intermediate task transfer

CS685 Fall 2020

Advanced Natural Language Processing

Mohit Iyyer

College of Information and Computer Sciences

University of Massachusetts Amherst

many slides from Tu Vu

Stuff from last time

- Too many readings!
- The mythical HW1
- Extra credit!

What is a task?

- *a description*

MNLI The Multi-Genre Natural Language Inference Corpus (Williams et al., 2018) is a crowd-sourced collection of sentence pairs with textual entailment annotations. Given a premise sentence and a hypothesis sentence, the task is to predict whether the premise entails the hypothesis (*entailment*), contradicts the hypothesis (*contradiction*), or neither (*neutral*). The premise sentences are gathered from ten different sources, including transcribed speech, fiction, and government reports.

- *a (sample) dataset*

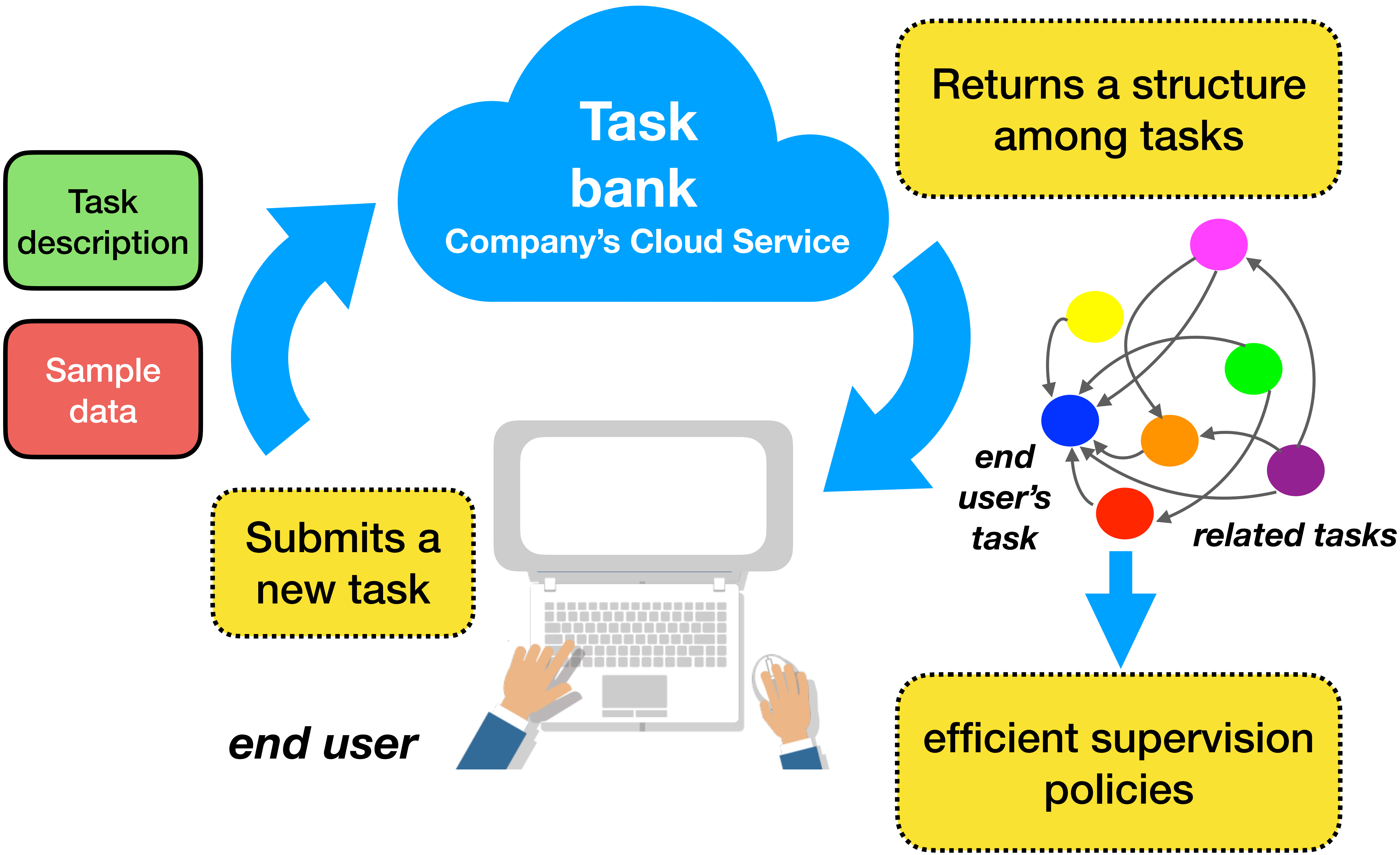
$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$$

Tasks can help each other!

- **classification**: supplementing language model (LM)-style pretraining with further training on intermediate tasks leads to improvements and reduced variance (Phang et al., 2019; arXiv)
- **sequence labeling**: pretraining on a closely related task yields better performance than LM pretraining when the pretraining dataset is fixed (Liu et al., 2019; NAACL)
- **machine comprehension**: pretraining on multiple related datasets leads to robust generalization and transfer (Talmor and Berant, 2019; ACL)

- Discover the space of language tasks
  - properties of individual tasks
  - task similarities and beneficial relations among tasks
- Practical application
  - reduce the need for supervision among related tasks
  - ***multi-task learning***: solve many tasks in one system
  - ***transfer learning***: select source tasks for a given task

A real-world scenario



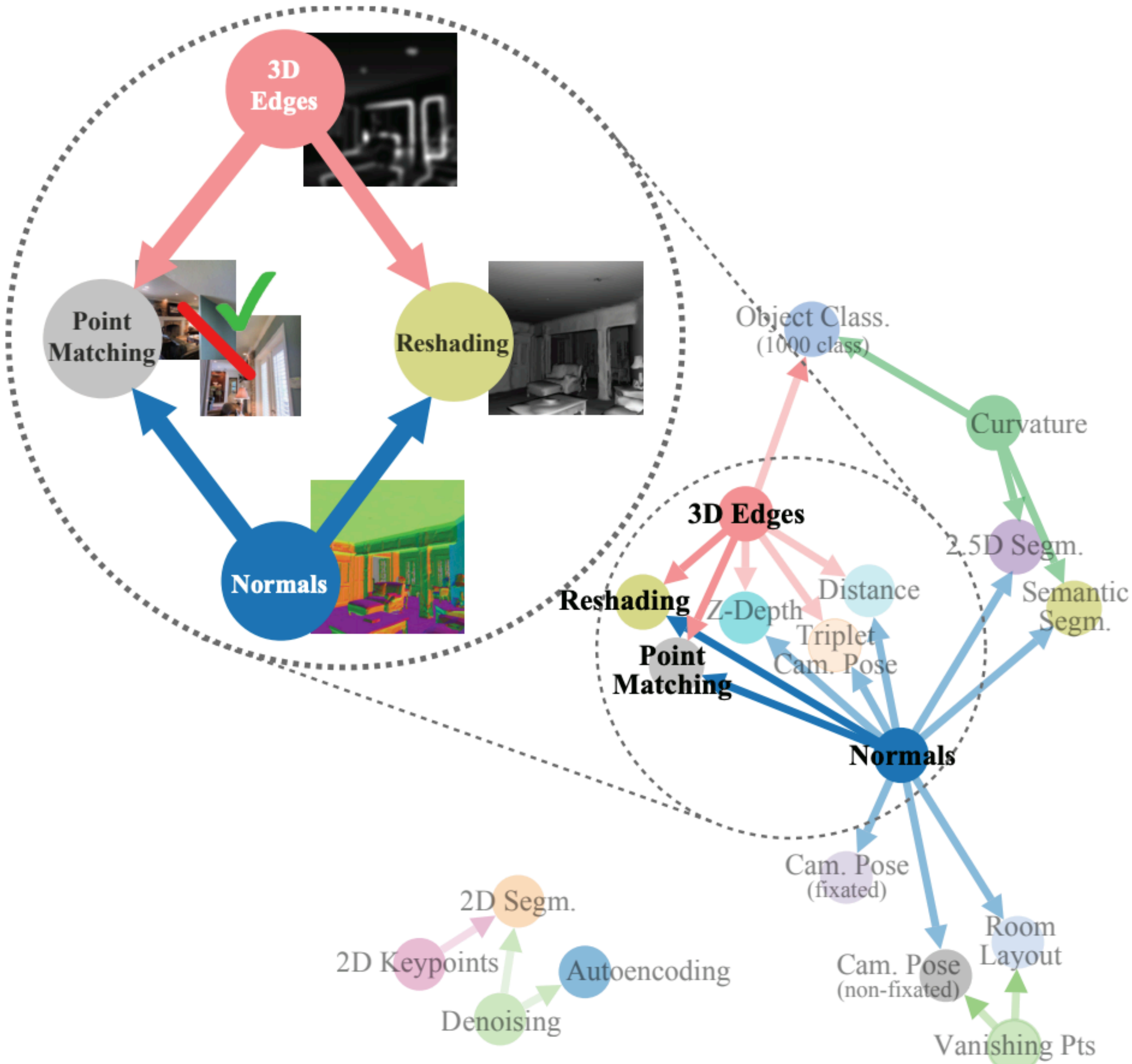
There are tons of NLP tasks!

