BERT Multi-Task Learning for Emotion Classification

CS769 Final Presentation

Team-ID: 8

Abhay Kumar, Neal Desai, Priyavarshini Murugan

Introduction & Motivation

- Understanding emotion in text has numerous applications
- Detection/intervention in social media users
- Sentiment of users of software applications
- Natural extension of the well-studied sentiment analysis problem
- Learning of granular and nuanced emotions is the goal

Prior Work/Background

- Fine tune BERT for text classification (Sun et al., 2019)
  - further pre-train BERT on within-task training data or in-domain data
  - fine-tune BERT for the target task.
  - fine-tuning BERT with multi-task learning if several related tasks are available

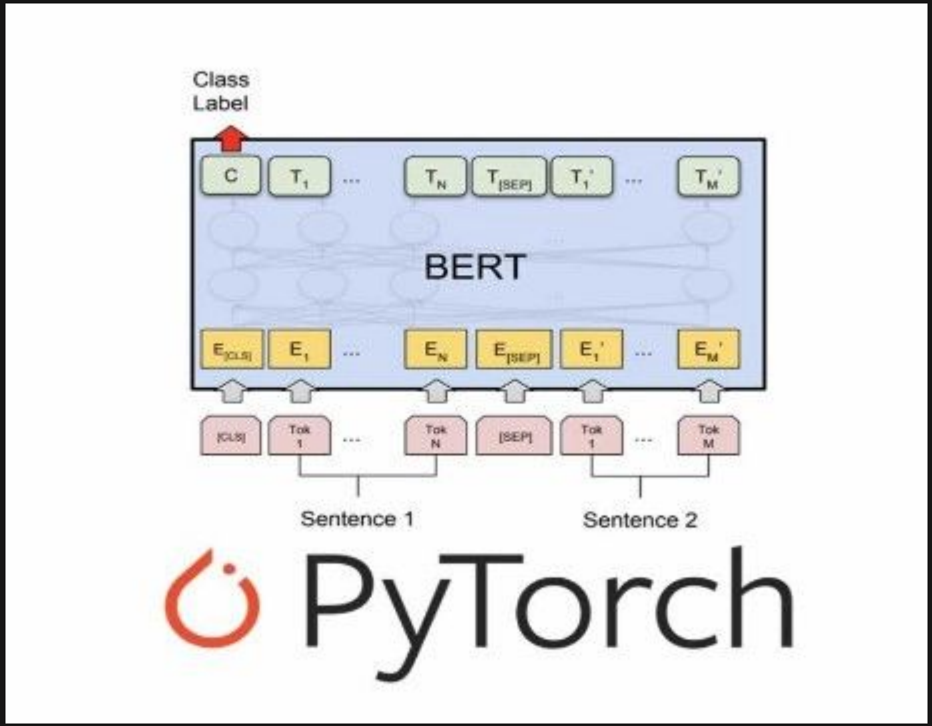
- Fine-grained emotion prediction by Modeling Emotion Definitions(Singh et al. 2021)
  - New transformer-based framework that uses semantic knowledge of emotion classes.

- Lexicon based feature extraction for emotion text classification (Bandhakavi et al., 2017)
  - uses Domain-Specific Emotion Lexicon as a direct tool for word and phrase-level emotion analysis, here we extract features to classify text into emotion classes

Datasets Used

- goEmotions: Large, manually annotated dataset of Reddit comments (Demeszky et al., 2020)
  - 3 taxonomy groupings (original, ekman, sentiment)
  - 27 emotion categories
- SST-2: Stanford sentiment treebank, 215k phrases from movie reviews.
- kaggle suicide and depression: Collection from SuicideWatch and depression subreddits over a 13-year span.

Experimental Setup (Single-Task)



We use BERT for classification in various labeled datasets.

Results (single task)

- goEmotions
  - BERT achieves very comparable results to Demezky et al for "original" grouping
  - Test accuracy of 0.44
- SST-2
  - Converges quickly to a test accuracy of 0.91, a bit less than latest state of the art of 0.95
- kaggle suicide and depression
  - Still in progress, but would expect close to state-of-the-art result.

