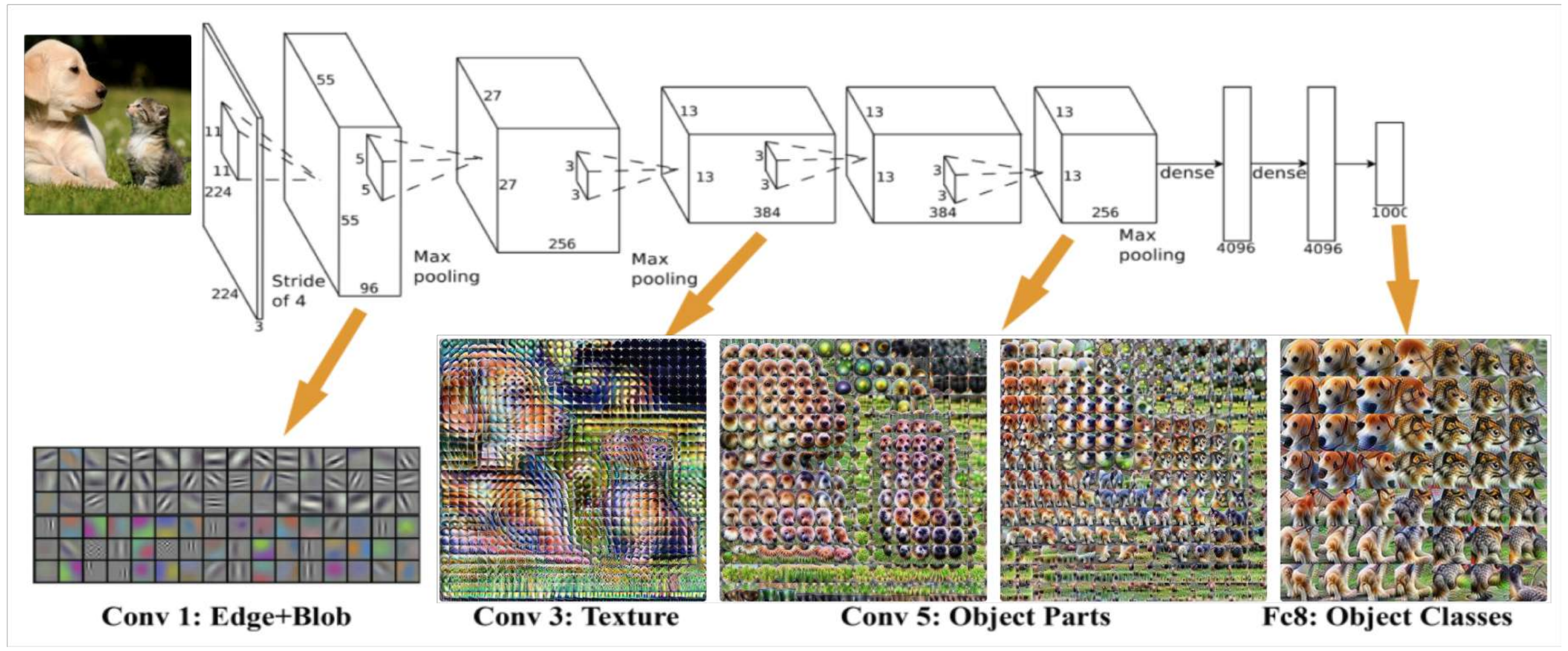
The Limits of Single-task Learning

- Great performance improvements in recent years given {dataset, task, model, metric}
- We can hill-climb to local optima as long as |dataset| > 1000xC
- For more general AI, we need continuous learning in a single model instead
- Models typically start from random or are only partly pre-trained → ☹

Pre-training and sharing knowledge is great!

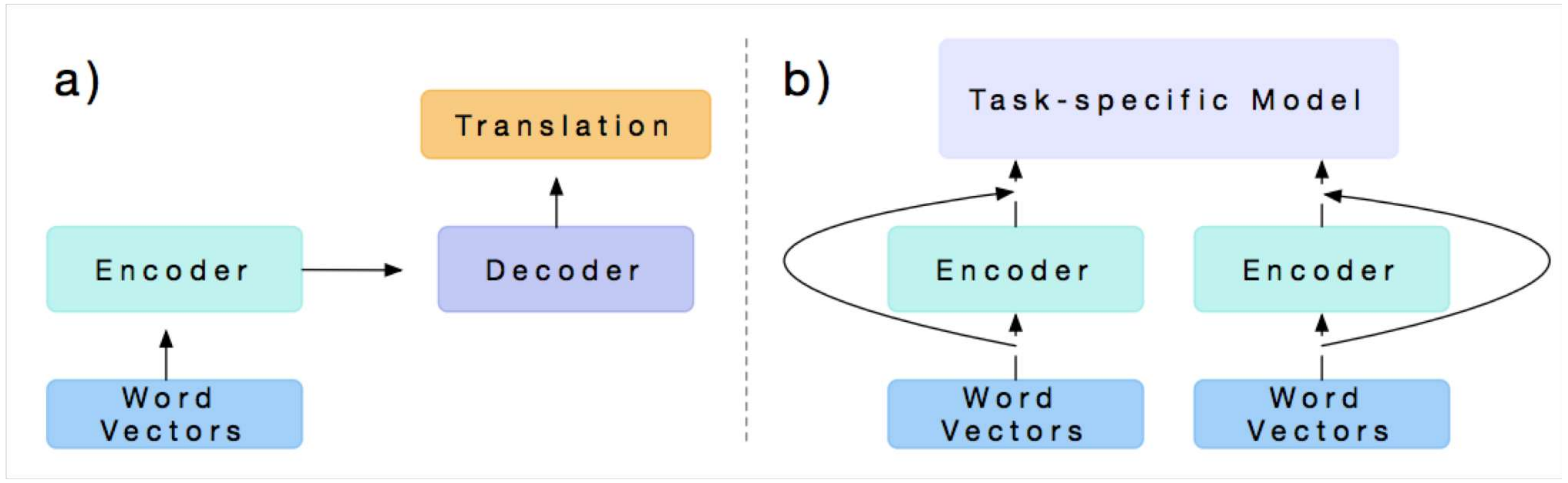
Computer Vision:

- ImageNet+CNN huge success →
- Classification was *the* blocking task in vision.



NLP:

- Word2Vec, GloVe, CoVe, ELMo, BERT
  → beginning success
- No single blocking task in natural language



Why has weight & model sharing not happened as much in NLP?

- NLP requires many types of reasoning:
  logical, linguistic, emotional, visual, ++
- Requires short and long term memory
- NLP had been divided into intermediate and separate tasks to make progress
  → Benchmark chasing in each community

- Can a single unsupervised task solve it all? No.
- Language clearly requires supervision in nature

Why a unified multi-task model for NLP?

- Multi-task learning is a blocker for general NLP systems

- Unified models can decide how to transfer knowledge (domain adaptation, weight sharing, transfer and zero shot learning)

- Unified, multi-task models can
  - More easily adapt to new tasks
  - Make deploying to production X times simpler
  - Lower the bar for more people to solve new tasks
  - Potentially move towards continual learning

How to express many NLP tasks in the same framework?

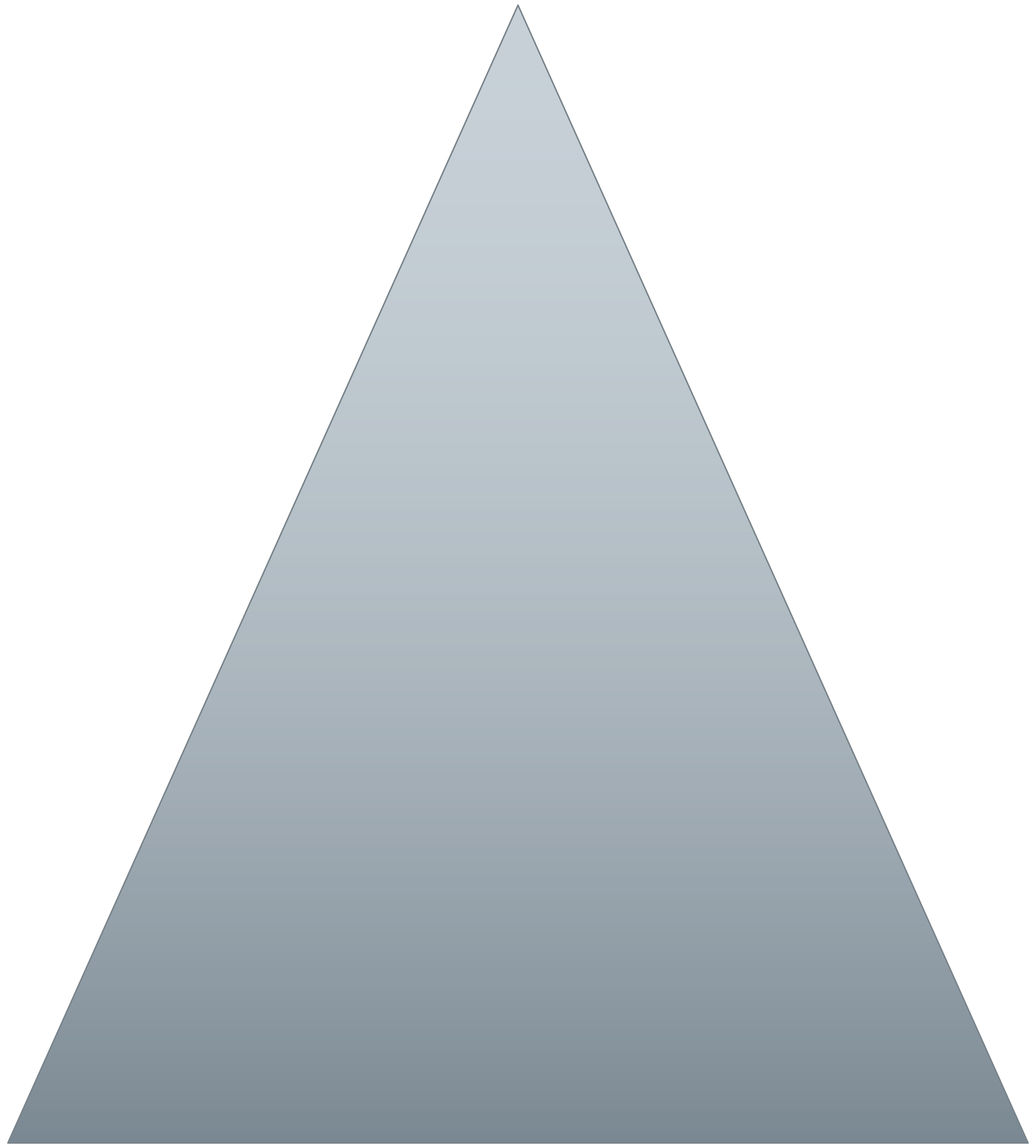
- Sequence tagging:
  named entity recognition, aspect specific sentiment
- Text classification:
  dialogue state tracking, sentiment classification
- Seq2seq:
  machine translation, summarization, question answering

3 equivalent Supertasks of NLP

Language Modeling

Question Answering

Dialogue



Usefulness and complexity
in their current interpretation

The Natural Language Decathlon (decaNLP)

Examples

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 Learning as Question Answering

- Question Answering
- Machine Translation
- Summarization
- Natural Language Inference
- Sentiment Classification
- Semantic Role Labeling
- Relation Extraction
- Dialogue
- Semantic Parsing
- Commonsense Reasoning

- Meta-Supervised learning: From {x, y} to {x, t, y} (t is the task)
- Use a question, q, as a natural description of the task, t, to allow the model to use linguistic information to connect tasks
- y is the answer to q and x is the context necessary to answer q

Designing a model for decaNLP

Specifications:

- No task-specific modules or parameters because we assume the task ID is not available