- ~ 100 tasks/datasets from various classes of problems

Single Sentence Classification	Sentence Pair Classification	Machine Comprehension	Sequence Labeling	Unsupervised Learning	Probing Tasks
CoLA	MRPC	SQuAD	CCG	LM	SentLen
SST-2	STS-B	NewsQA	POS	autoencoding	WC
20 Newsgroups	QQP	SearchQA	Chunk	next sentence	TreeDepth
TREC-6	MNLI	TriviaQA	NER	real/fake	TopConst
IMDB	QNLI	HotpotQA	ST	discourse relations	BShift
Yelp-2	RTE	CQ	GED	…	Tense
Yelp-full	WNLI	CWQ	PS		SubjNum
AG	BoolQ	ComQA	EF		ObjNum
DBPedia	CB	WikiHOP	Parent		SOMO
Sogou News	WiC	DROP	Conj		CoordInv
…	…	…	…		…

Taskonomy for vision tasks

- *Zamir et al. (2018); CVPR: A library of 26 tasks covering common themes in computer vision (2D, 3D, semantics, etc.)*

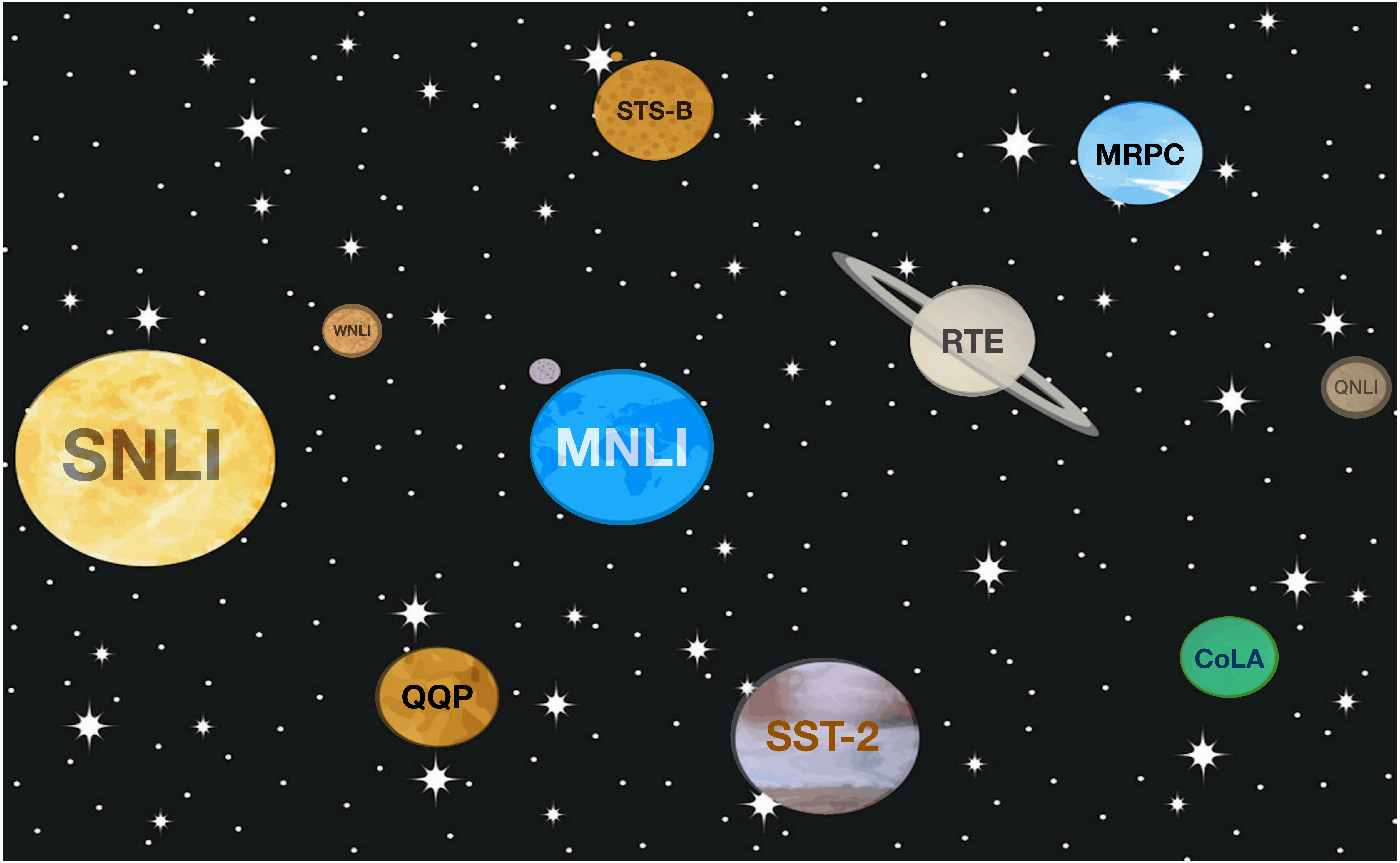


A research question

- *What criteria can be used to predict which combinations of source/intermediate and target tasks should work well?*

Create **task embeddings**

- fixed-length dense vector representations of *tasks*
- The vector space can tell us how closely related two tasks are (i.e., via cosine distance)


STS-B
MRPC
RTE
WNLI
QNLI
SNLI
MNLI
CoLA
QQP
SST-2

Previous work on exploring the relations between NLP tasks

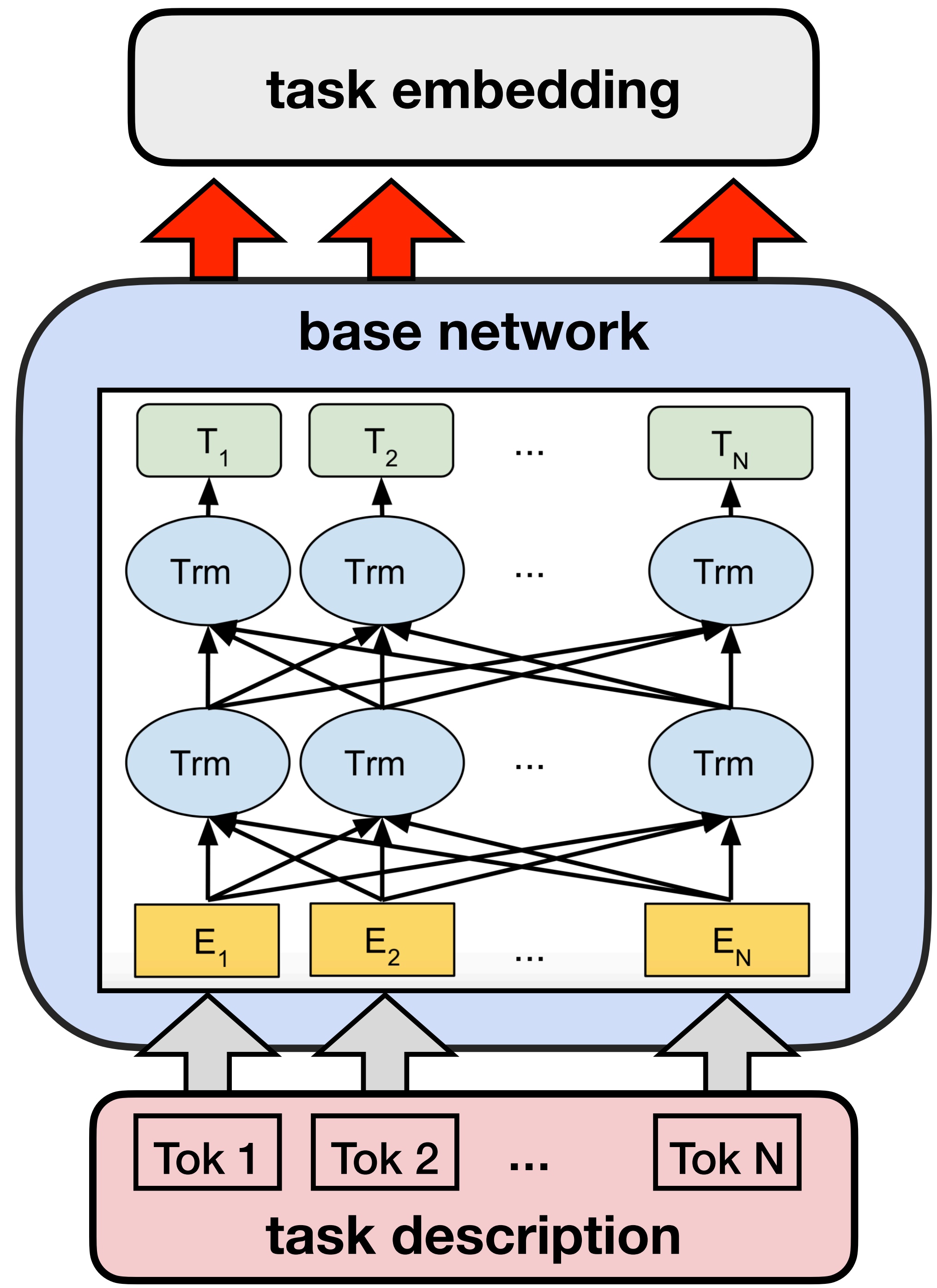
- *Bingel and Søgaard (2017); EACL: 10 main sequence labeling tasks, 90 task pairs for multi-task learning*
- *Talmor and Berant (2019); ACL: 10 main reading comprehension tasks*