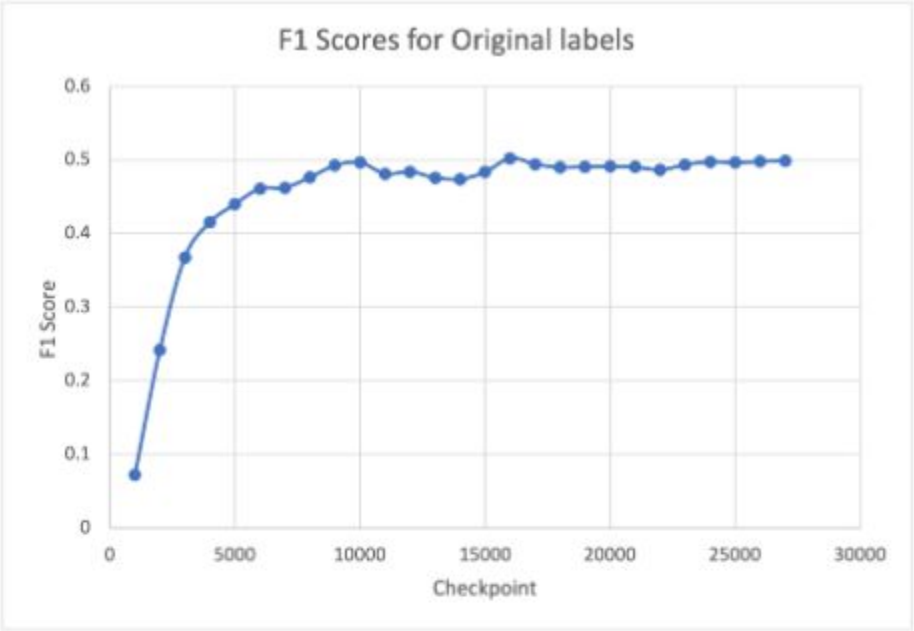
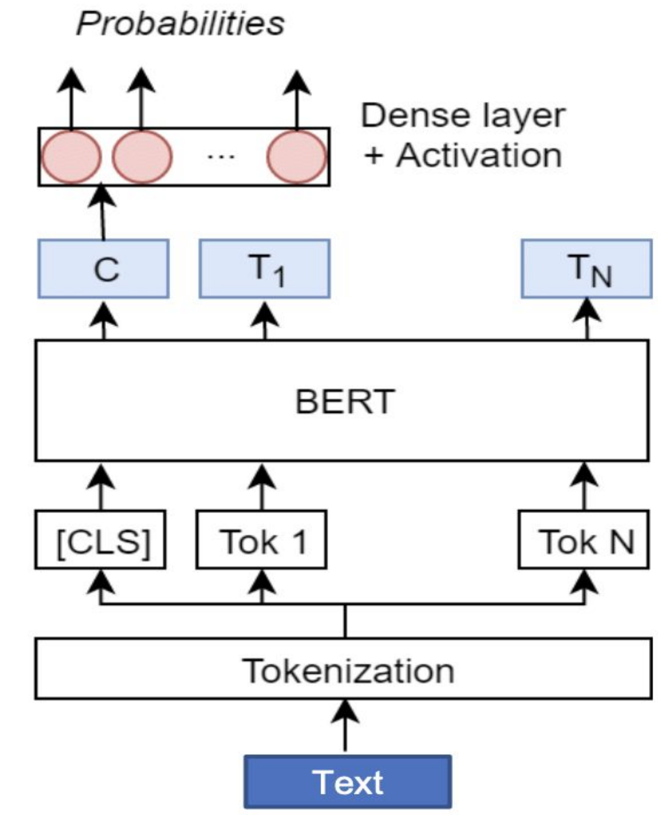
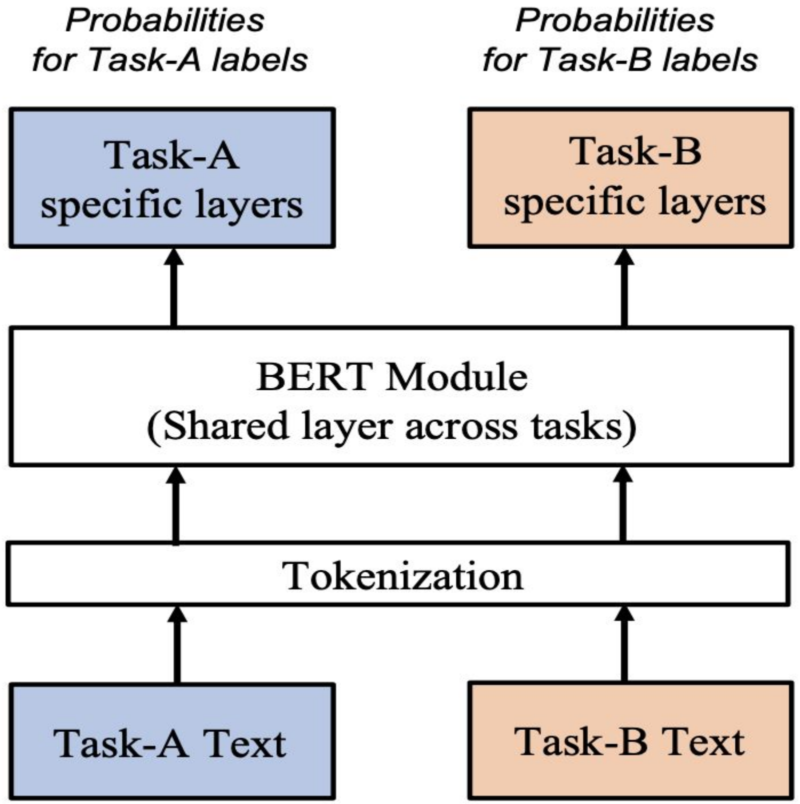