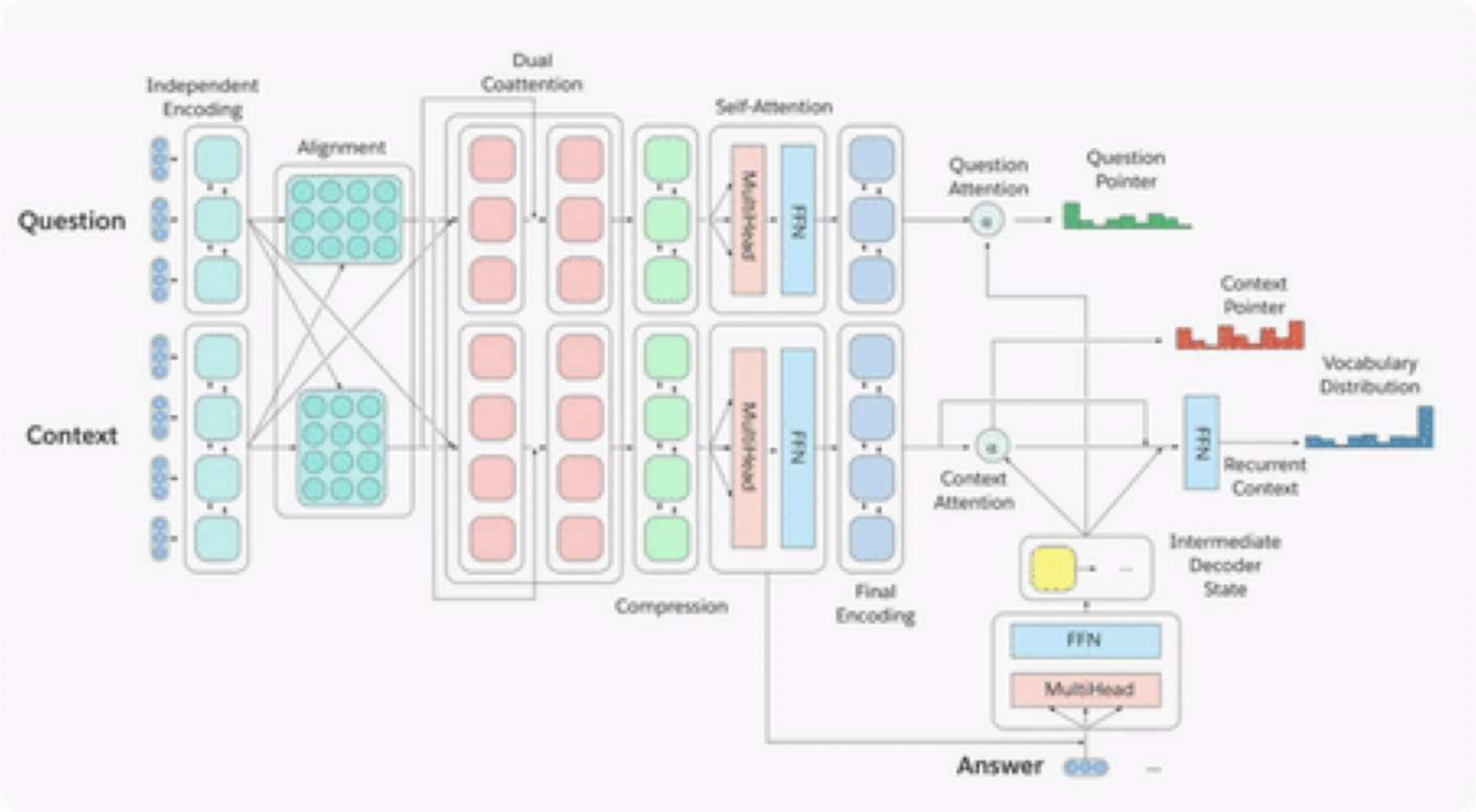
- Must be able to adjust internally to perform disparate tasks

- Should leave open the possibility of zero-shot inference for unseen tasks

A Multitask Question Answering Network for decaNLP

Context

Southern California, often abbreviated SoCal, is a geographic and cultural region that generally comprises California's southernmost 10 counties. The region is traditionally described as "eight counties", based on demographics and economic ties: Imperial, Los Angeles, Orange, Riverside, San Bernardino, San Diego, Santa Barbara, and Ventura. The more extensive 10-county definition, including Kern and San Luis Obispo counties, is also used based on historical political divisions. Southern California is a major economic center for the state of California and the United States.

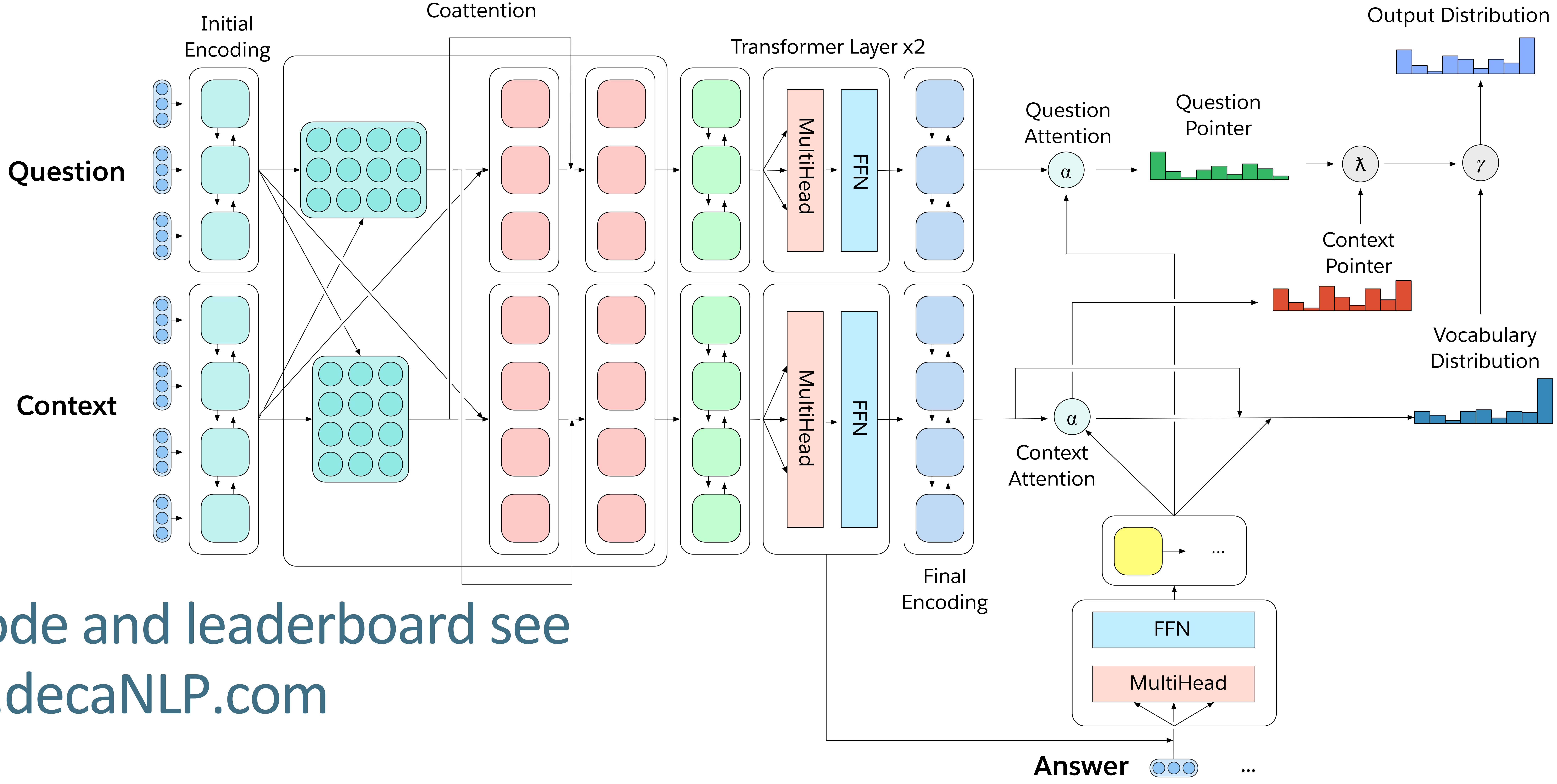


Task: Question Answering

- Start with a context
- Ask a question
- Generate the answer one word at a time by
  - Pointing to context
  - Pointing to question
  - Or choosing a word from an external vocabulary
- Pointer Switch is choosing between those three options for each output word

salesforce

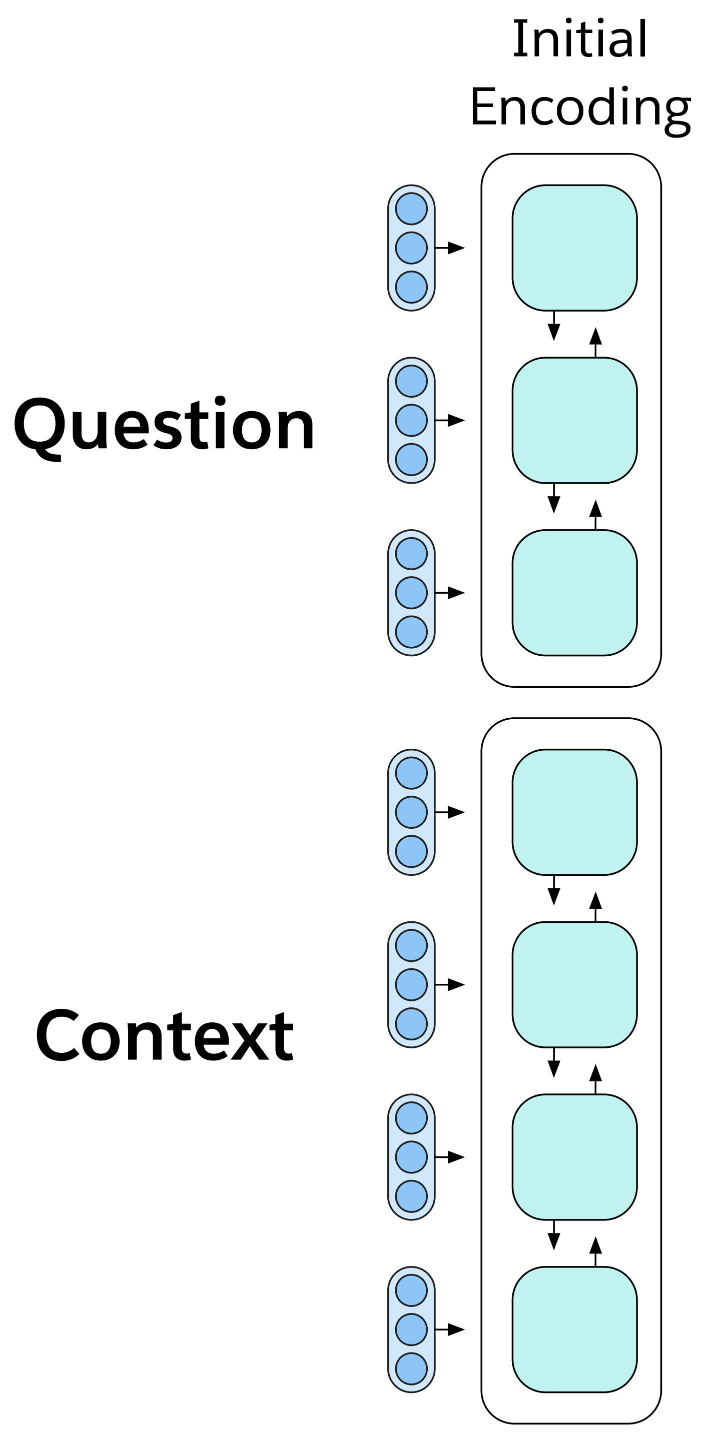
Multitask Question Answering Network (MQAN)



