Speech Representation Learning through Self-supervised Pretraining and Multi-task Finetuning

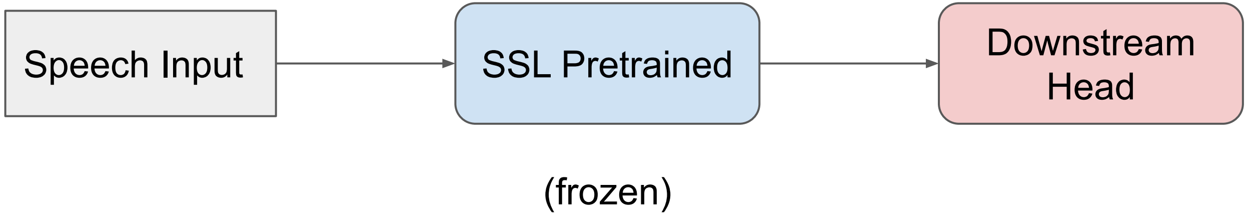
Yi-Chen Chen, Shu-wen Yang, Cheng-Kuang Lee, Simon See, Hung-yi Lee

Motivation - Representation Learning via Pretraining

- For example,
  - ImageNet pretraining in CV -> object detection
  - BERT in NLP -> question answering
  - wav2vec in speech -> automatic speech recognition

Motivation - Representation Learning via SSL Pretraining

- Self-supervised Learning (SSL)
  - Generative losses
    - APC, Mockingjay, Tera, DeCoAR, …
  - Discriminative losses
    - CPC, wav2vec (2.0), HuBERT, …
  - Multiple losses
    - PASE(+), …



Motivation - Representation Learning via Multi-task Learning

- General representation learning for a variety of speech processing tasks
- Supervised multi-task learning (MTL) is to train a shared model on various downstream tasks.
- There has not been a systematic study of general representation learning models trained by MTL of various speech processing tasks.
- We want to investigate if MTL on various downstream tasks can further improve the representations from SSL.
- SUPERB Benchmark (Unconstrained)