	CCG	CHU	COM	FNT	POS	HYP	KEY	MWE	SEM	STR
CCG		1.4	0.45	0.58	1.8	0.24	0.3	0.45	1.4	0.84
CHU	-0.052		-0.15	-0.12	-0.45	-0.5	-0.22	-0.27	-0.099	-0.32
COM	-5	1.3		1.3	-1.4	-2.4	-4.8	0.82	-3	-0.63
FNT	-5.8	-1	-6.1		-9.4	-5.7	-3.6	-9.4	-3	-0.68
POS	4.9	2.9	1.9	0.9		-0.85	-0.26	1.3	3.4	2.9
HYP	12	4	-11	9.2	22		1.5	-7.7	23	8.1
KEY	5.7	3.2	-1	-0.43	-1.3	-2.6		-4.7	0.59	0.69
MWE	18	20	7.4	5.5	1.6	-3.8	-5.8		16	8.6
SEM	-5	-0.76	-1.2	-0.81	-0.85	-1.3	-0.83	-1.1		-1.7
STR	-1.7	1.5	-0.26	-0.72	0.037	-1.5	-1.4	-1.6	1.7	

	CQ	CWQ	ComQA	WikiHop	DROP	SQuAD	NewsQA	SearchQA	TQA-G	TQA-W	HotpotQA
SQuAD	18.0	10.1	16.1	4.2	2.4	-	**23.4**	9.5	32.0	20.9	**7.6**
NewsQA	14.9	8.2	13.5	4.8	3.0	**41.9**	-	7.7	25.3	19.9	5.3
SearchQA	29.2	16.1	24.6	8.1	2.3	17.4	10.8	-	**50.3**	**28.9**	4.5
TQA-G	**30.3**	**17.8**	**29.4**	**9.2**	3.0	30.2	15.5	**38.5**	-	-	7.2
TQA-W	24.6	14.5	17.9	8.4	2.9	24.8	15.0	20.5	-	-	6.5
HotpotQA	24.6	14.9	21.2	8.5	**7.7**	38.3	16.9	13.5	36.8	26.0	-
Multi-75K	32.8	17.9	26.7	7.4	4.3	-	-	-	-	-	-
Self	24.1	24.9	45.2	41.7	15.6	68.0	36.5	51.3	58.9	41.6	22.5
SQuAD	23.6	12.0	20.0	4.6	5.5	-	**31.8**	8.4	37.8	33.4	**11.8**
NewsQA	24.1	12.4	18.9	7.1	4.4	**60.4**	-	10.1	37.6	28.4	8.0
SearchQA	30.3	18.5	25.8	12.4	2.8	23.3	12.7	-	**53.2**	**35.4**	5.2
TQA-G	**35.4**	**19.7**	**28.6**	6.3	3.6	36.3	18.8	**39.2**	-	-	8.8
TQA-W	30.3	16.5	23.6	**12.6**	5.1	35.5	19.4	27.8	-	-	8.7
HotpotQA	27.7	15.5	22.1	10.2	**9.1**	54.5	25.6	19.6	37.3	34.9	-
Multi-75K	34.0	18.2	30.9	11.7	8.6	-	-	-	-	-	-
Self	30.8	27.1	51.6	52.9	17.9	78.0	46.0	52.2	60.7	50.1	24.2

A simple approach

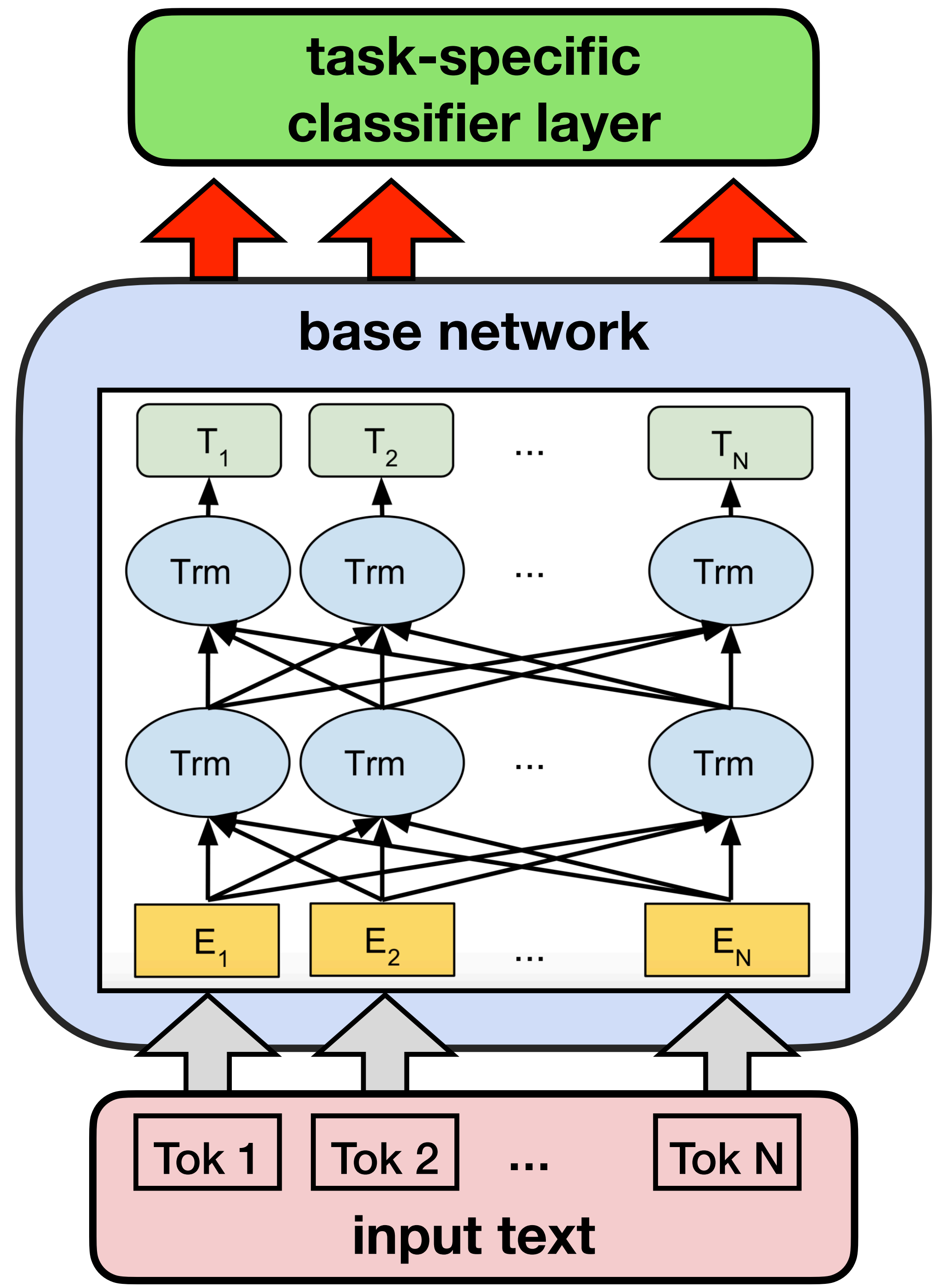
- use the task description only (i.e., a paragraph describing the task)
- Limitation: requires a clear description for each task in the library