For code and leaderboard see www.decaNLP.com

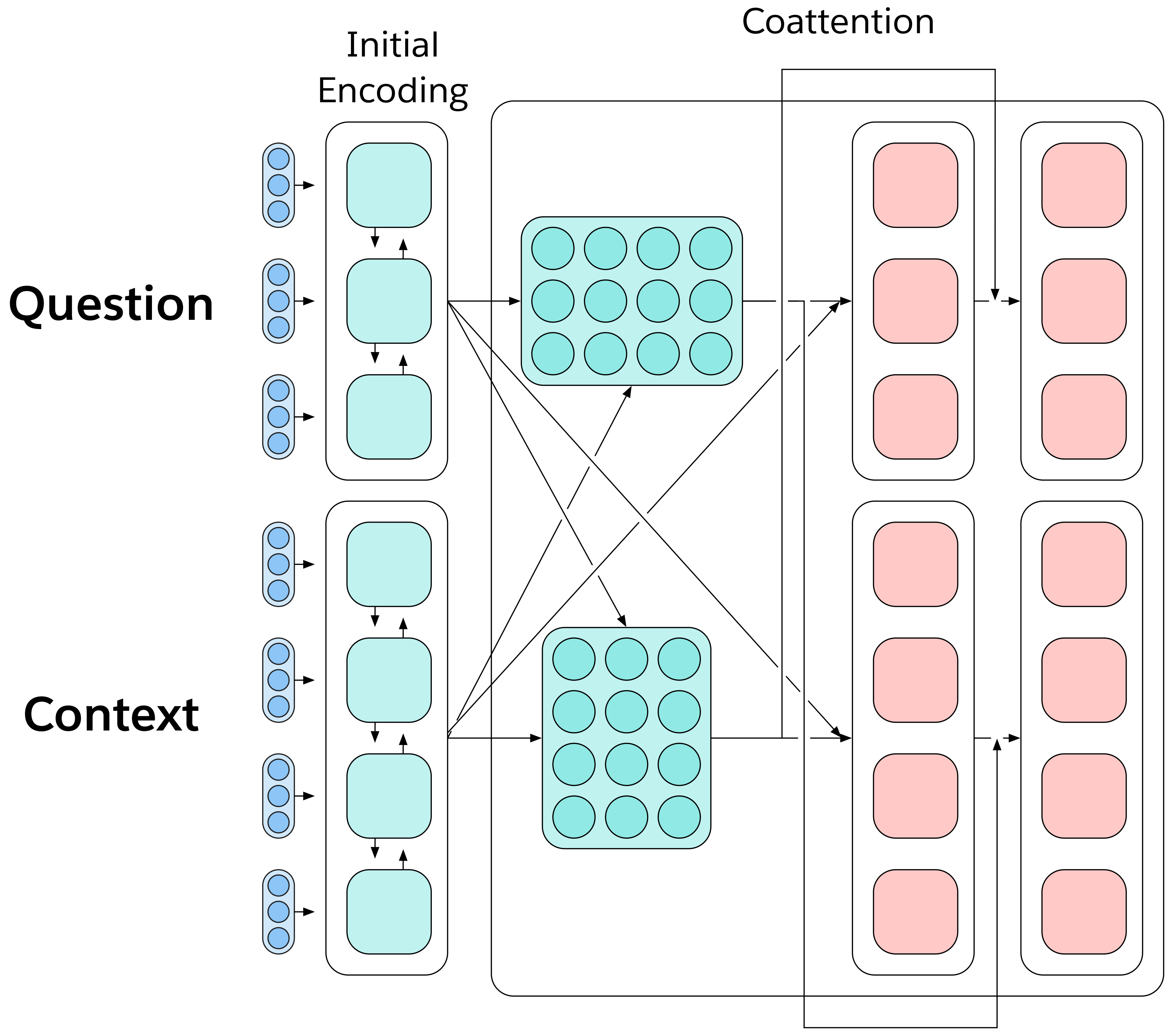
salesforce

Multitask Question Answering Network (MQAN)



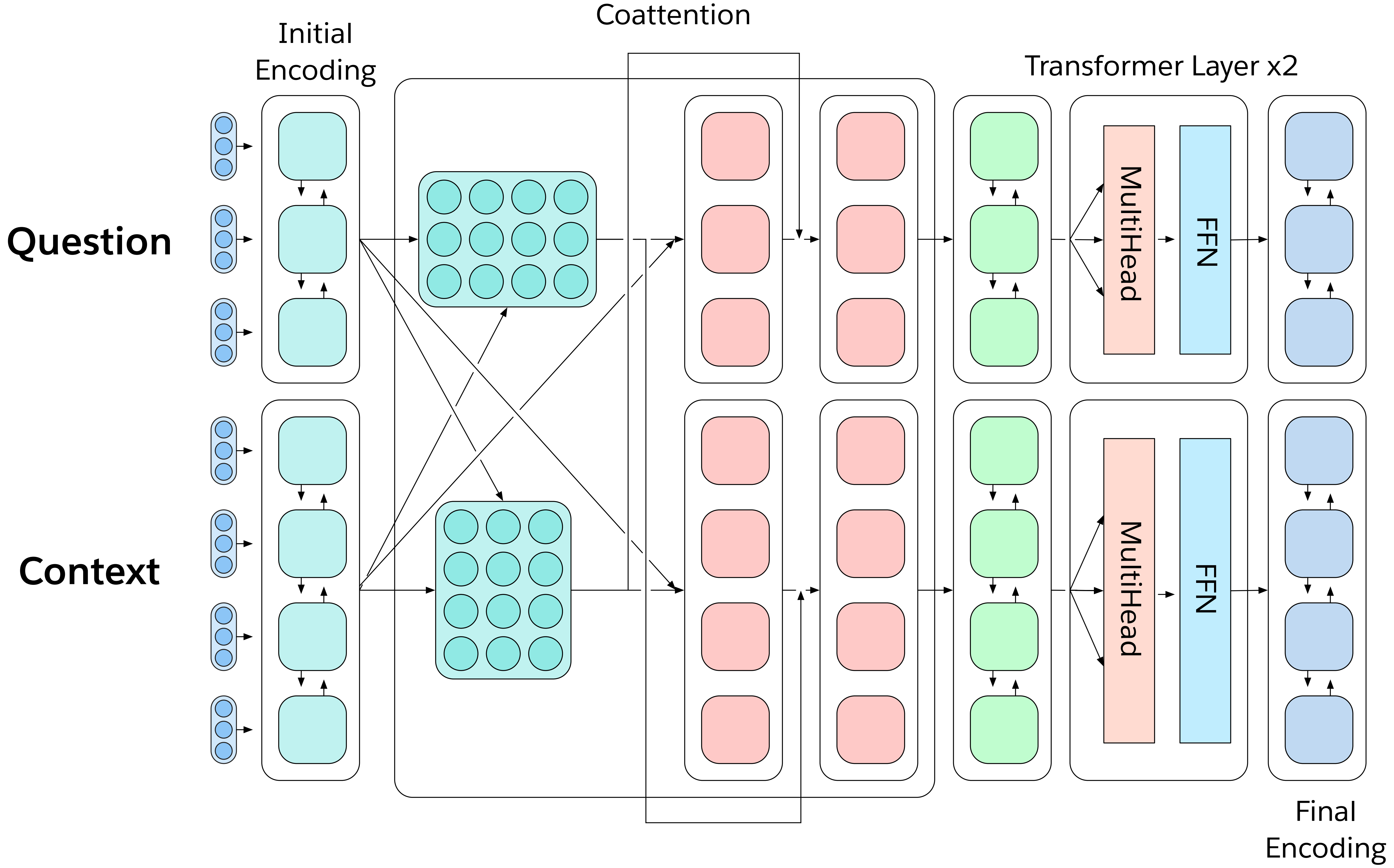
Fixed Glove+Character n-gram embeddings → Linear → Shared BiLSTM with skip connection

Multitask Question Answering Network (MQAN)



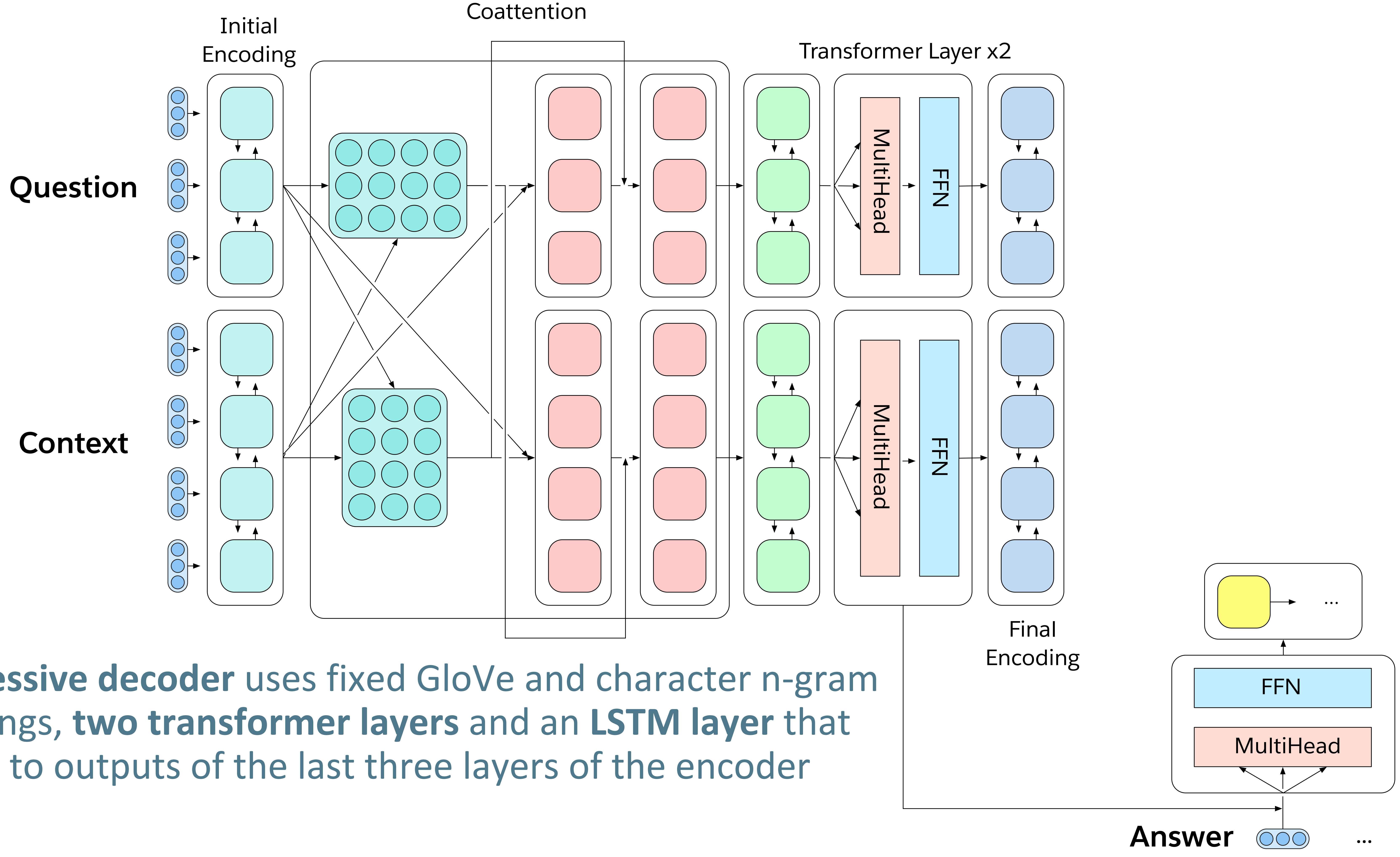
Attention summations from one sequence to the other and back again with skip connections

Multitask Question Answering Network (MQAN)