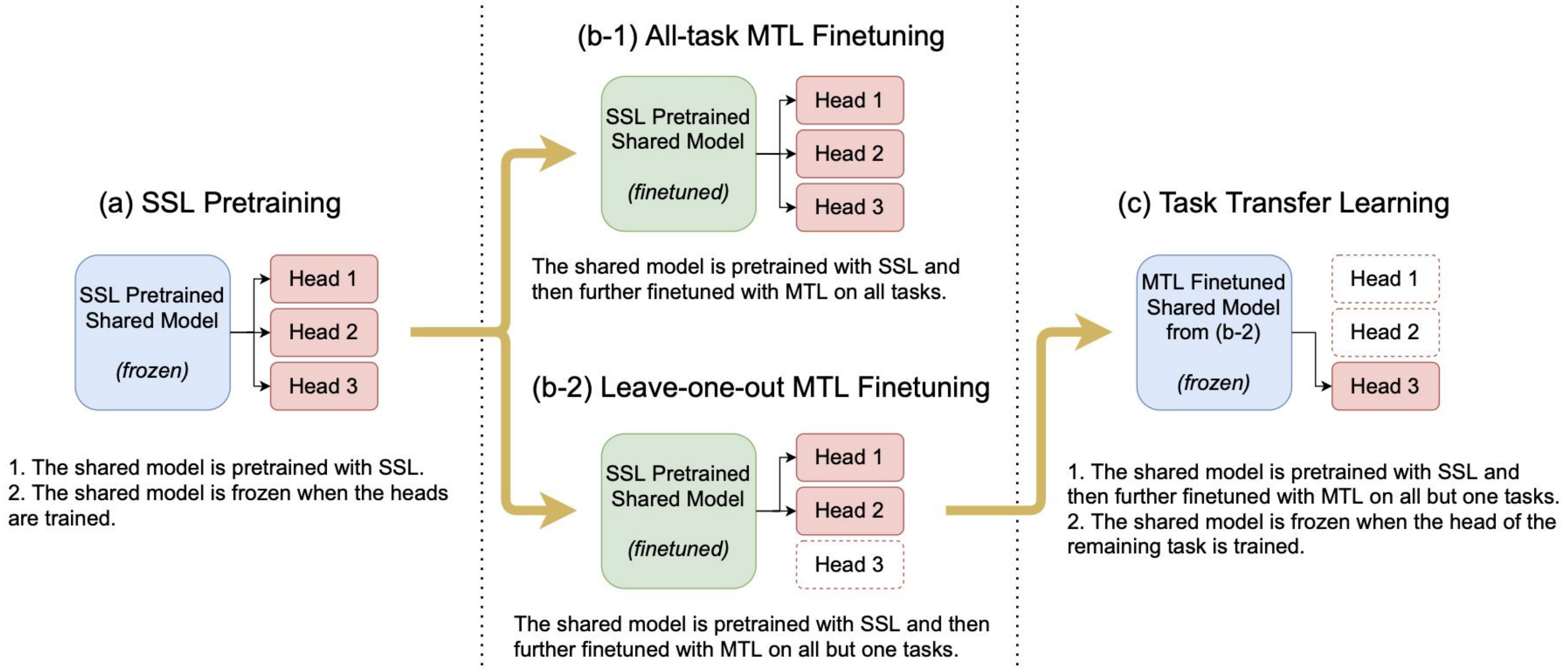
SUPERB

- Since no downstream model training is required in QbE, we only perform MTL experiments and compare the results on the other nine tasks.
  - Content
    - Phoneme Recognition (PR)
    - Automatic Speech Recognition (ASR)
    - Keyword Spotting (KS)
  - Speaker
    - Speaker Identification (SID)
    - Automatic Speaker Verification (ASV)
    - Speaker Diarization (SD)
  - Semantics
    - Intent Classification (IC)
    - Slot Filling (SF)
  - Paralinguistics
    - Emotion Recognition (ER)

Experimental Setup

- The SSL pretraining approach used in experiments:
  - HuBERT achieves the overall best performance on SUPERB.
  - We use a weighted sum of hidden representations of all layers in the HuBERT model as the representations for downstream heads, as in SUPERB.
- The model architecture and implementation details:
  - HuBERT Base is adopted as our shared model architecture.
  - For task-specific head architectures, we simply follow the settings in SUPERB.
  - The code is released: https://github.com/s3prl/s3prl/tree/multi-task-distributed

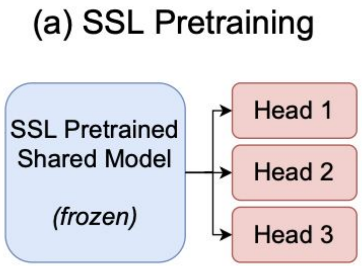
Training Scenarios



Experimental Results

Scenario	Tasks for MTL Finetuning	**ASR**	**PR**	**SF**		**SD**	**ER**	**IC**	**KS**	**ASV**	**SID**
		WER↓	**PER↓**	**F1↑**	**CER↓**	**DER↓**	**ACC↑**	**ACC↑**	**ACC↑**	**EER↓**	**ACC↑**
(a) SSL	N/A	6.42	5.41	88.53	25.20	5.88	64.24	98.34	96.30	5.11	81.42
(b-1) SSL+MTL	all	6.22	3.61	87.56	26.76	4.93	67.28	99.60	97.34	6.76	90.86
(b-2): SSL+MTL	all but ASR	X	3.63	87.28	27.11	**4.89**	65.07	**99.63**	**97.57**	7.78	90.69
	all but PR	6.79	X	86.94	27.66	**4.81**	66.73	**99.66**	**97.44**	7.94	**91.16**
	all but SF	**6.10**	**3.39**	X	X	**4.73**	65.71	99.58	97.18	7.61	90.70
	all but SD	6.28	**3.54**	**87.94**	**26.31**	X	66.73	**99.63**	97.11	7.49	90.79
	all but ER	**6.17**	**3.40**	87.45	26.90	**4.77**	X	99.55	97.27	7.19	90.51
	all but IC	**6.13**	**3.34**	**87.65**	26.94	**4.78**	66.08	X	97.27	**6.74**	90.55
	all but KS	**6.17**	**3.55**	**87.83**	26.88	**4.91**	66.27	**99.71**	X	7.86	90.67
	all but ASV	**5.90**	**2.79**	**87.88**	**26.52**	**3.61**	64.88	99.58	**97.44**	X	85.06
	all but SID	**5.95**	**3.25**	87.33	27.39	**4.50**	**68.66**	99.55	97.27	9.00	X
(c) Task Transfer	N/A	6.27	5.79	88.14	26.24	5.80	64.24	97.42	96.33	7.55	62.05