Gradient-based methods

- use a single base network
- add a task-specific layer for a given task
- pass the entire dataset forward through the network only once
- during backpropagation:

  either use **training labels** or **sample from the model's predictive distribution** to compute gradients w.r.t. the model's parameters (**weights**) or outputs (**activations**)



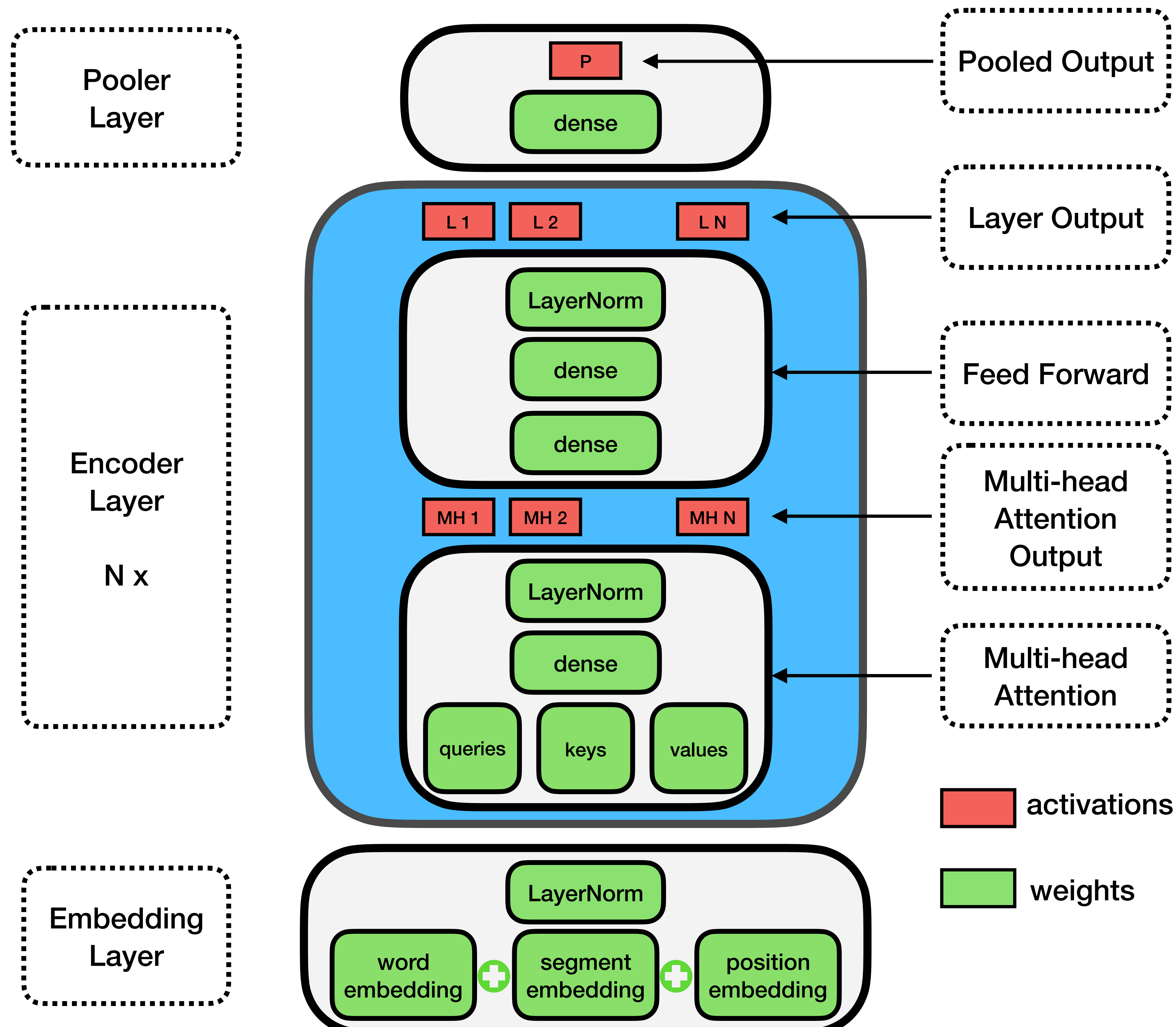
What is the base network?

- a pre-trained model, e.g., BERT, XLNet, RoBERTa

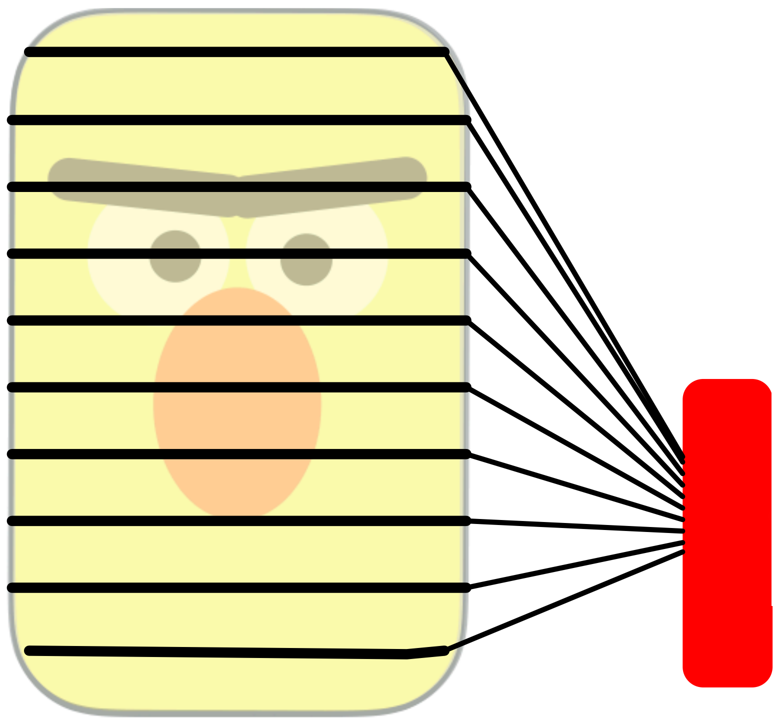
How to get gradient information?

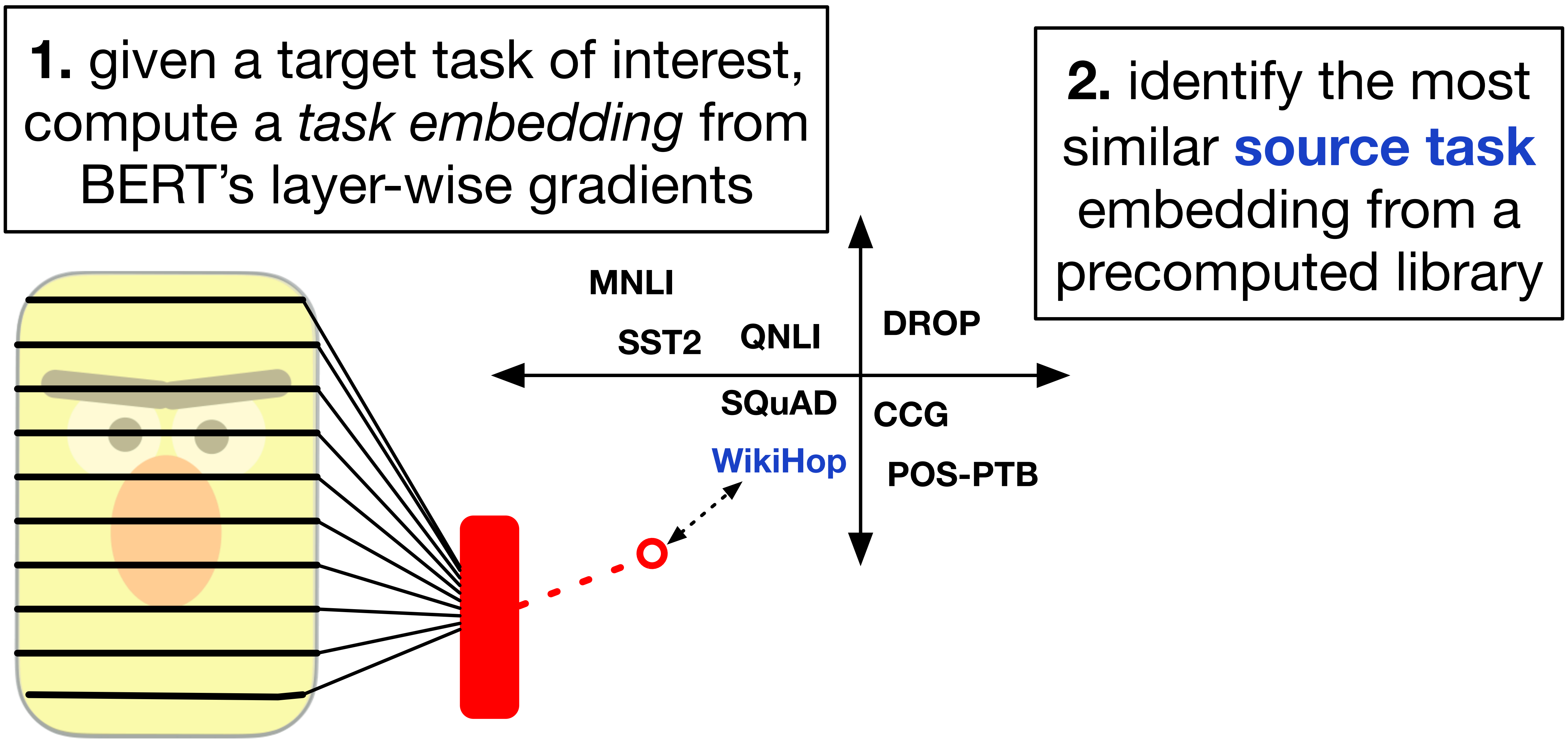
- **use training labels**
  - original gradients $\nabla_\theta \log p_\theta(y_n|x_n)$
  - use the **empirical Fisher** $\nabla_\theta \log p_\theta(y_n|x_n)\, \nabla_\theta \log p_\theta(y_n|x_n)^\top$
- **sample from the model's predictive distribution**
  - original gradients
    $$\mathbb{E}_{p_\theta(y|x_n)}\left[\nabla_\theta \log p_\theta(y|x_n)\right]$$
  - use the **theoretical Fisher**
    $$\mathbb{E}_{p_\theta(y|x_n)}\left[\nabla_\theta \log p_\theta(y|x_n)\, \nabla_\theta \log p_\theta(y|x_n)^\top\right]$$