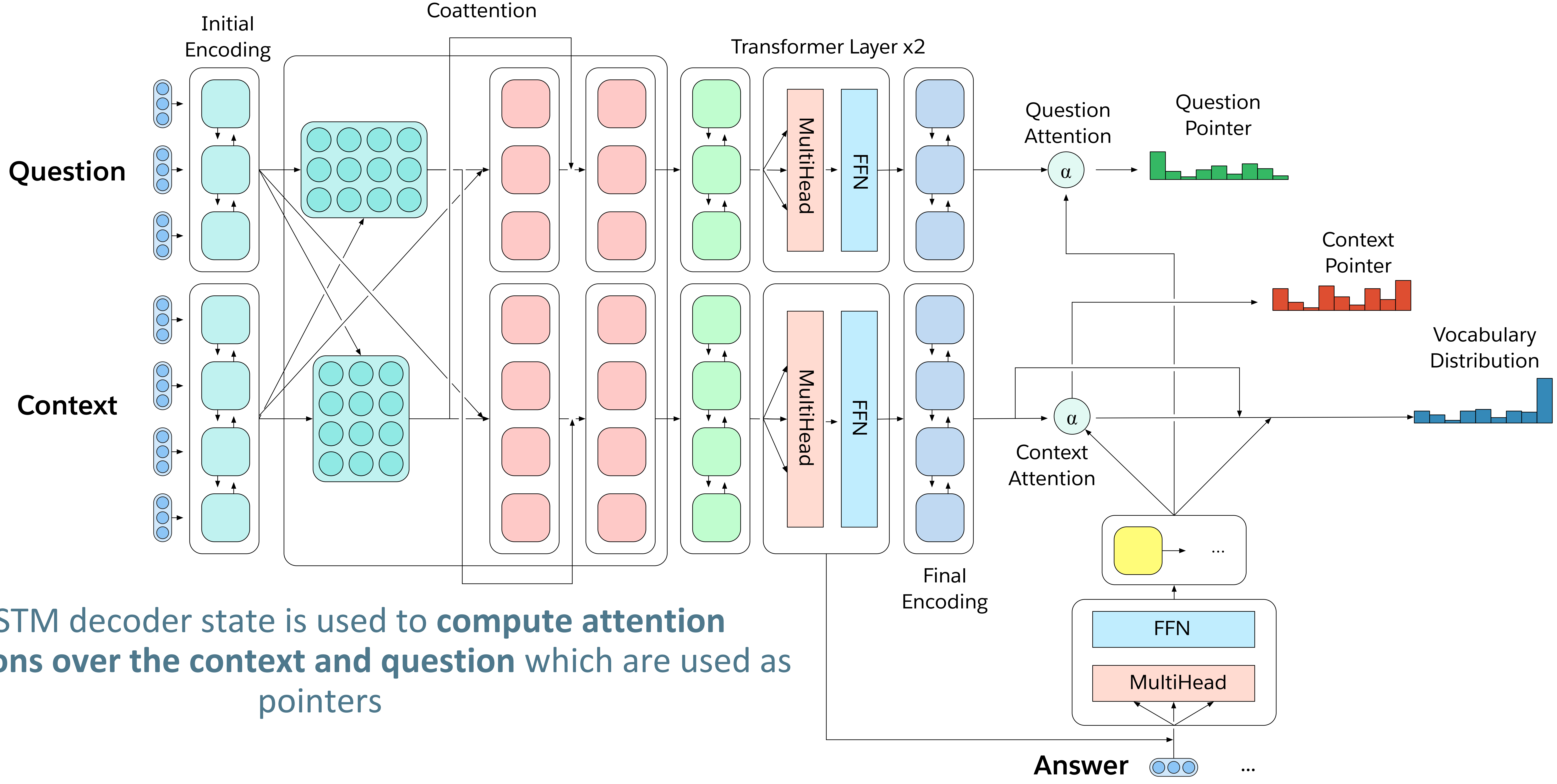
Separate BiLSTMs to reduce dimensionality, two transformer layers, another BiLSTM

Multitask Question Answering Network (MQAN)



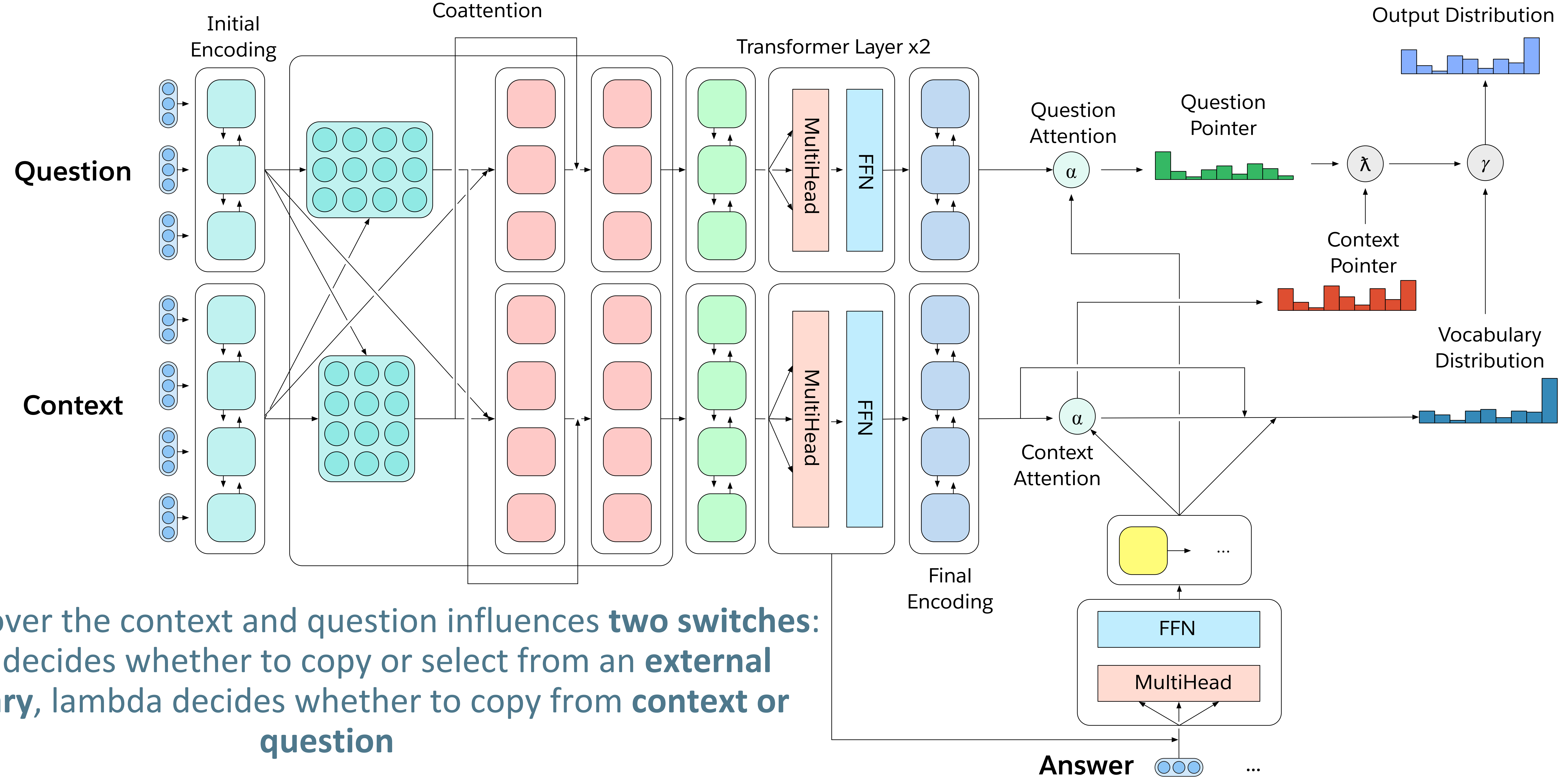
Auto-regressive decoder uses fixed GloVe and character n-gram embeddings, **two transformer layers** and an **LSTM layer** that attend to outputs of the last three layers of the encoder

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

Multitask Question Answering Network (MQAN)



Attention over the context and question influences **two switches**: gamma decides whether to copy or select from an **external vocabulary**, lambda decides whether to copy from **context or question**

Evaluation	Dataset	Metric
Question Answering	SQuAD	nF1
Machine Translation	IWSLT En — De	BLEU
Summarization	CNN/DailyMail	ROUGE
Natural Language Inference	MultiNLI	EM
Sentiment Analysis	SST2	EM
Semantic Role Labeling	QA-SRL	nF1
Relation Extraction	QA-ZRE	cF1
Goal-Oriented Dialogue	WOZ	dsEM
Semantic Parsing	WikiSQL	lfEM
Pronoun Resolution	Winograd Schemas	EM

nF1 = normalized word-level F1
(case insensitive , no punctutation or articles)
ROUGE = average of ROUGE-1, 2, and L
EM = exact match

cF1 = corpus-level F1
(accounts for unanswerable questions)
dsEM = dialogue state EM
lfEM = logical form EM

Evaluation

Question Answering	SQuAD	nF1
Machine Translation	IWSLT En — De	BLEU
Summarization	CNN/DailyMail	ROUGE
Natural Language Inference	MultiNLI	EM
Sentiment Analysis	SST2	EM
Semantic Role Labeling	QA-SRL	nF1
Relation Extraction	QA-ZRE	cF1
Goal-Oriented Dialogue	WOZ	dsEM
Semantic Parsing	WikiSQL	lfEM
Pronoun Resolution	Winograd Schemas	EM
Natural Language Decathlon		decaScore

decaScore = sum of task-specific metrics

Dataset	Single-task Performance S2S	+SelfAtt	+CoAtt	+QPtr	Multitask Performance S2S	+SelfAtt	+CoAtt	+QPtr
SQuAD	48.2	68.2	74.6	75.5	47.5	66.8	71.8	70.8
IWSLT En — De	25.0	23.3	26.0	25.5	14.2	13.6	9.00	16.1
CNN/DailyMail	19.0	20.0	25.1	24.0	25.7	14.0	15.7	23.9
MultiNLI	67.5	68.5	34.7	72.8	60.9	69.0	70.4	70.5
SST2	86.4	86.8	86.2	88.1	85.9	84.7	86.5	86.2
QA-SRL	63.5	67.8	74.8	75.2	68.7	75.1	76.1	75.8
QA-ZRE	20.0	19.9	16.6	15.6	28.5	31.7	28.5	28.0
WOZ	85.3	86.0	86.5	84.4	84.0	82.8	75.1	80.6
WikiSQL	60.0	72.4	72.3	72.6	45.8	64.8	62.9	62.0
Winograd Schemas	43.9	46.3	40.4	52.4	52.4	43.9	37.8	48.8
decaScore					513.6	546.4	533.8	562.7

- S2S = Seq2Seq
- +SelfAtt = plus self attention
- +CoAtt = plus coattention
- +QPtr = plus question pointer == MQAN

	Single-task Performance				Multitask Performance			
Dataset	S2S	+SelfAtt	+CoAtt	+QPtr	S2S	+SelfAtt	+CoAtt	+QPtr
SQuAD	48.2	68.2	74.6	75.5	47.5	66.8	71.8	70.8
IWSLT En — De	25.0	23.3	26.0	25.5	14.2	13.6	9.00	16.1
CNN/DailyMail	19.0	20.0	25.1	24.0	25.7	14.0	15.7	23.9
MultiNLI	67.5	68.5	34.7	72.8	60.9	69.0	70.4	70.5
SST2	86.4	86.8	86.2	88.1	85.9	84.7	86.5	86.2
QA-SRL	63.5	67.8	74.8	75.2	68.7	75.1	76.1	75.8
QA-ZRE	20.0	19.9	16.6	15.6	28.5	31.7	28.5	28.0
WOZ	85.3	86.0	86.5	84.4	84.0	82.8	75.1	80.6
WikiSQL	60.0	72.4	72.3	72.6	45.8	64.8	62.9	62.0
Winograd Schemas	43.9	46.3	40.4	52.4	52.4	43.9	37.8	48.8
decaScore					513.6	546.4	533.8	562.7

- Transformer layers yield benefits in single-task and multitask setting

	Single-task Performance				Multitask Performance			
Dataset	S2S	+SelfAtt	+CoAtt	+QPtr	S2S	+SelfAtt	+CoAtt	+QPtr
SQuAD	48.2	68.2	74.6	75.5	47.5	66.8	71.8	70.8
IWSLT En — De	25.0	23.3	26.0	25.5	14.2	13.6	9.00	16.1
CNN/DailyMail	19.0	20.0	25.1	24.0	25.7	14.0	15.7	23.9
MultiNLI	67.5	68.5	34.7	72.8	60.9	69.0	70.4	70.5
SST2	86.4	86.8	86.2	88.1	85.9	84.7	86.5	86.2
QA-SRL	63.5	67.8	74.8	75.2	68.7	75.1	76.1	75.8
QA-ZRE	20.0	19.9	16.6	15.6	28.5	31.7	28.5	28.0
WOZ	85.3	86.0	86.5	84.4	84.0	82.8	75.1	80.6
WikiSQL	60.0	72.4	72.3	72.6	45.8	64.8	62.9	62.0
Winograd Schemas	43.9	46.3	40.4	52.4	52.4	43.9	37.8	48.8
decaScore					513.6	546.4	533.8	562.7

- Transformer layers yield benefits in single-task and multitask setting
- QA and SRL have a strong connection

Dataset	Single-task Performance S2S	+SelfAtt	+CoAtt	+QPtr	Multitask Performance S2S	+SelfAtt	+CoAtt	+QPtr
SQuAD	48.2	68.2	74.6	75.5	47.5	66.8	71.8	70.8
IWSLT En — De	25.0	23.3	26.0	25.5	14.2	13.6	9.00	16.1
CNN/DailyMail	19.0	20.0	25.1	24.0	25.7	14.0	15.7	23.9
MultiNLI	67.5	68.5	34.7	72.8	60.9	69.0	70.4	70.5
SST2	86.4	86.8	86.2	88.1	85.9	84.7	86.5	86.2
QA-SRL	63.5	67.8	74.8	75.2	68.7	75.1	76.1	75.8
QA-ZRE	20.0	19.9	16.6	15.6	28.5	31.7	28.5	28.0
WOZ	85.3	86.0	86.5	84.4	84.0	82.8	75.1	80.6
WikiSQL	60.0	72.4	72.3	72.6	45.8	64.8	62.9	62.0
Winograd Schemas	43.9	46.3	40.4	52.4	52.4	43.9	37.8	48.8
decaScore					513.6	546.4	533.8	562.7