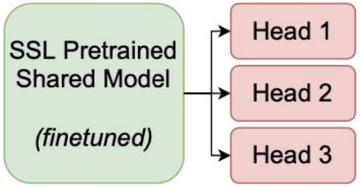
MTL Scenarios - SSL Pretraining



1. The shared model is pretrained with SSL.
2. The shared model is frozen when the heads are trained.

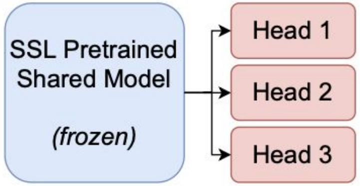
MTL Scenarios - SSL Pretraining

(b-1) All-task MTL Finetuning



The shared model is pretrained with SSL and then further finetuned with MTL on all tasks.

(a) SSL Pretraining



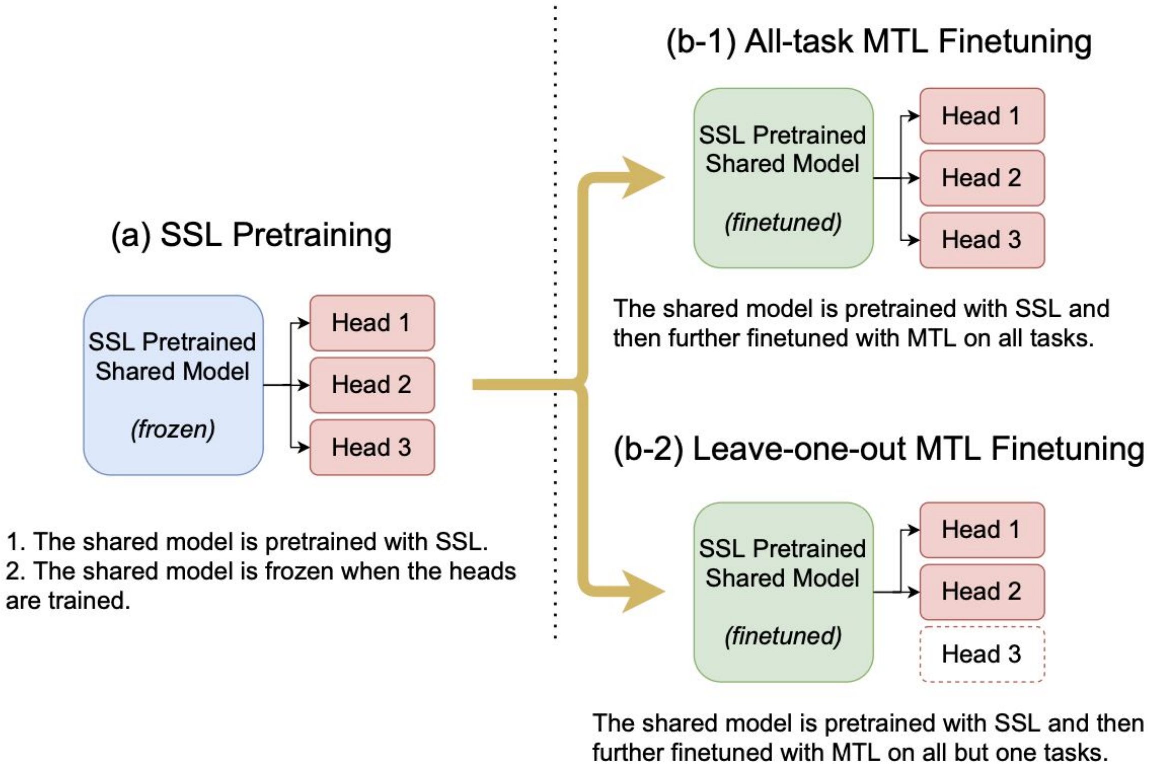
1. The shared model is pretrained with SSL.
2. The shared model is frozen when the heads are trained.

Experimental Results - All-task MTL Finetuning

Scenario	Tasks for MTL Finetuning	ASR	PR	SF		SD	ER	IC	KS	ASV	SID
		WER↓	**PER↓**	**F1↑**	**CER↓**	**DER↓**	**ACC↑**	**ACC↑**	**ACC↑**	**EER↓**	**ACC↑**
(a) SSL	N/A	6.42	5.41	88.53	25.20	5.88	64.24	98.34	96.30	5.11	81.42
(b-1) SSL+MTL	all	6.22	3.61	87.56	26.76	4.93	67.28	99.60	97.34	6.76	90.86

- Serve as a strong baseline for SSL pretraining or other representation learning approaches.
- How to select the model checkpoint based on validation scores of tasks?
  - ASV suffers from overfitting.