Figure 2: F1-score for original grouping over training checkpoints.

Experimental setup (multi-task)



(a) Overview of BERT classification model for a single task.



(b) Overview of Multi-Task Learning (MTL) framework.

Why multi-task learning?

- **Implicit Data Augmentation**
  - averaging out task dependent noise patterns
- **Attention Focusing**
  - focuses on relevant features for all tasks
- **Representation Bias**
  - Prefer generalized representation
  - Aids in future generalization to new tasks
- **Regularization**
  - reduces the risk of overfitting

Results and Discussion

Task Weighting

- Weighted loss to balance tasks

$$\mathcal{L}_{\mathrm{MTL}} = \sum_{t=1}^{T} \lambda_t \ell_t(\mathcal{D}_t; \theta, \psi_t)$$

- Difficult task should be upweighted in MTL setting.

Auxiliary Task/Dataset

- Accuracies in MTL setting have improved about 2-4% as compared to standalone task.

Auxiliary Tasks/Dataset	Dataset Size (train)	Auxiliary Task	Goemotion Original
SST-2	67,349	0.92(MTL) 0.93 (STL)	**0.63**(MTL) 0.59 (STL)
Twitter Sentiment140	1,600,000	**0.87** 0.85 (STL)	**0.63**(MTL) 0.59 (STL)
Goemotion-Ekman	43,410	**0.70** 0.68 (STL)	**0.61**(MTL) 0.59 (STL)

Task Sampling/Batching

- Tasks with skewed training dataset sizes. Task-1(43410) Task-2(160,000)
- Accuracies for MTL (Multi-Task learning) & STL (Single Task Learning)

Sampling Method		Task-1 GoEmotion Original		Task-2 Twitter Sentiment140	
Proportional	$p_i \propto N_i$	0.61 (MTL)	0.59 (STL)	0.87 (MTL)	0.85 (STL)
Square root sampling	$p_i \propto N_i^{\alpha}$	0.62 (MTL)	0.59 (STL)	0.87 (MTL)	0.85 (STL)
Annealed Sampling	$\alpha = 1 - 0.8\frac{e-1}{E-1}$	**0.63** (MTL)	0.59 (STL)	0.86 (MTL)	0.85 (STL)

GoEmotion Taxonomy

- Hierarchical grouping of emotion labels.
- **Original**: 27 categories
- **Ekman**: 6 categories

Split	train (43,410), test (5,427), dev (5,426)
Labels (Original)	admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise and neutral
Sentiment taxonomy	*positive, negative, ambiguous* and *neutral*
Ekman taxonomy	*Neutral* label & 6 groups: *anger* (anger, annoyance, disapproval), *disgust* (disgust), *fear* (fear, nervousness), *joy* (all positive emotions), *sadness* (adness, disappointment, embarrassment, grief, remorse) and *surprise* (all ambiguous emotions)
Number of labels per example	1: 83%, 2: 15%, 3: 2%, 4+: 0.2% (rounded)