- Transformer layers yield benefits in single-task and multitask setting
- QA and SRL have a strong connection
- Pointing to the question is essential

	Single-task Performance				Multitask Performance			
Dataset	S2S	+SelfAtt	+CoAtt	+QPtr	S2S	+SelfAtt	+CoAtt	+QPtr
SQuAD	48.2	68.2	74.6	75.5	47.5	66.8	71.8	70.8
IWSLT En — De	25.0	23.3	26.0	25.5	14.2	13.6	9.00	16.1
CNN/DailyMail	19.0	20.0	25.1	24.0	25.7	14.0	15.7	23.9
MultiNLI	67.5	68.5	34.7	72.8	60.9	69.0	70.4	70.5
SST2	86.4	86.8	86.2	88.1	85.9	84.7	86.5	86.2
QA-SRL	63.5	67.8	74.8	75.2	68.7	75.1	76.1	75.8
QA-ZRE	20.0	19.9	16.6	15.6	28.5	31.7	28.5	28.0
WOZ	85.3	86.0	86.5	84.4	84.0	82.8	75.1	80.6
WikiSQL	60.0	72.4	72.3	72.6	45.8	64.8	62.9	62.0
Winograd Schemas	43.9	46.3	40.4	52.4	52.4	43.9	37.8	48.8
decaScore					513.6	546.4	533.8	562.7

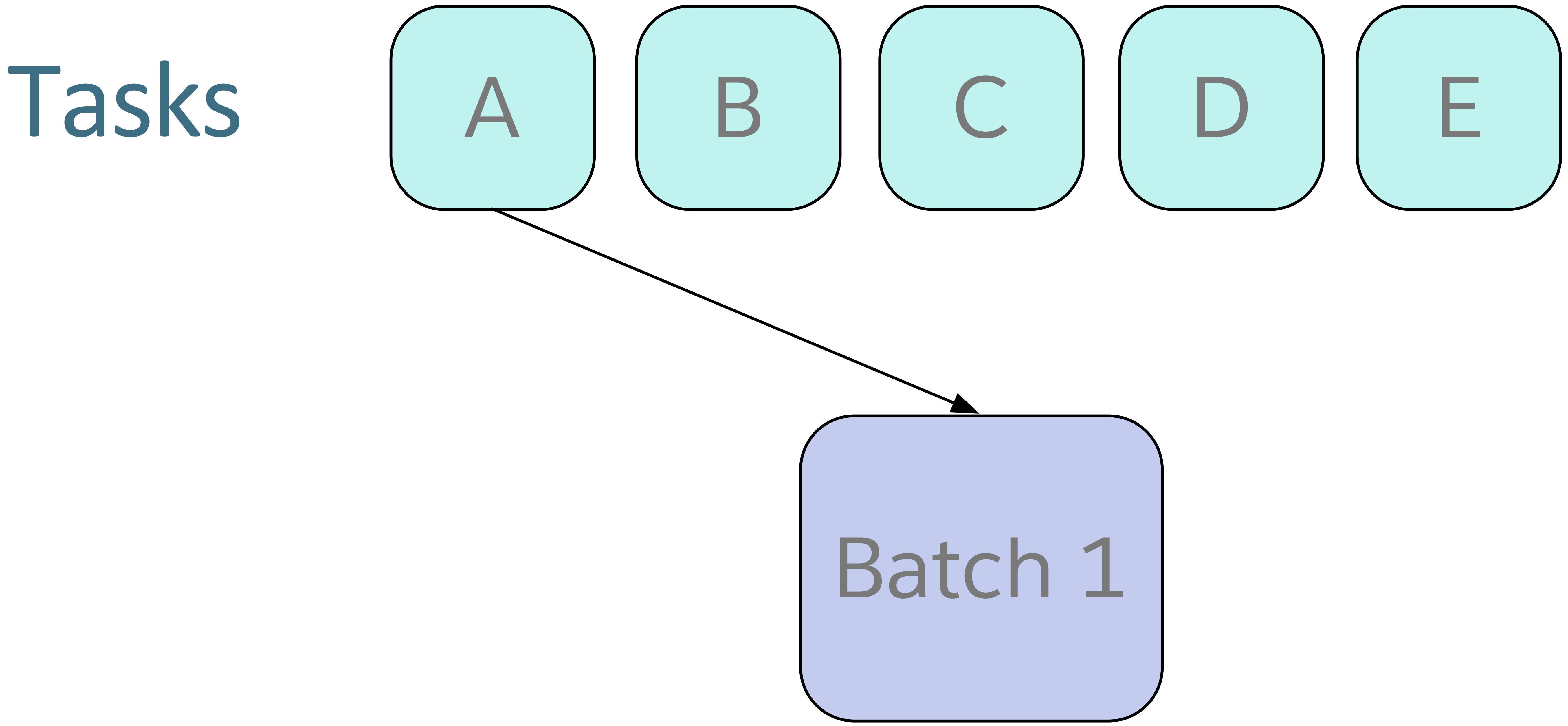
- Transformer layers yield benefits in single-task and multitask setting
- QA and SRL have a strong connection
- Pointing to the question is essential
- Multitasking helps zero-shot

	Single-task Performance				Multitask Performance			
Dataset	S2S	+SelfAtt	+CoAtt	+QPtr	S2S	+SelfAtt	+CoAtt	+QPtr
SQuAD	48.2	68.2	74.6	75.5	47.5	66.8	71.8	70.8
IWSLT En — De	25.0	23.3	26.0	25.5	14.2	13.6	9.00	16.1
CNN/DailyMail	19.0	20.0	25.1	24.0	25.7	14.0	15.7	23.9
MultiNLI	67.5	68.5	34.7	72.8	60.9	69.0	70.4	70.5
SST2	86.4	86.8	86.2	88.1	85.9	84.7	86.5	86.2
QA-SRL	63.5	67.8	74.8	75.2	68.7	75.1	76.1	75.8
QA-ZRE	20.0	19.9	16.6	15.6	28.5	31.7	28.5	28.0
WOZ	85.3	86.0	86.5	84.4	84.0	82.8	75.1	80.6
WikiSQL	60.0	72.4	72.3	72.6	45.8	64.8	62.9	62.0
Winograd Schemas	43.9	46.3	40.4	52.4	52.4	43.9	37.8	48.8
decaScore				(586.1)	513.6	546.4	533.8	562.7

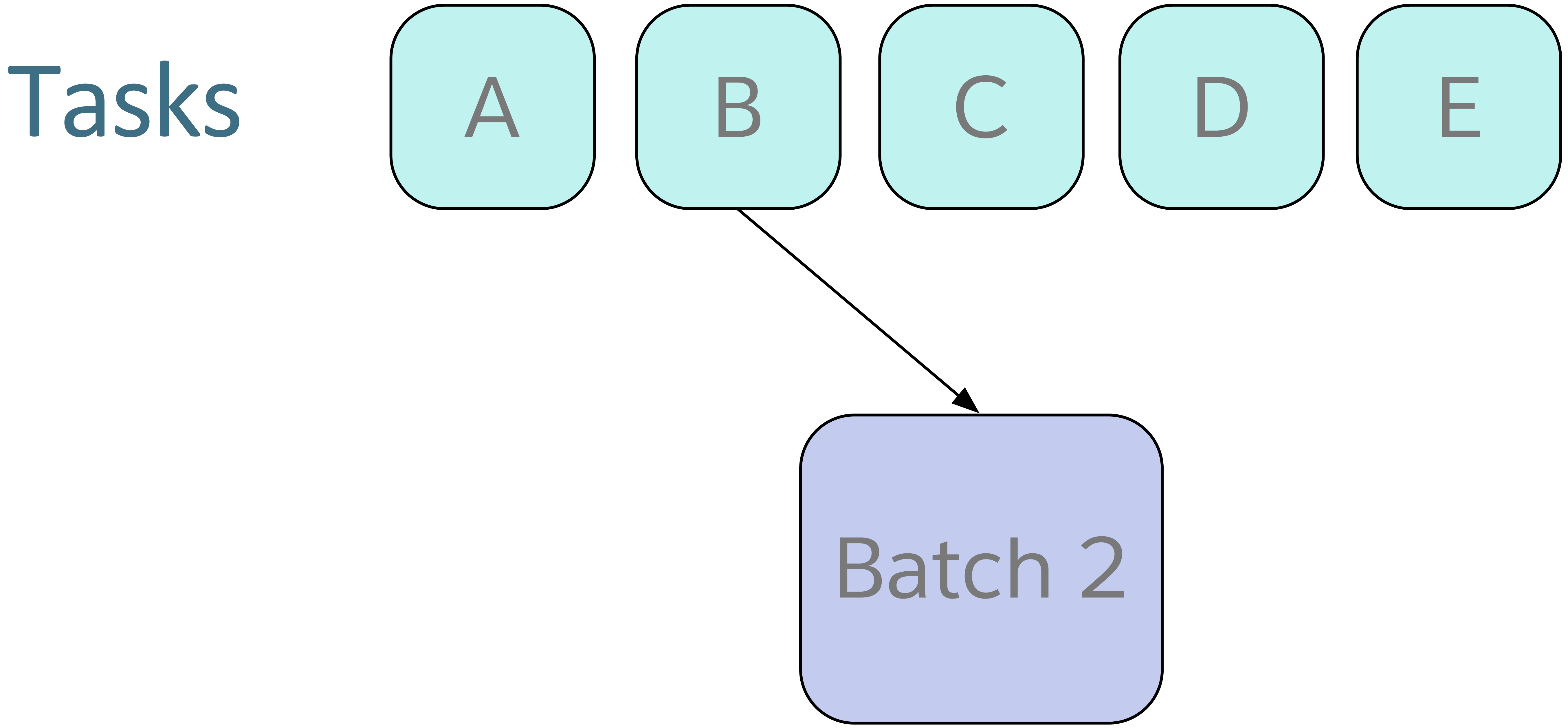
- Transformer layers yield benefits in single-task and multitask setting
- QA and SRL have a strong connection
- Pointing to the question is essential
- Multitasking helps zero-shot
- There is a gap between the combined single-task models and the single multitask model