Various gradient types

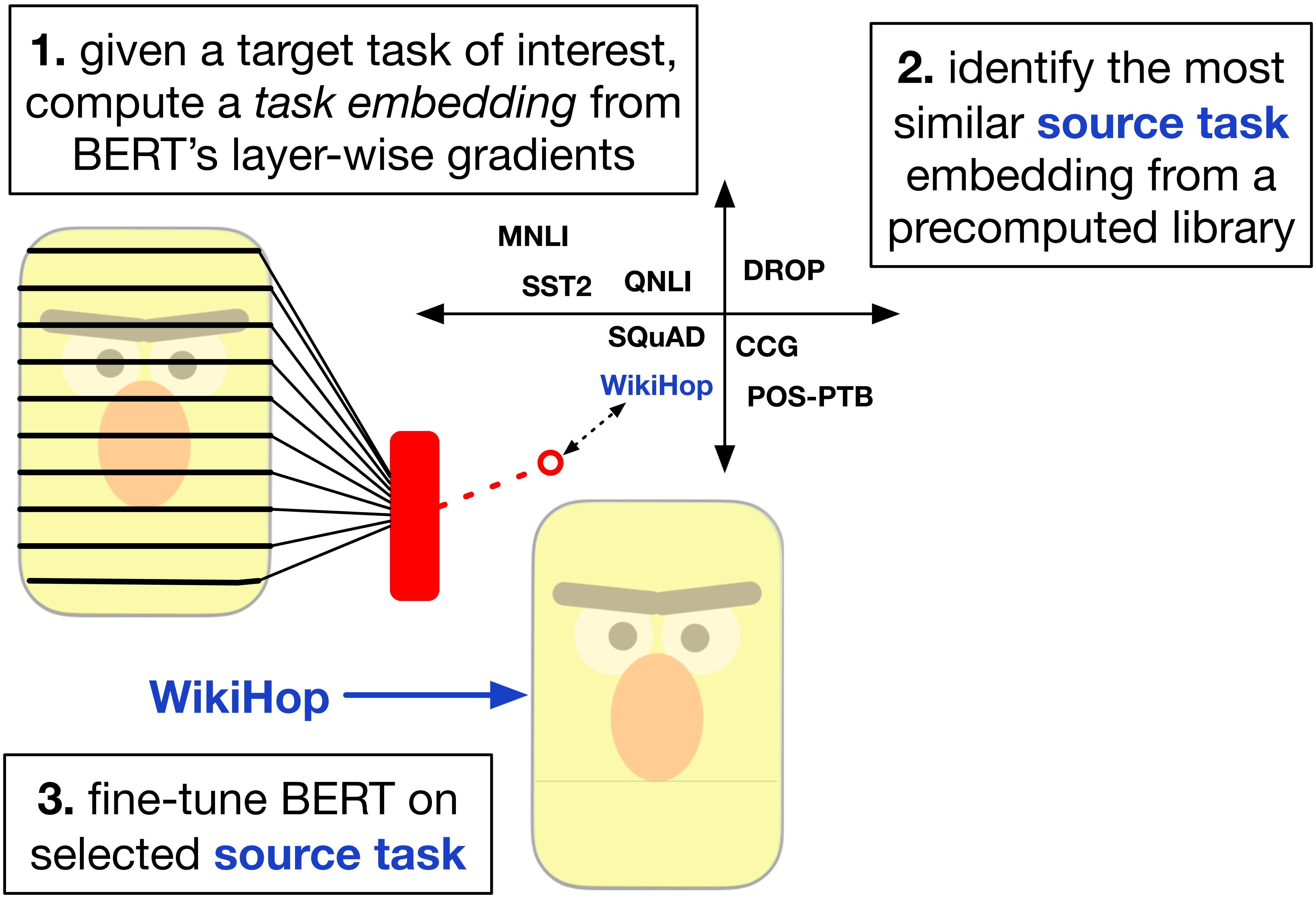


1. given a target task of interest, compute a *task embedding* from BERT's layer-wise gradients