Table 1: Goemotion Dataset Summary

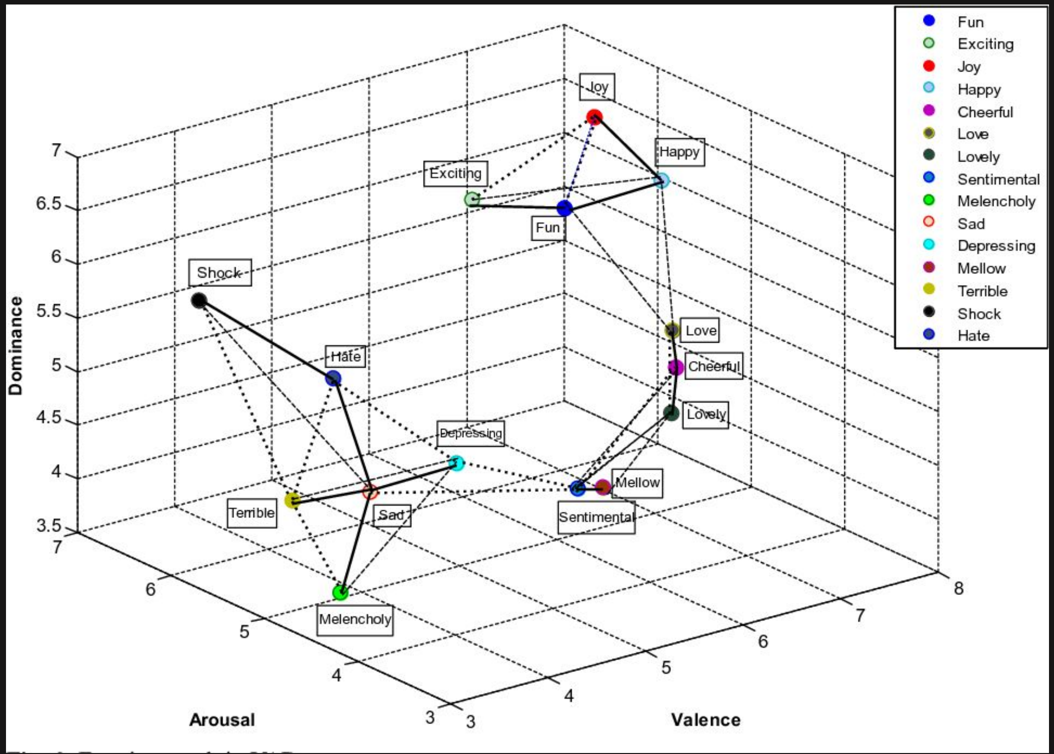
Task Sampling/Batching

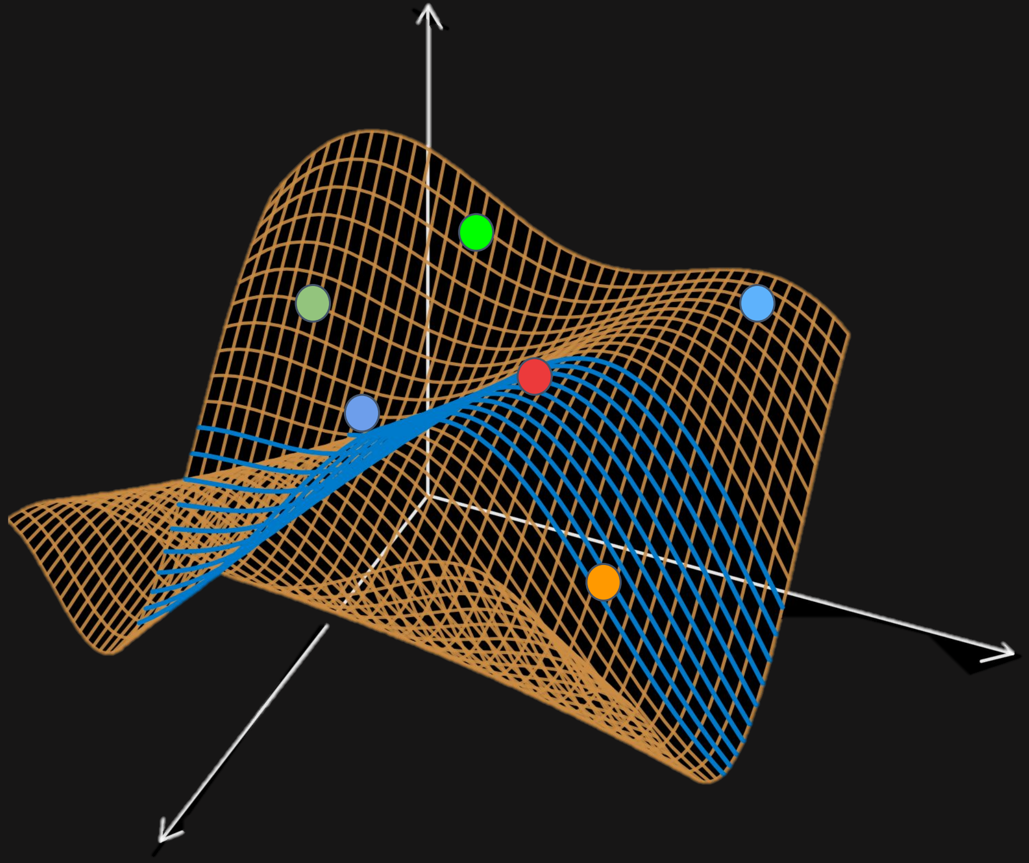
- Both tasks of similar/same training dataset size.
- Accuracies for MTL (Multi-Task learning) & STL (Single Task Learning)