salesforce

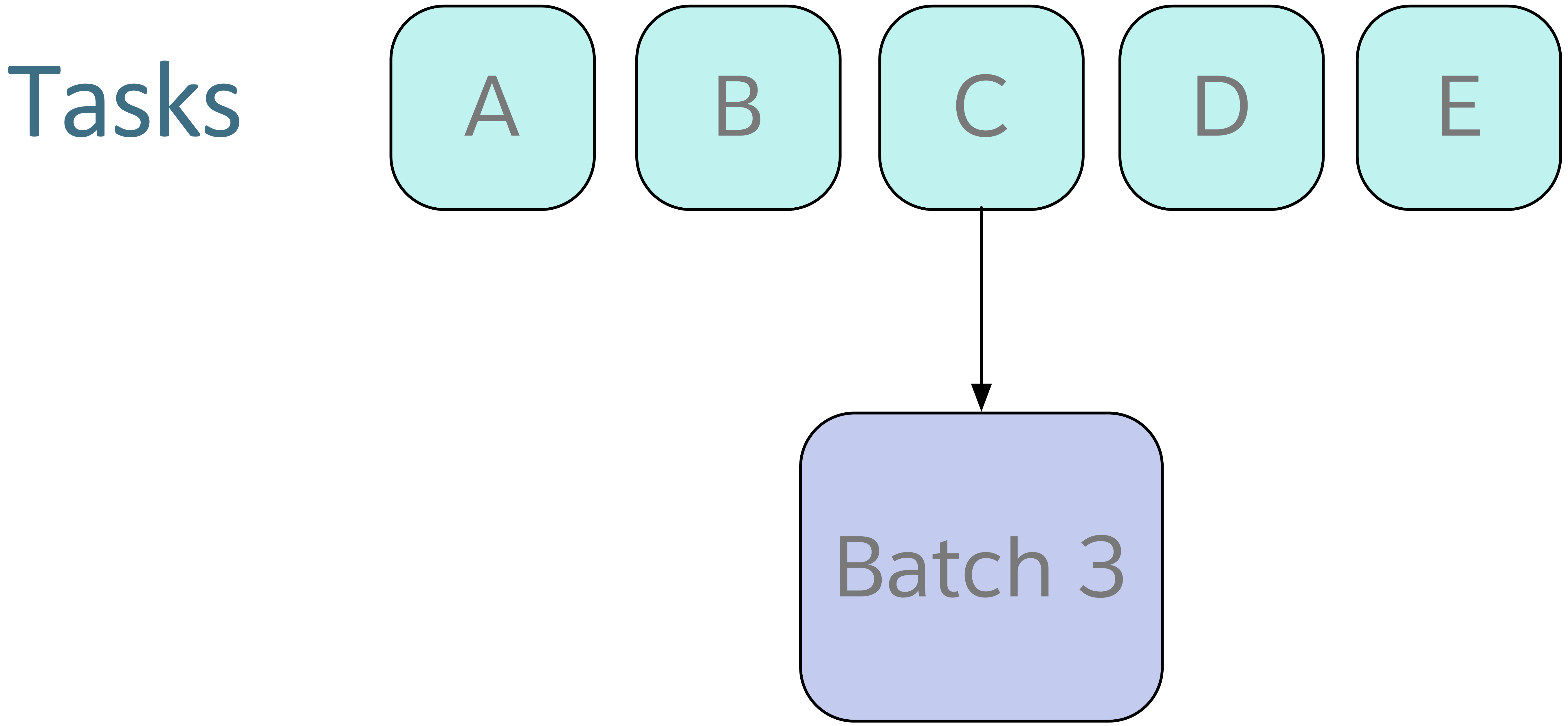
Training Strategies: Fully Joint



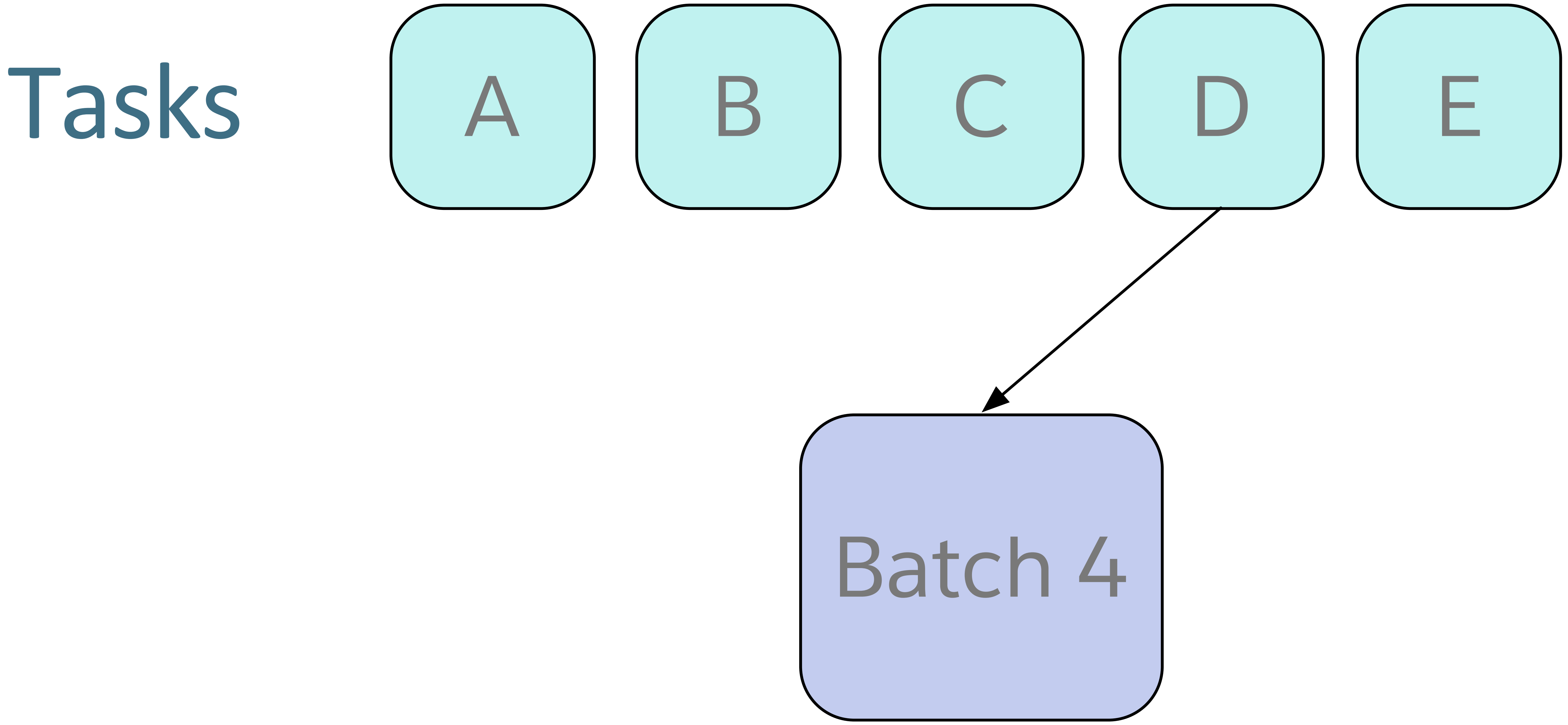
Training Strategies: Fully Joint



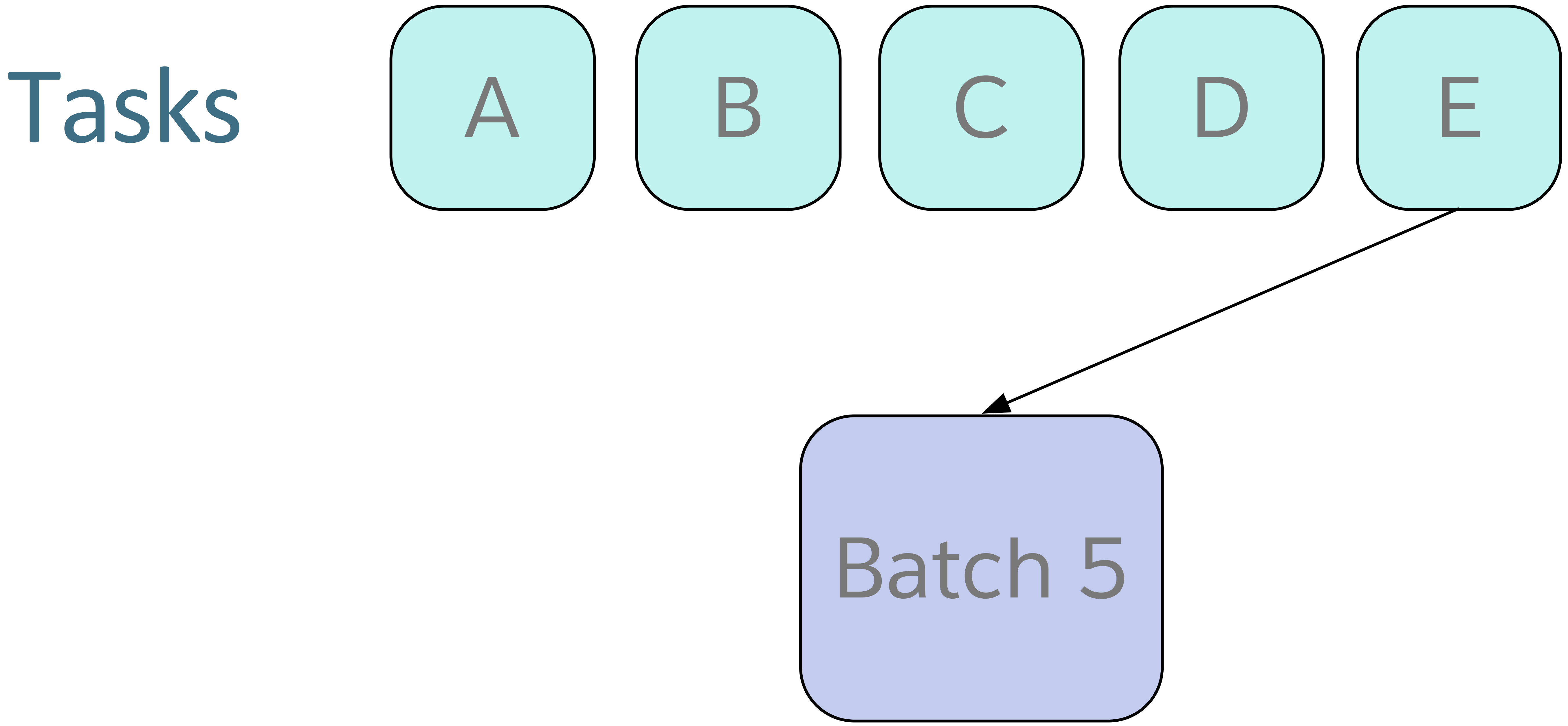
Training Strategies: Fully Joint



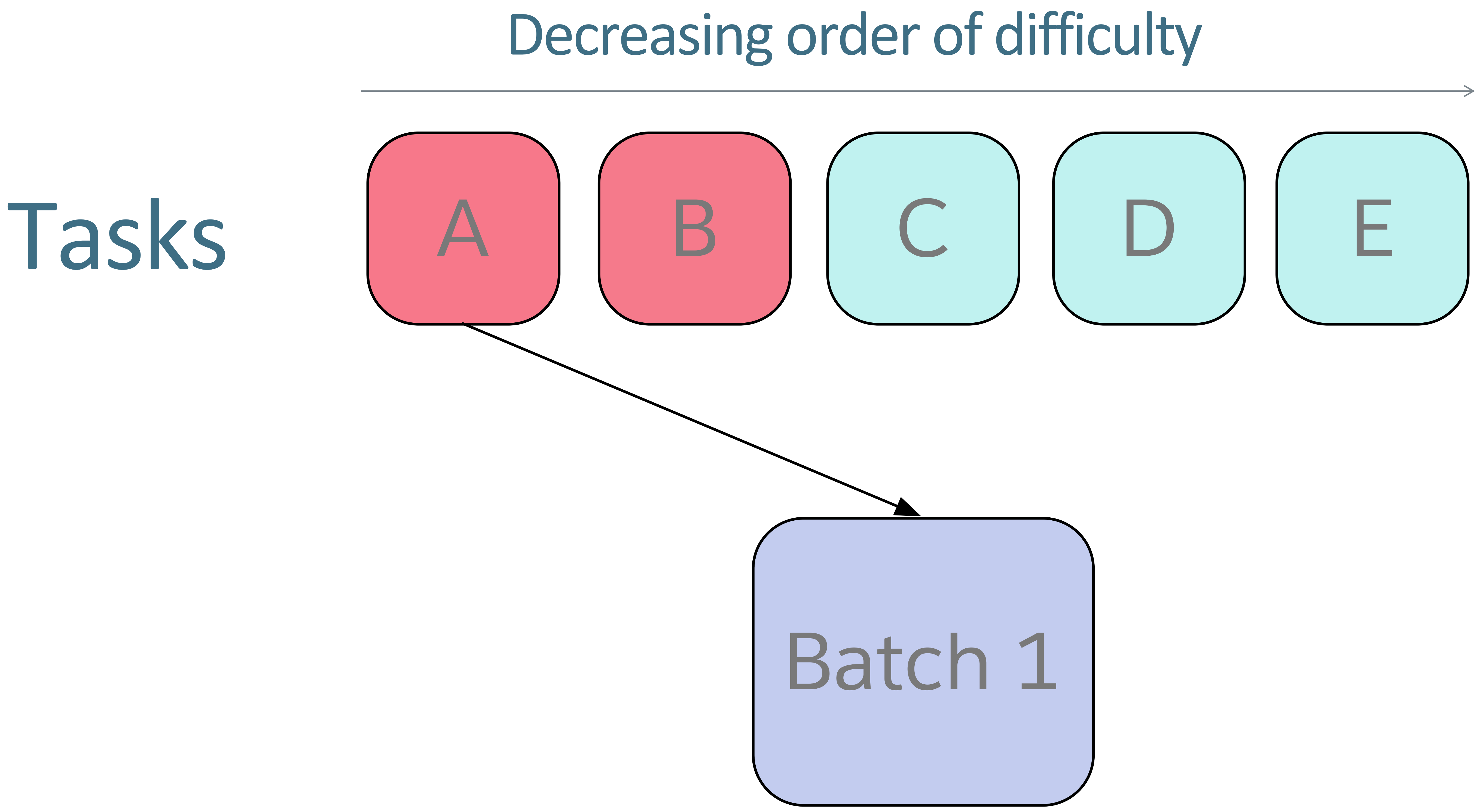
Training Strategies: Fully Joint



Training Strategies: Fully Joint



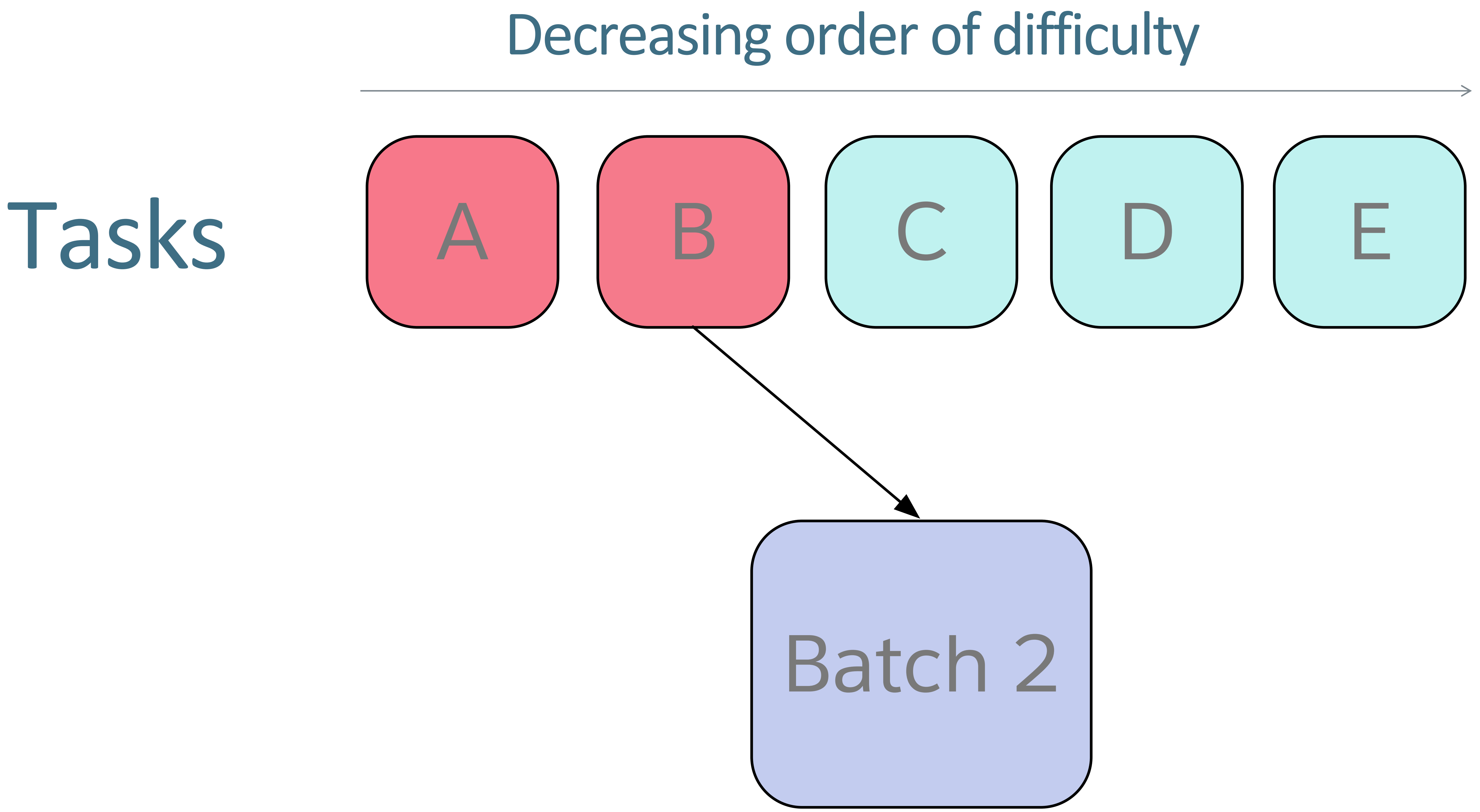
Training Strategies: Anti-Curriculum Pre-training