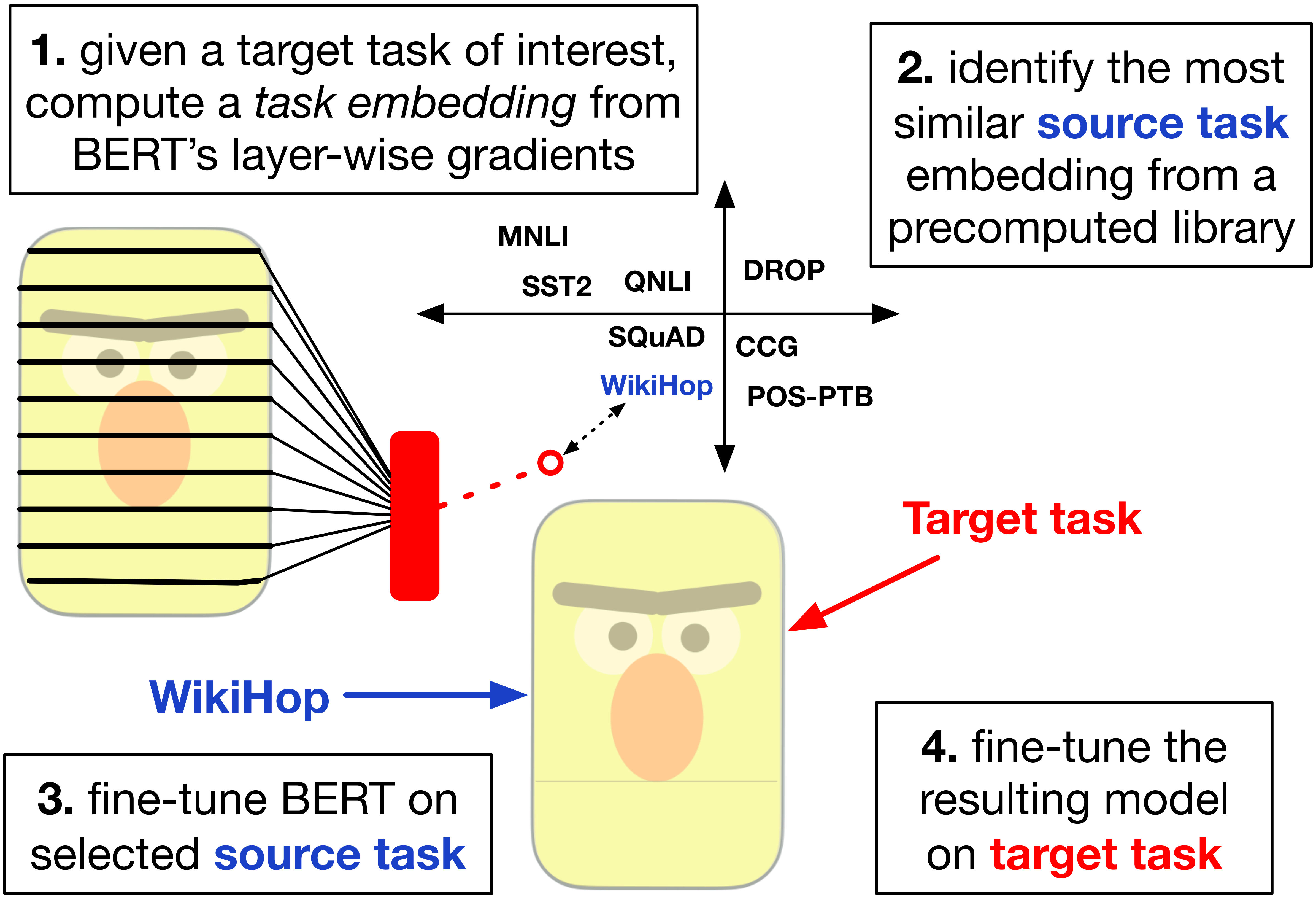


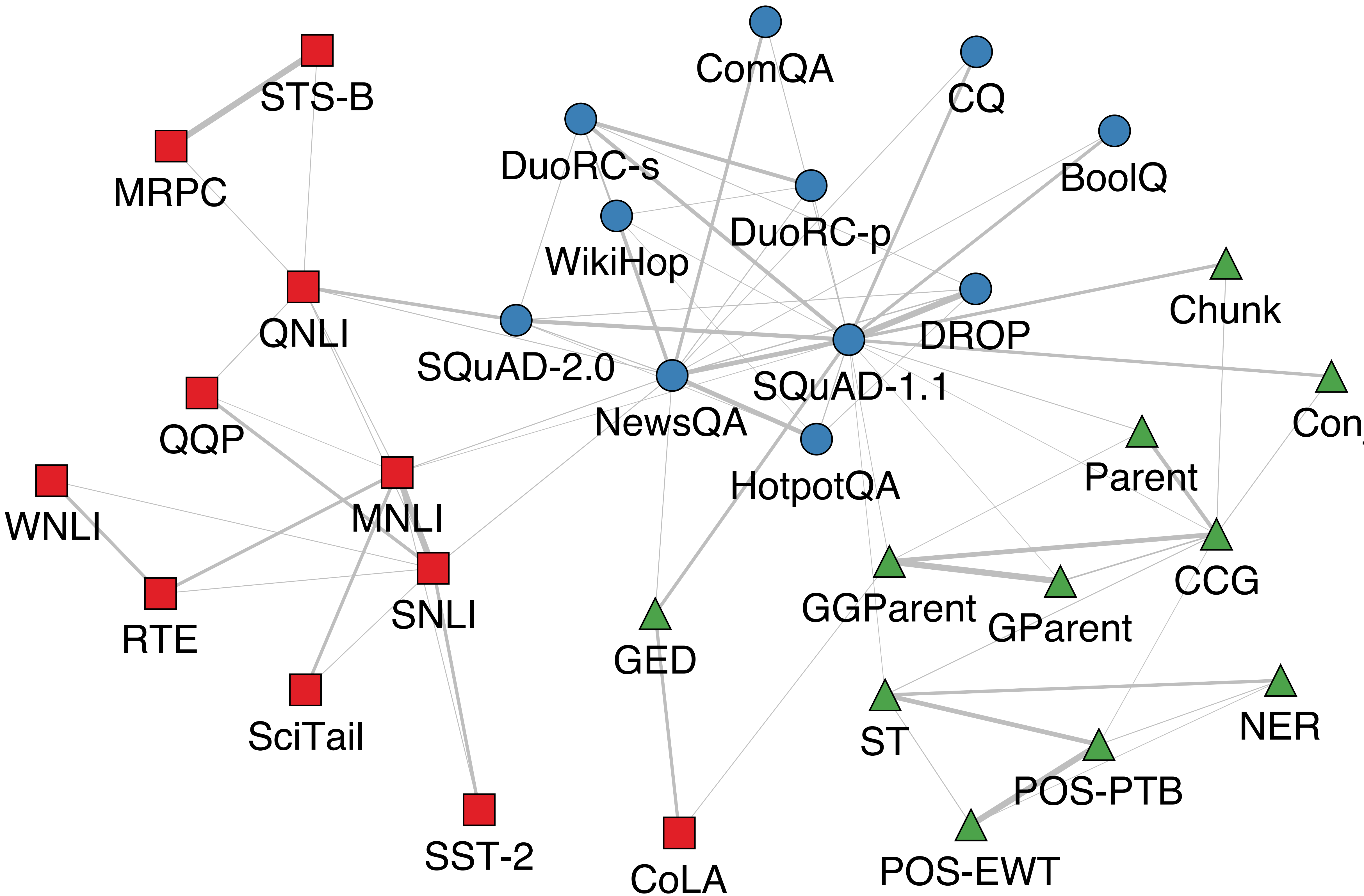


1. given a target task of interest, compute a task embedding from BERT's layer-wise gradients
2. identify the most similar source task embedding from a precomputed library
MNLI
SST2
QNLI
DROP
SQuAD
CCG
WikiHop
POS-PTB

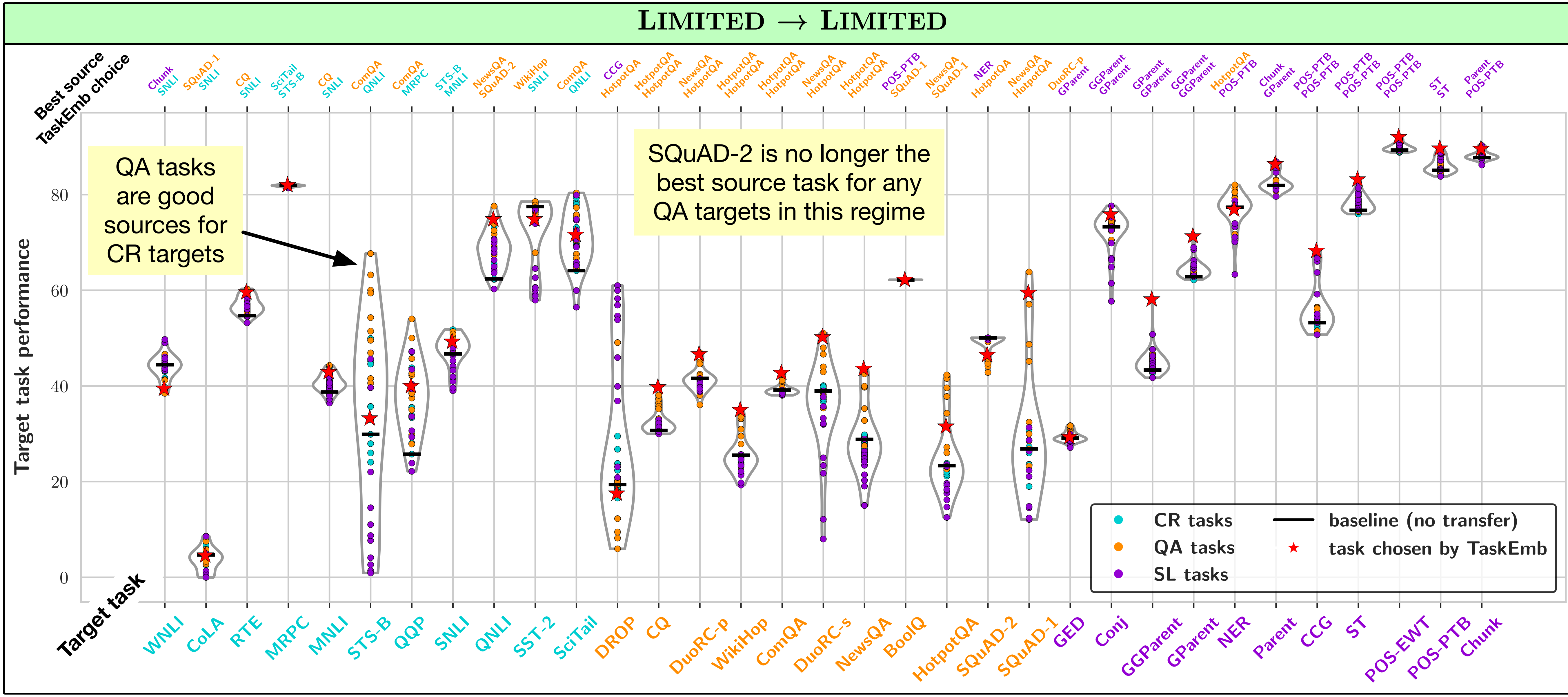


1. given a target task of interest, compute a *task embedding* from BERT's layer-wise gradients
2. identify the most similar **source task** embedding from a precomputed library
MNLI
SST2
QNLI
DROP
SQuAD
CCG
WikiHop
POS-PTB
WikiHop
3. fine-tune BERT on selected **source task**