Sampling Method	Task-1 **GoEmotion Original**	Task-2 **GoEmotion Ekman**
Proportional $p_i \propto N_i$	0.61 (MTL) 0.59 (STL)	0.69 (MTL) 0.68 (STL)
Square root sampling $p_i \propto N_i^\alpha$	0.60 (MTL) 0.59 (STL)	0.70 (MTL) 0.68 (STL)
Annealed Sampling $\alpha = 1 - 0.8\frac{e-1}{E-1}$	0.61 (MTL) 0.59 (STL)	0.70 (MTL) 0.68 (STL)

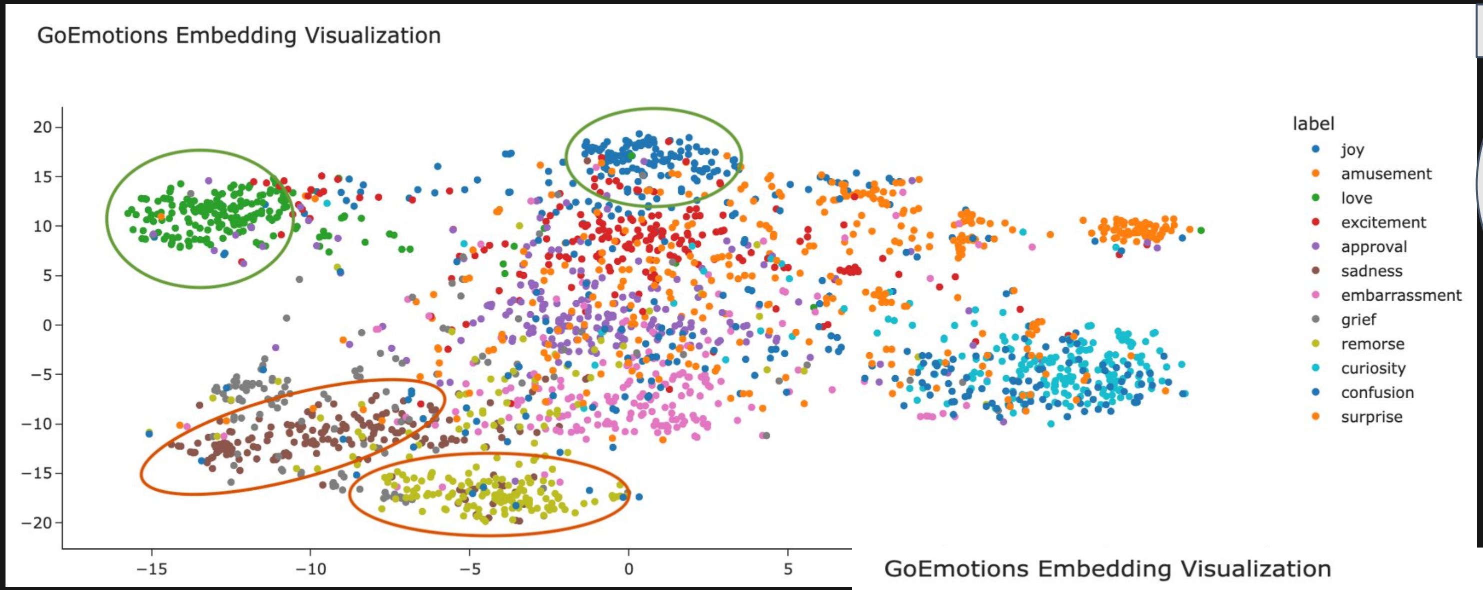
Contrastive-learning type effect!