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

Reddish Tasks: tasks included in the pretraining phase

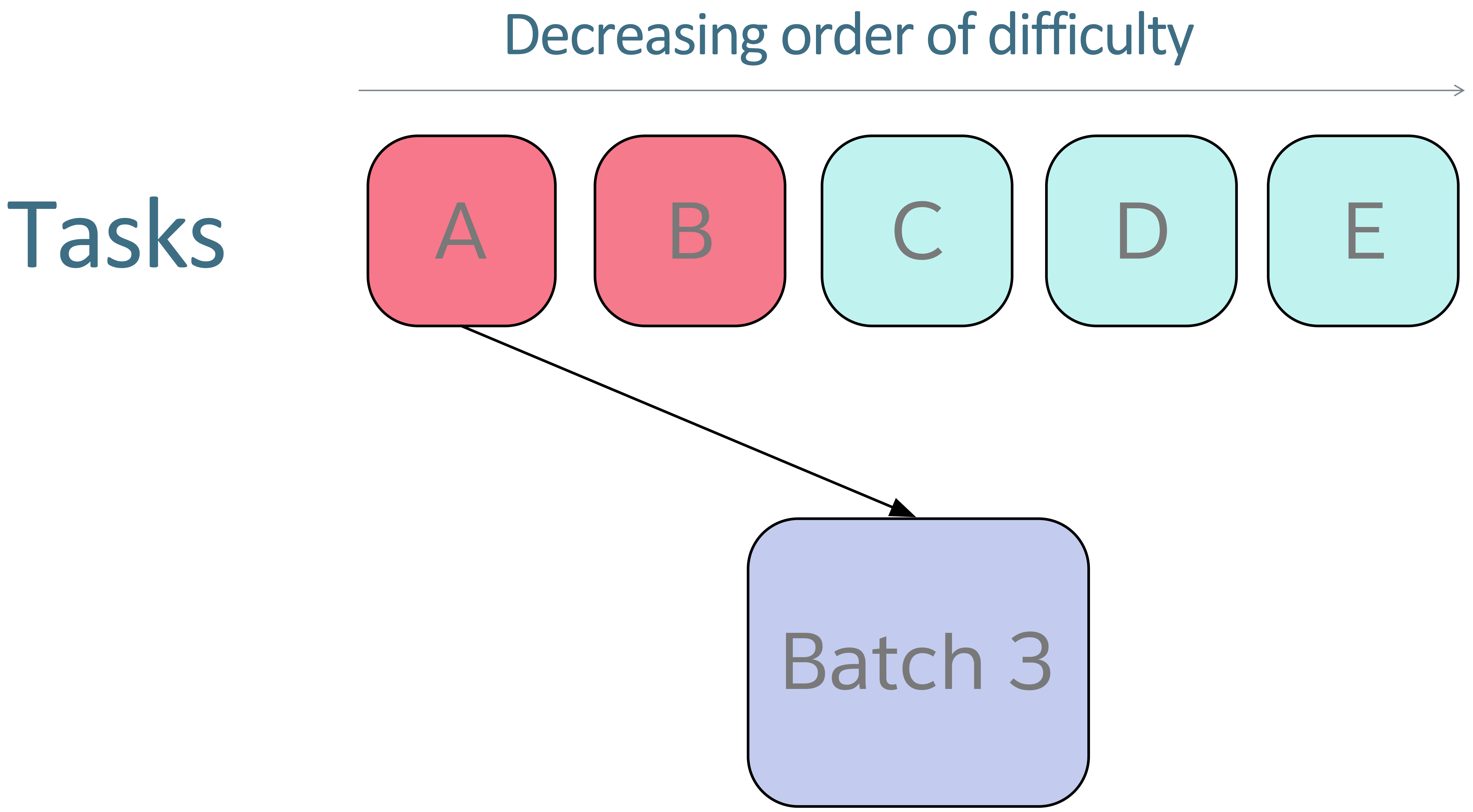
Training Strategies: Anti-Curriculum Pre-training



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

Reddish Tasks: tasks included in the pretraining phase

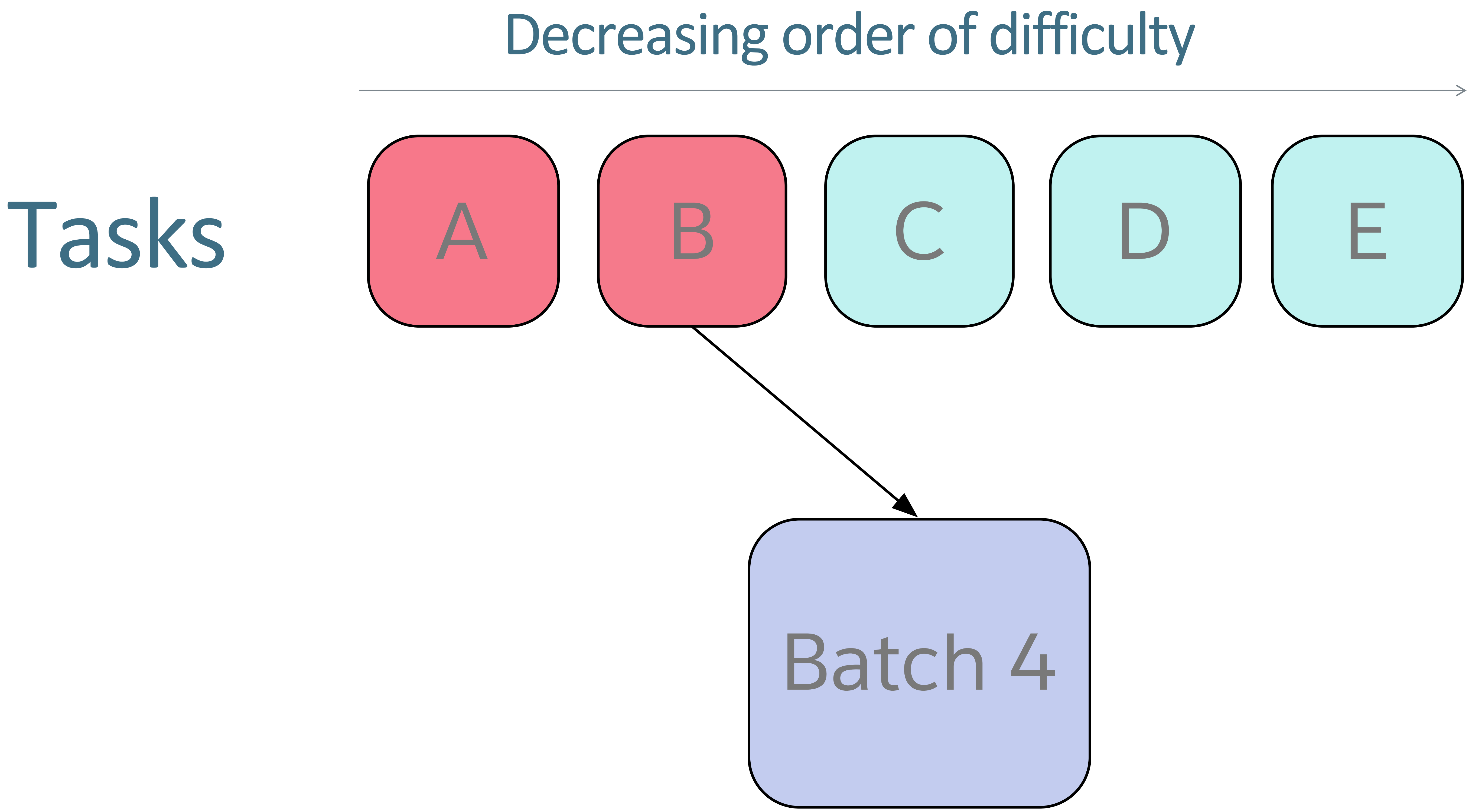
Training Strategies: Anti-Curriculum Pre-training