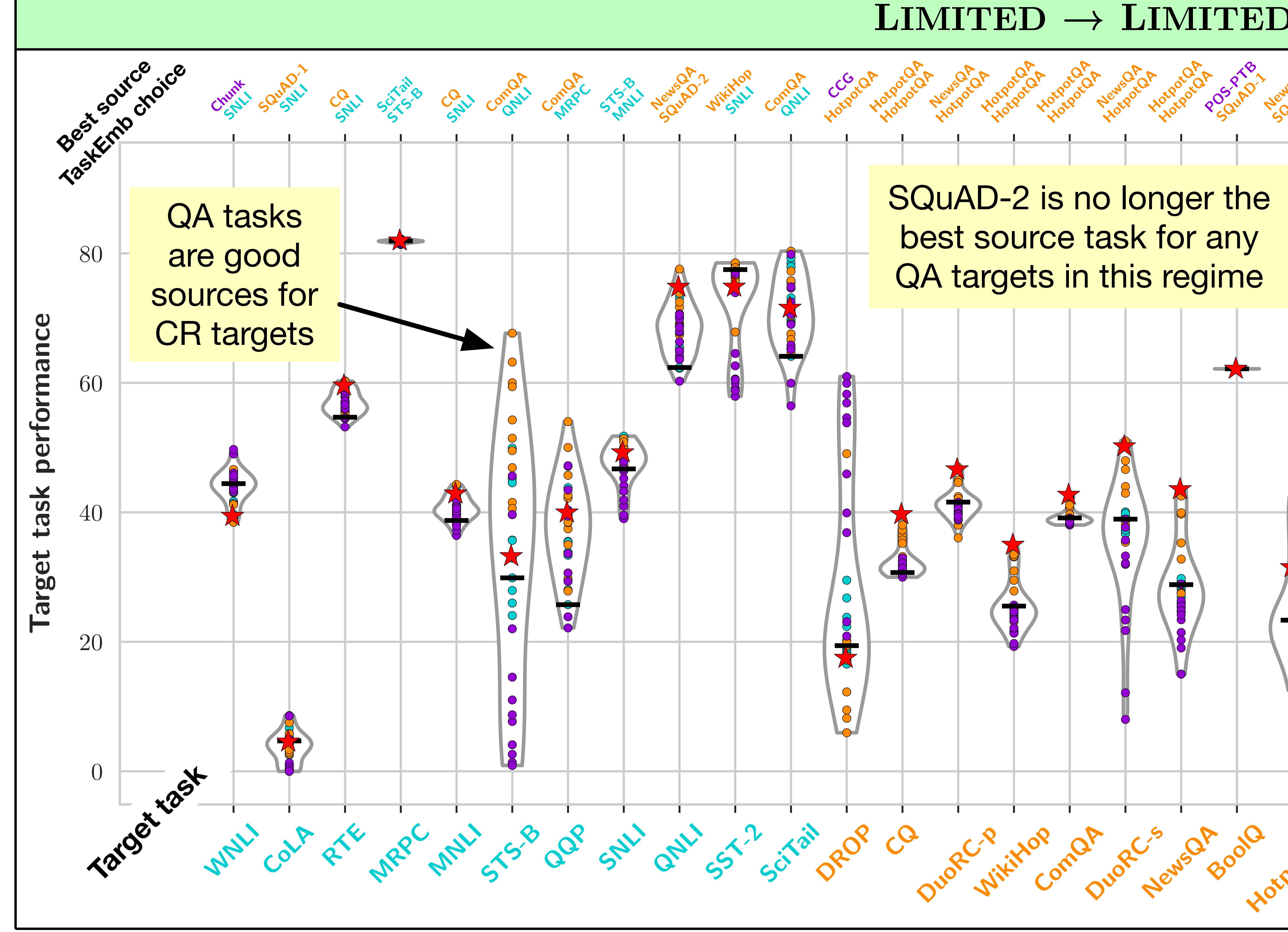

1. given a target task of interest, compute a *task embedding* from BERT's layer-wise gradients
2. identify the most similar source task embedding from a precomputed library
MNLI
SST2
QNLI
DROP
SQuAD
CCG
WikiHop
POS-PTB
Target task
WikiHop
3. fine-tune BERT on selected source task
4. fine-tune the resulting model on target task



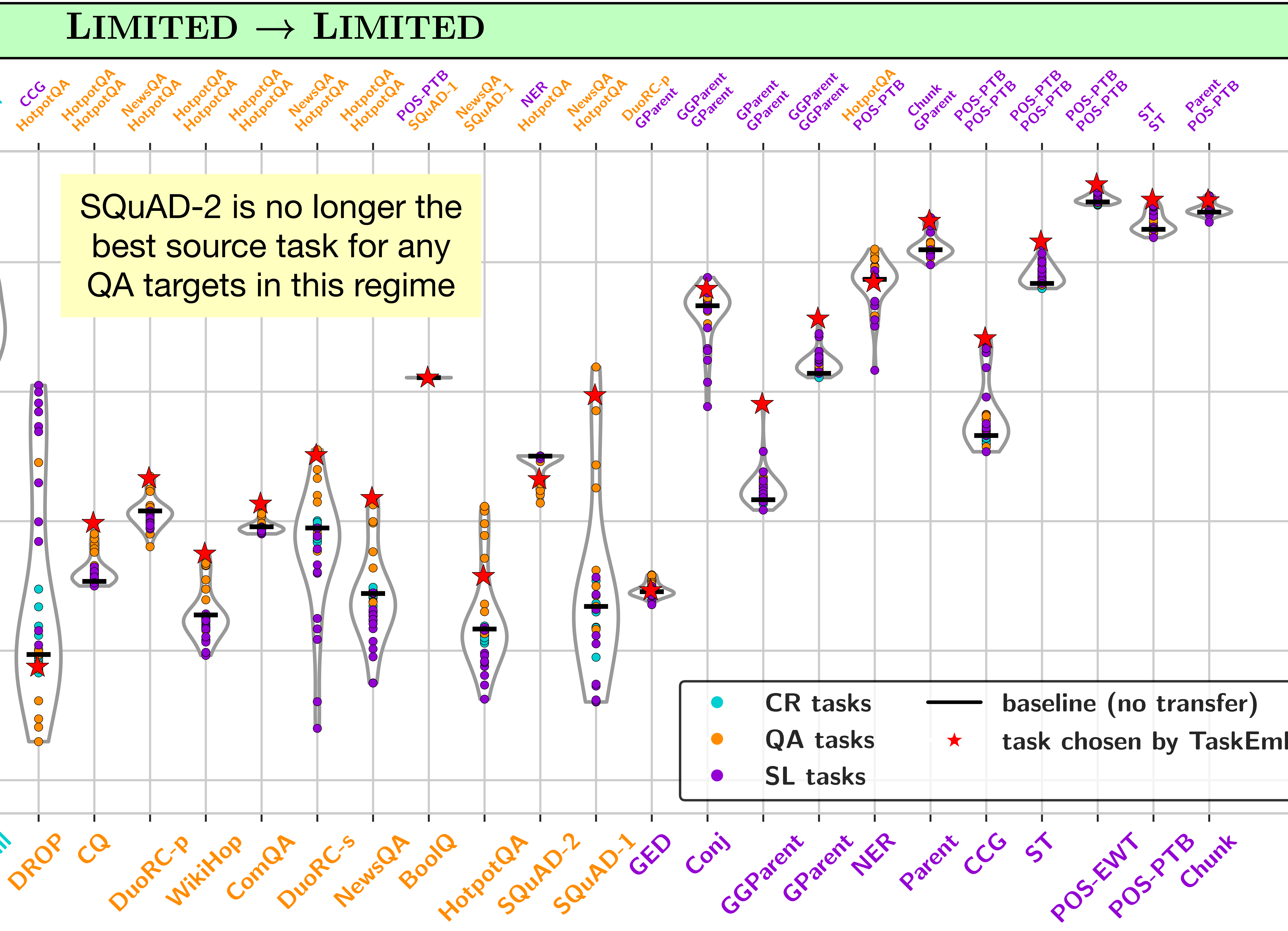
ComQA
CQ
STS-B
MRPC
DuoRC-s
BoolQ
DuoRC-p
WikiHop
Chunk
QNLI
DROP
SQuAD-2.0
SQuAD-1.1
NewsQA
Conj
QQP
Parent
HotpotQA
WNLI
MNLI
CCG
SNLI
RTE
GGParent
GParent
GED
SciTail
ST
NER
POS-PTB
SST-2
CoLA
POS-EWT