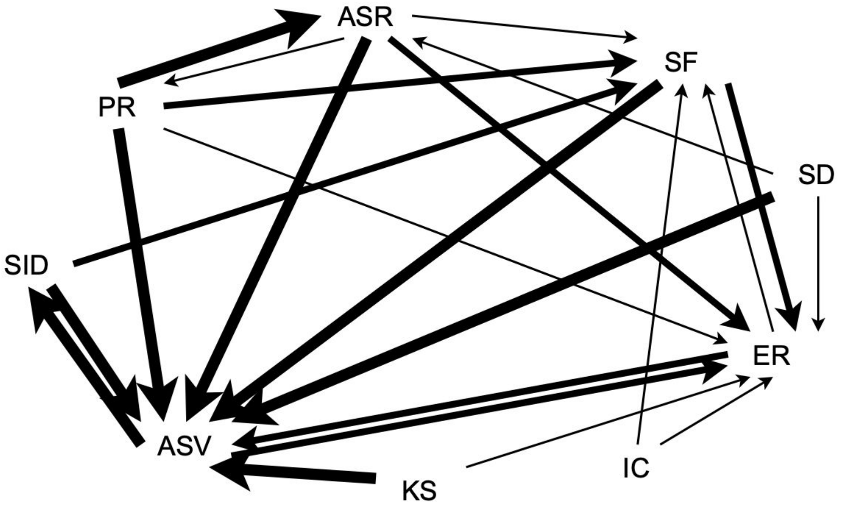
MTL Scenarios - SSL Pretraining



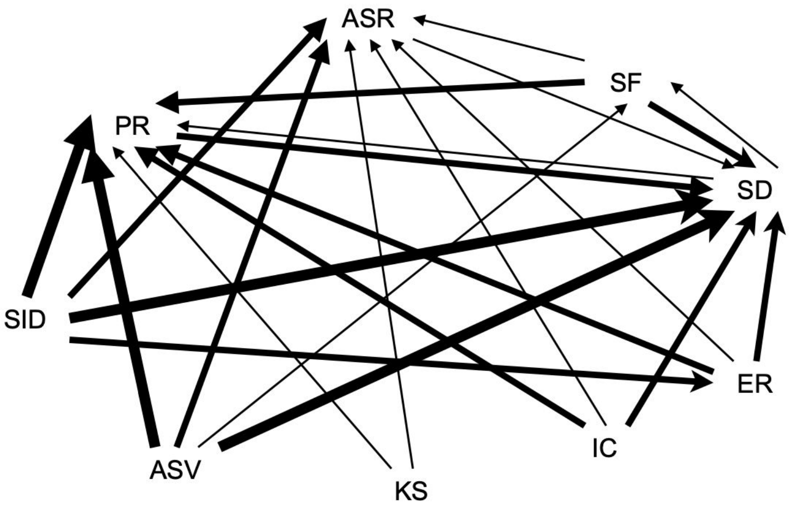
Experimental Results - Removing One Task In MTL

Scenario	Tasks for MTL Finetuning	ASR	PR	SF		SD	ER	IC	KS	ASV	SID
		WER↓	PER↓	F1↑	CER↓	DER↓	ACC↑	ACC↑	ACC↑	EER↓	ACC↑
(b-1) SSL+MTL	all	6.22	3.61	87.56	26.76	4.93	67.28	99.60	97.34	6.76	90.86
(b-2): SSL+MTL	all but ASR	X	3.63	87.28	27.11	**4.89**	65.07	**99.63**	**97.57**	7.78	90.69
	all but PR	6.79	X	86.94	27.66	**4.81**	66.73	**99.66**	**97.44**	7.94	**91.16**
	all but SF	**6.10**	**3.39**	X	X	**4.73**	65.71	99.58	97.18	7.61	90.70
	all but SD	6.28	**3.54**	**87.94**	**26.31**	X	66.73	**99.63**	97.11	7.49	90.79
	all but ER	**6.17**	**3.40**	87.45	26.90	**4.77**	X	99.55	97.27	7.19	90.51
	all but IC	**6.13**	**3.34**	**87.65**	26.94	**4.78**	66.08	X	97.27	**6.74**	90.55
	all but KS	**6.17**	**3.55**	**87.83**	26.88	**4.91**	66.27	**99.71**	X	7.86	90.67
	all but ASV	**5.90**	**2.79**	**87.88**	**26.52**	**3.61**	64.88	99.58	**97.44**	X	85.06
	all but SID	**5.95**	**3.25**	87.33	27.39	**4.50**	**68.66**	99.55	97.27	9.00	X