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

Reddish Tasks: tasks included in the pretraining phase

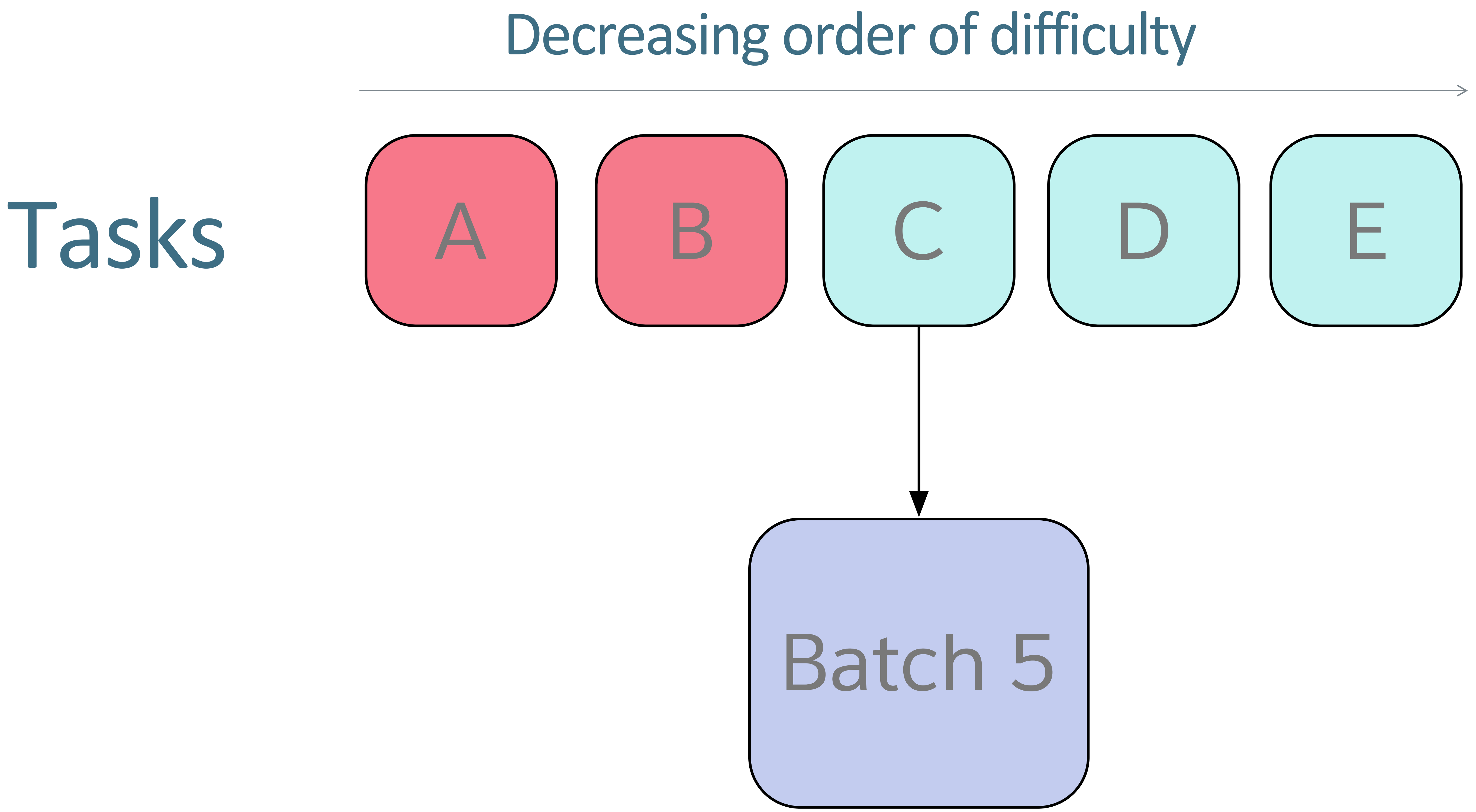
Training Strategies: Anti-Curriculum Pre-training



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

	Single-task Performance				Multitask Performance				
Dataset	S2S	+SelfAtt	+CoAtt	+QPtr	S2S	+SelfAtt	+CoAtt	+QPtr	+ACurr
SQuAD	48.2	68.2	74.6	75.5	47.5	66.8	71.8	70.8	74.3
IWSLT En — De	25.0	23.3	26.0	25.5	14.2	13.6	9.00	16.1	13.7
CNN/DailyMail	19.0	20.0	25.1	24.0	25.7	14.0	15.7	23.9	24.6
MultiNLI	67.5	68.5	34.7	72.8	60.9	69.0	70.4	70.5	69.2
SST2	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
Winograd Schemas	43.9	46.3	40.4	52.4	52.4	43.9	37.8	48.8	48.4
decaScore				(586.1)	513.6	546.4	533.8	562.7	571.7

- Anti-curriculum pre-training for QA improves over fully joint training
- But MT was still bad

Closing the Gap: Some Recent Experiments

MQAN at ~563 with fully joint training, Set of Single Models (SOSM) started at 586.1
-- the gap started at 23

MQAN at ~571 with anti-curriculum training (SQuAD pre-training)
--dropped the gap to 15.

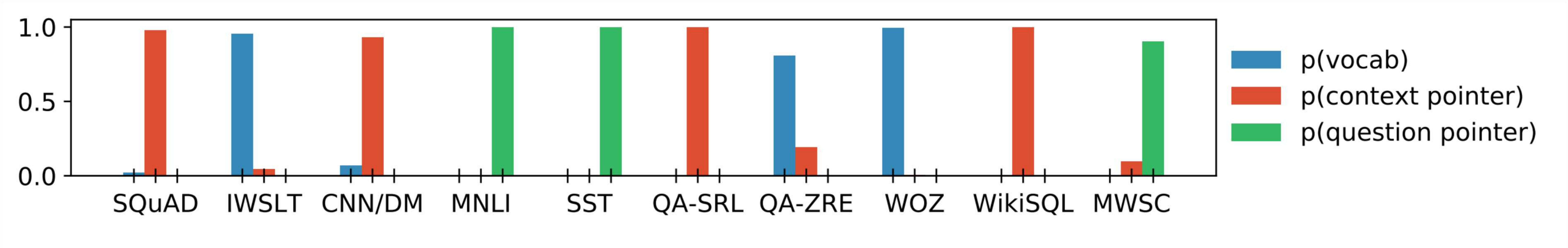
MQAN at~593 and BOSM ~618 with CoVe
--increased the gap from 15 to 25, but raised overall performance

MQAN at ~609 by including more tasks in the first phase of anti-curriculum pretraining
-- dropped the gap to about 5 points.

MQAN at ~617 by oversampling on IWSLT
--dropped the gap to 1 point

	Single-task Performance					Multitask Performance					
Dataset	S2S	+SelfAtt	+CoAtt	+QPtr	+CoVe	S2S	+SelfAtt	+CoAtt	+QPtr	+ACurr	+Cove+Tune
SQuAD	48.2	68.2	74.6	75.5	77.2	47.5	66.8	71.8	70.8	74.3	77.1
IWSLT En — De	25.0	23.3	26.0	25.5	28.2	14.2	13.6	9.00	16.1	13.7	21.4
CNN/DailyMail	19.0	20.0	25.1	24.0	26.0	25.7	14.0	15.7	23.9	24.6	23.8
MultiNLI	67.5	68.5	34.7	72.8	76.5	60.9	69.0	70.4	70.5	69.2	73.9
SST2	86.4	86.8	86.2	88.1	88.2	85.9	84.7	86.5	86.2	86.4	87.0
QA-SRL	63.5	67.8	74.8	75.2	79.2	68.7	75.1	76.1	75.8	77.6	80.4
QA-ZRE	20.0	19.9	16.6	15.6	27.0	28.5	31.7	28.5	28.0	34.7	47.0
WOZ	85.3	86.0	86.5	84.4	89.2	84.0	82.8	75.1	80.6	84.1	86.9
WikiSQL	60.0	72.4	72.3	72.6	73.0	45.8	64.8	62.9	62.0	58.7	69.7
Winograd Schemas	43.9	46.3	40.4	52.4	53.7	52.4	43.9	37.8	48.8	48.4	49.6
decaScore				(586.1)	(618.2)	$\boxed{513.6}$	546.4	533.8	562.7	571.7.	616.8

Where MQAN Points



- Answers are correctly copied from either context or question
- No confusion over which task the model should perform or which output space to use

Pretraining on decaNLP improves final performance

- For e.g. additional IWSLT language pairs
- Or new tasks like named entity recognition.

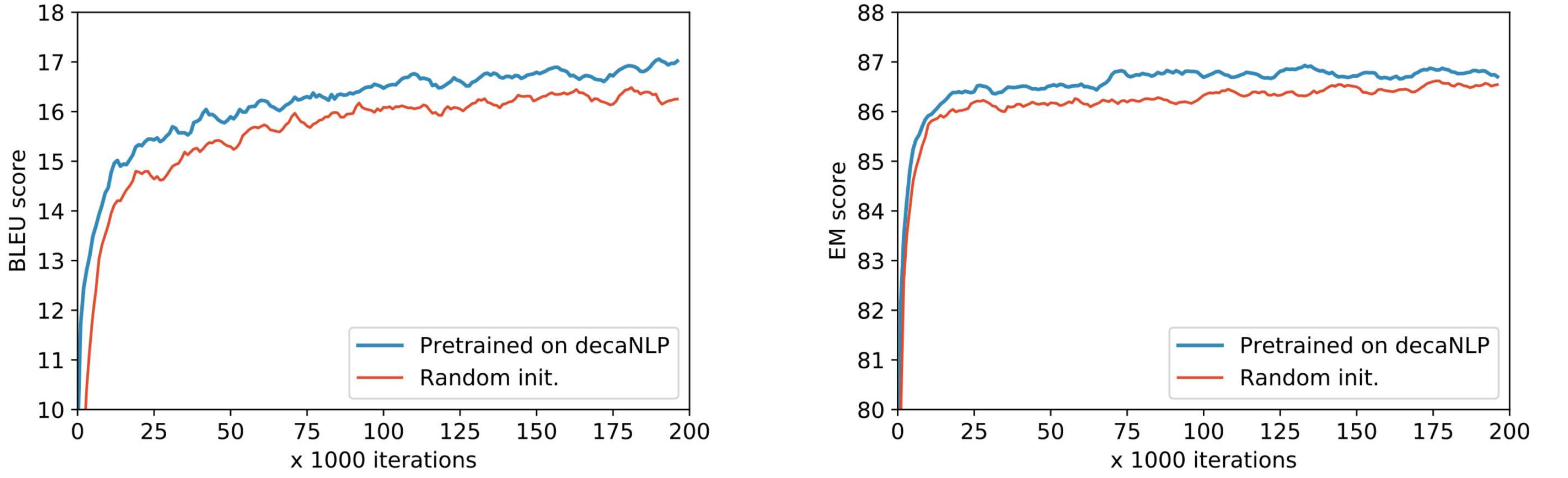


Figure 4: MQAN pretrained on decaNLP outperforms random initialization when adapting to new domains and learning new tasks. Left: training on a new language pair – English to Czech, right: training on a new task – Named Entity Recognition (NER).

Zero-Shot Domain Adaptation of pretrained MQAN:

- Achieves 80% accuracy on Amazon and Yelp reviews

- Achieves 62% on SNLI
  (87% with fine-tuning, a 2 point gain over random initialization)

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

decaNLP: A Benchmark for Generalized NLP

- Train single question answering model for multiple NLP tasks (aka questions)
- Framework for tackling
  - more general language understanding
  - multitask learning
  - domain adaptation
  - transfer learning
  - weight sharing, pre-training, fine-tuning (towards ImageNet-CNN of NLP?)
  - zero-shot learning

Related Work (tiny subset)

Multitask Learning

Collobert and J. Weston. A unified architecture for natural language processing: deep neural networks with multitask learning. In ICML, 2008.

M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. B. Viégas, M. Wattenberg, G. S. Corrado, M. Hughes, and J. Dean. Google's multilingual neural machine translation system: Enabling zero-shot translation. TACL, 5:339–351, 2017.

M.-T. Luong, Q. V. Le, I. Sutskever, O. Vinyals, and L. Kaiser. Multi-task sequence to sequence learning. CoRR, abs/1511.06114, 2015a.

L. Kaiser, A. N. Gomez, N. Shazeer, A. Vaswani, N. Parmar, L. Jones, and J. Uszkoreit. One model to learn them all. CoRR, abs/1706.05137, 2017.

Model

A. See, P. J. Liu, and C. D. Manning. Get to the point: Summarization with pointer-generator networks. In ACL, 2017.

Training

Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In ICML, 2009.

What’s next for NLP?

Thank you ☺

We are hiring, see https://einstein.ai