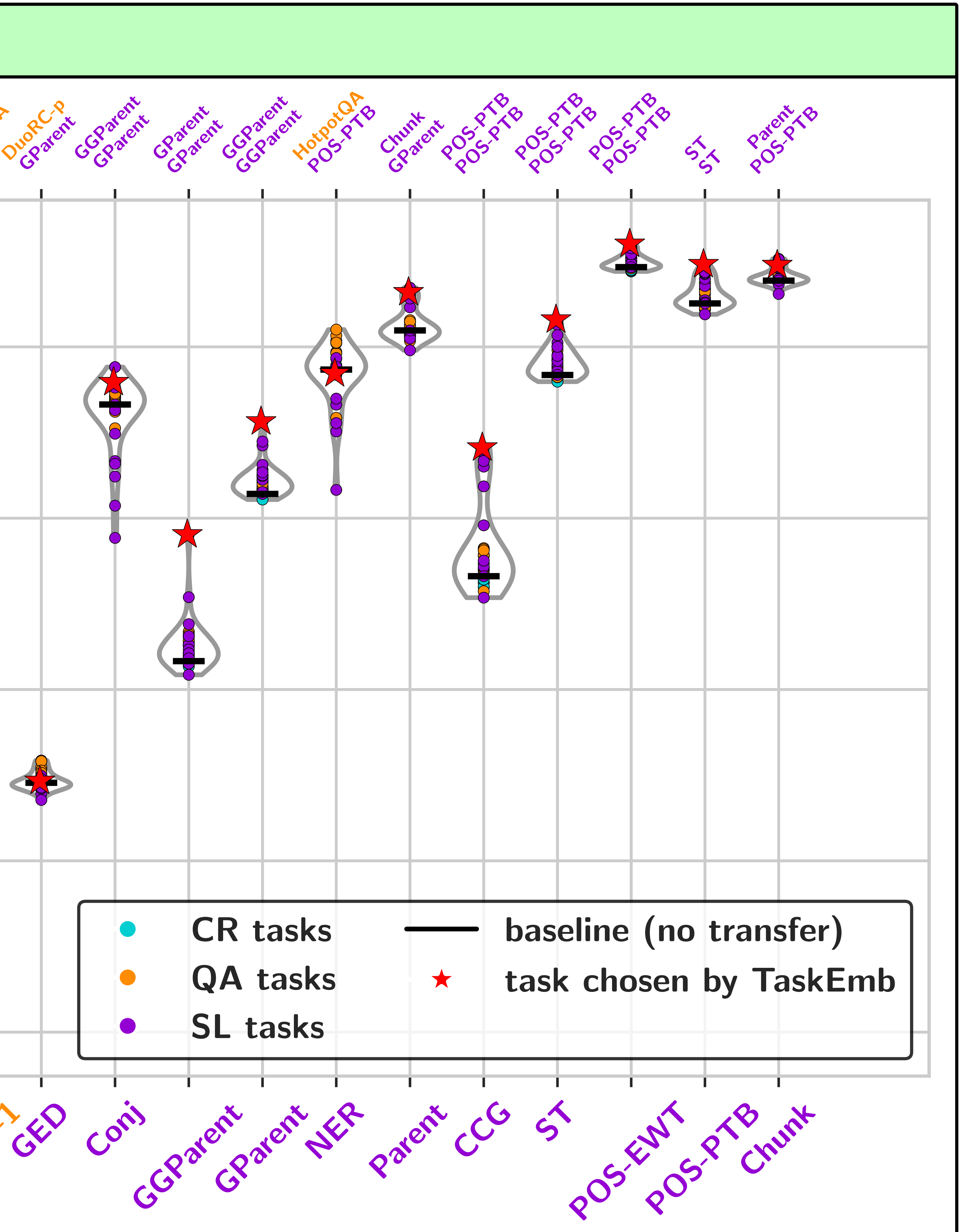
LIMITED → LIMITED
Best source
TaskEmb choice
Chunk SNLI
SQuAD-1 SNLI
CQ SNLI
SciTail STS-B
CQ SNLI
ComQA QNLI
ComQA MRPC
STS-B MNLI
NewsQA SQuAD-2
WikiHop SNLI
ComQA QNLI
CCG HotpotQA
HotpotQA HotpotQA
NewsQA HotpotQA
HotpotQA HotpotQA
HotpotQA HotpotQA
NewsQA HotpotQA
HotpotQA HotpotQA
POS-PTB SQuAD-1
NewsQA SQuAD-1
NER HotpotQA
NewsQA HotpotQA
DuoRC-p GParent
GGParent GParent
GParent GParent
GGParent GGParent
HotpotQA POS-PTB
Chunk GParent
POS-PTB POS-PTB
POS-PTB POS-PTB
POS-PTB POS-PTB
ST ST
Parent POS-PTB
QA tasks are good sources for CR targets
SQuAD-2 is no longer the best source task for any QA targets in this regime
Target task performance
0
20
40
60
80
CR tasks
QA tasks
SL tasks
baseline (no transfer)
task chosen by TaskEmb
Target task
WNLI
CoLA
RTE
MRPC
MNLI
STS-B
QQP
SNLI
QNLI
SST-2
SciTail
DROP
CQ
DuoRC-p
WikiHop
ComQA
DuoRC-s
NewsQA
BoolQ
HotpotQA
SQuAD-2
SQuAD-1
GED
Conj
GGParent
GParent
NER
Parent
CCG
ST
POS-EWT
POS-PTB
Chunk



LIMITED → LIMITED
Best source
TaskEmb choice
Chunk SNLI
SQuAD-1 SNLI
CQ SNLI
SciTail STS-B
CQ SNLI
ComQA QNLI
ComQA MRPC
STS-B MNLI
NewsQA SQuAD-2
WikiHop SNLI
ComQA QNLI
CCG HotpotQA
HotpotQA HotpotQA
NewsQA HotpotQA
HotpotQA HotpotQA
HotpotQA HotpotQA
NewsQA HotpotQA
HotpotQA HotpotQA
POS-PTB SQuAD-1
QA tasks are good sources for CR targets
SQuAD-2 is no longer the best source task for any QA targets in this regime
Target task performance
0
20
40
60
80
Target task
WNLI
CoLA
RTE
MRPC
MNLI
STS-B
QQP
SNLI
QNLI
SST-2
SciTail
DROP
CQ
DuoRC-p
WikiHop
ComQA
DuoRC-s
NewsQA
BoolQ



LIMITED → LIMITED
SQuAD-2 is no longer the best source task for any QA targets in this regime
CCG
HotpotQA
HotpotQA
HotpotQA
NewsQA
HotpotQA
HotpotQA
HotpotQA
HotpotQA
HotpotQA
NewsQA
HotpotQA
HotpotQA
HotpotQA
POS-PTB
SQuAD-1
NewsQA
SQuAD-1
NER
HotpotQA
NewsQA
HotpotQA
DuoRC-p
GParent
GGParent
GParent
GParent
GParent
GGParent
GGParent
HotpotQA
POS-PTB
Chunk
GParent
POS-PTB
POS-PTB
POS-PTB
POS-PTB
POS-PTB
POS-PTB
ST
ST
Parent
POS-PTB
DROP
CQ
DuoRC-p
WikiHop
ComQA
DuoRC-s
NewsQA
BoolQ
HotpotQA
SQuAD-2
SQuAD-1
GED
Conj
GGParent
GParent
NER
Parent
CCG
ST
POS-EWT
POS-PTB
Chunk
CR tasks
QA tasks
SL tasks
baseline (no transfer)
task chosen by TaskEm


DuoRC-p
GParent
GGParent
GParent
GParent
GParent
GGParent
GGParent
HotpotQA
POS-PTB
Chunk
GParent
POS-PTB
POS-PTB
POS-PTB
POS-PTB
POS-PTB
POS-PTB
ST
ST
Parent
POS-PTB
CR tasks
QA tasks
SL tasks
baseline (no transfer)
task chosen by TaskEmb
GED
Conj
GGParent
GParent
NER
Parent
CCG
ST
POS-EWT
POS-PTB
Chunk