Experimental Results - Removing One Task In MTL



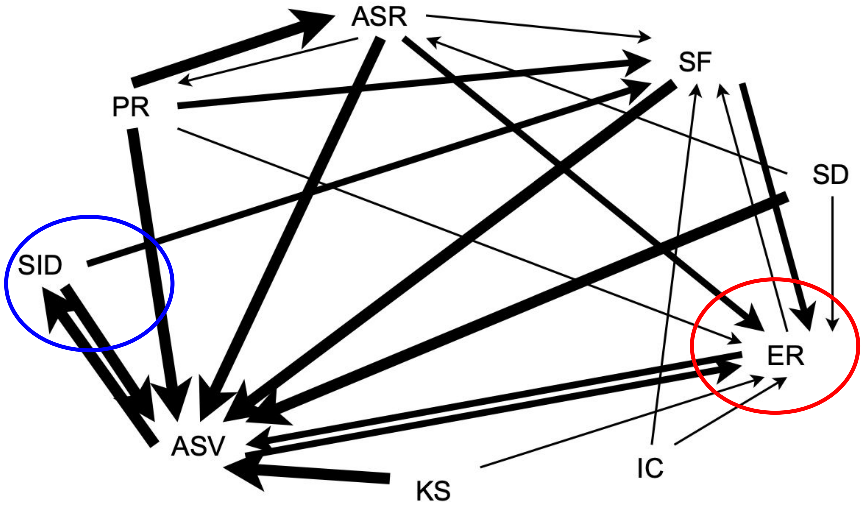
The improvement relations of tasks.



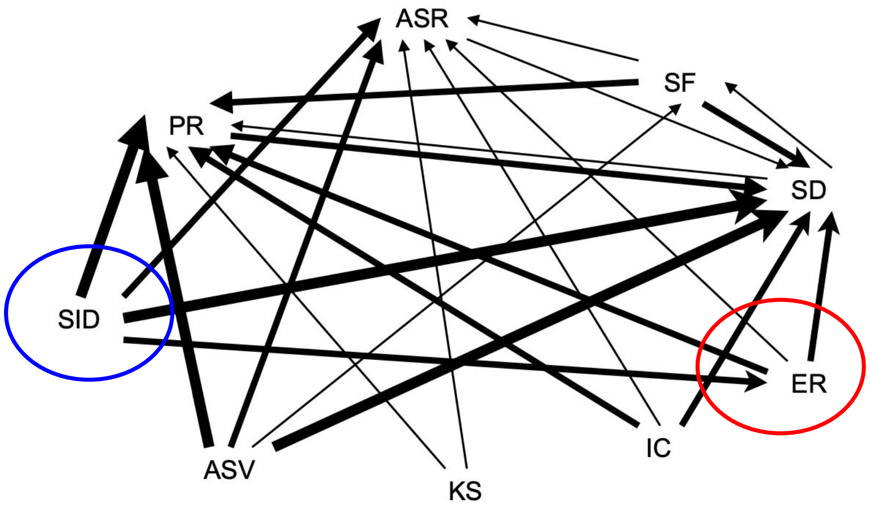
The hurt relations of tasks.

- If we focus on a certain primary task, we may select proper auxiliary tasks to assist the primary task based on these MTL experimental results.