- Emotion: VAD (Valence, Arousal, Dominance) space.
- In Embedding (sentence representation) space, similar VAD emotion classes ***may or may not*** be together!



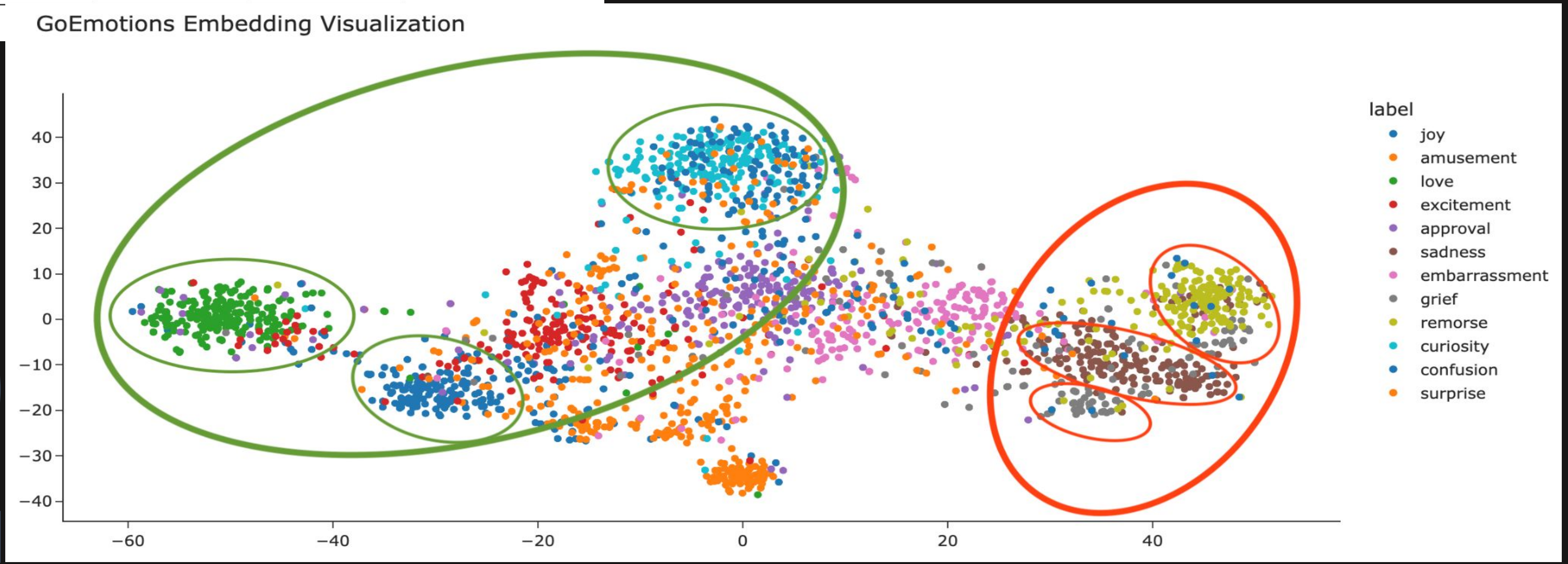


Contrastive-learning type effect!



Single Task Setting

Positive and negative emotions are inter-mixed together.



Positive and negative emotions are grouped separately.

Multi Task Setting

Conclusion

Task Weighting	Depending on task importance, task difficulty level etc
Auxiliary Task Selection	Selection of similar and related tasks could help get better representation bias
Contrastive-like effect (hierarchical tasks in MTL setting)	Training hierarchical tasks together can implicitly enforce the hierarchy on the embedding manifold.
Task Sampling	When training on tasks with skewed dataset sizes, annealed sampling should be helpful.