Experimental Results - Removing One Task In MTL



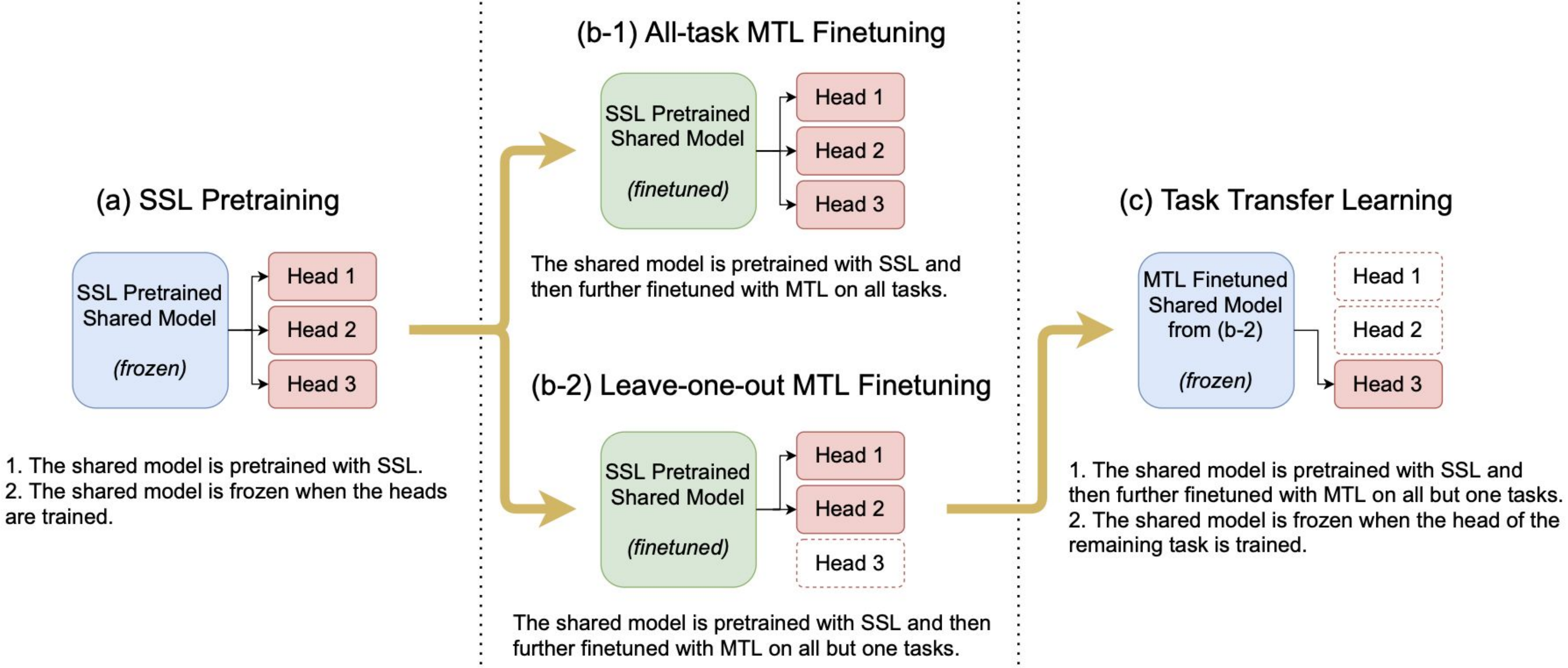
The improvement relations of tasks.



The hurt relations of tasks.

- If we focus on a certain primary task, we may select proper auxiliary tasks to assist the primary task based on these MTL experimental results.

MTL Scenarios - SSL Pretraining



Experimental Results - Task Transfer Learning

Scenario	Tasks for MTL Finetuning	ASR	PR	SF		SD	ER	IC	KS	ASV	SID
		WER↓	PER↓	F1↑	CER↓	DER↓	ACC↑	ACC↑	ACC↑	EER↓	ACC↑
(a) SSL	N/A	6.42	5.41	88.53	25.20	5.88	64.24	98.34	96.30	5.11	81.42
(b-1) SSL+MTL	all	6.22	3.61	87.56	26.76	4.93	67.28	99.60	97.34	6.76	90.86
(c) Task Transfer	N/A	6.27	5.79	88.14	26.24	5.80	64.24	97.42	96.33	7.55	62.05

- Whether a task is involved in MTL is crucial to the performance of this task.

Conclusion and Discussion

- In this work, we investigate different training scenarios of supervised MTL as a speech representation learning approach along with SSL pretraining on a benchmark with various speech processing tasks.
  - We analyze the generalizability of representations learned with supervised MTL empirically.
- The performance of MTL is dependent on many factors.
  - the amount of data
  - task relationships
  - noise
  - These factors should be isolated and investigated with more analyses.
- Future directions:
  - exploring a better method to select the model checkpoint with MTL
  - more in-depth research of MTL and its optimization on speech processing tasks
  - trying to train the shared model with both SSL, MTL and self training simultaneously as a semi-supervised representation learning approach