07/12/2022

Imagination-Augmented Natural Language Understanding

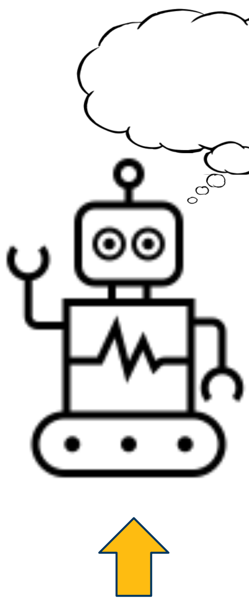
Yujie Lu, Wanrong Zhu, Xin Eric Wang, Miguel Eckstein, William Yang Wang





How Do Humans Understand Natural Language?

Visual Imagination

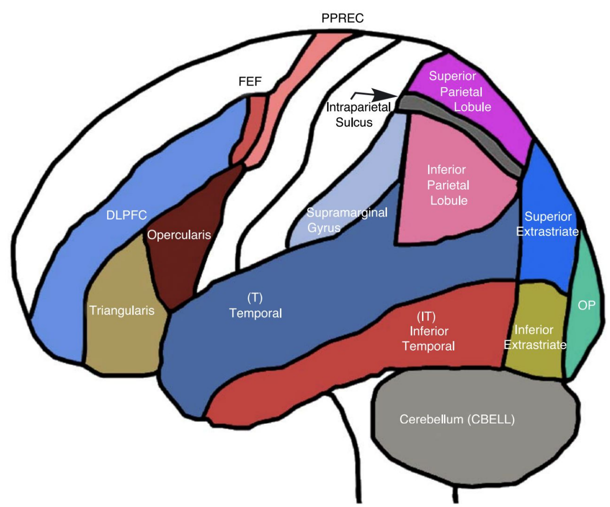


A senior is waiting at the window of a restaurant that serves sandwiches.

Background in Cognitive Neuroscience

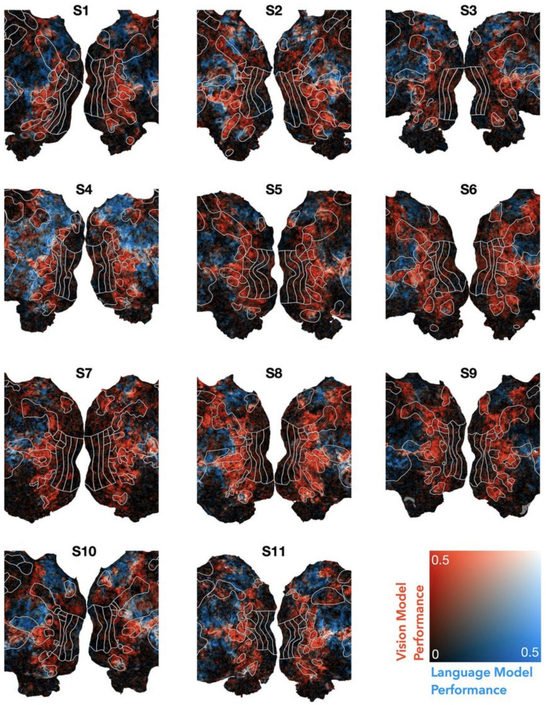
- **Imagery in Sentence Comprehension**

Neural activation in vision-related brain areas when reading texts (Marcel et al., 2004)



- **Visual and linguistic**

semantic representations are aligned at the border of human visual cortex (Sara et al., 2021)

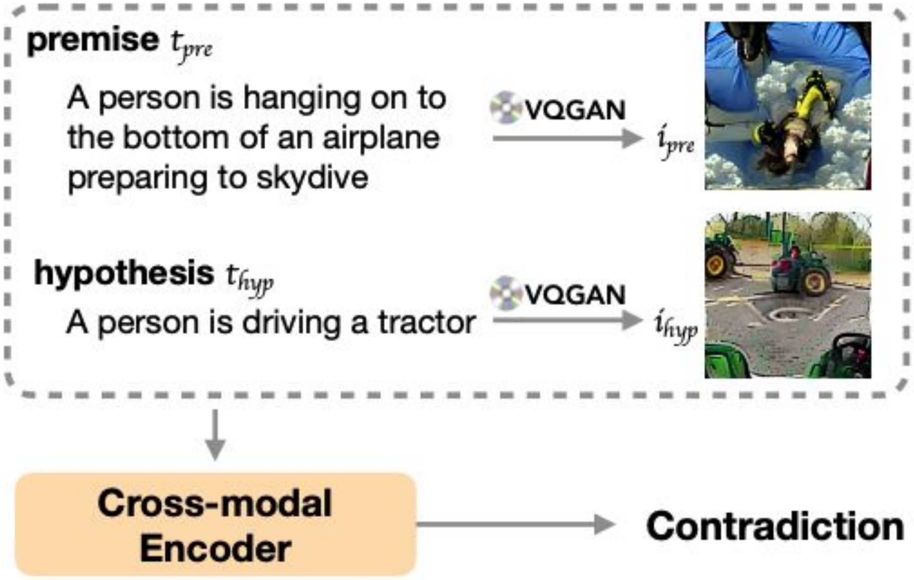


- Visual imagery improves comprehension during human language processing. (Mark et al., 1994)

How Does Visual Supervision Help NLU?

Such **imagination** empowers human brains with **generalization** capability to solve problems with **limited supervision or data samples**.

- Pure-language based
- No explicit visual supervision in downstream tasks



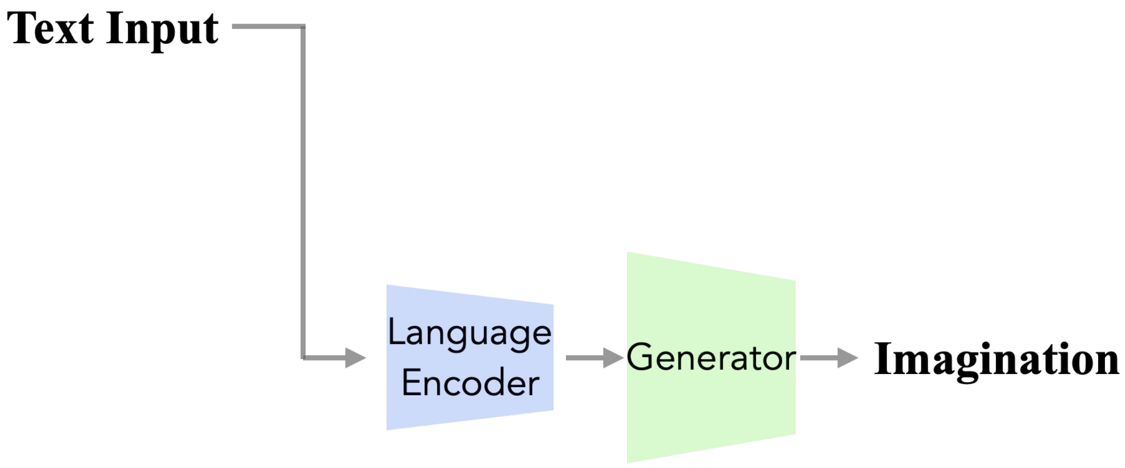
Generating Images or Retrieving Images?

Down by the salley gardens my love and I did meet

Down by the salley gardens my love and I did meet

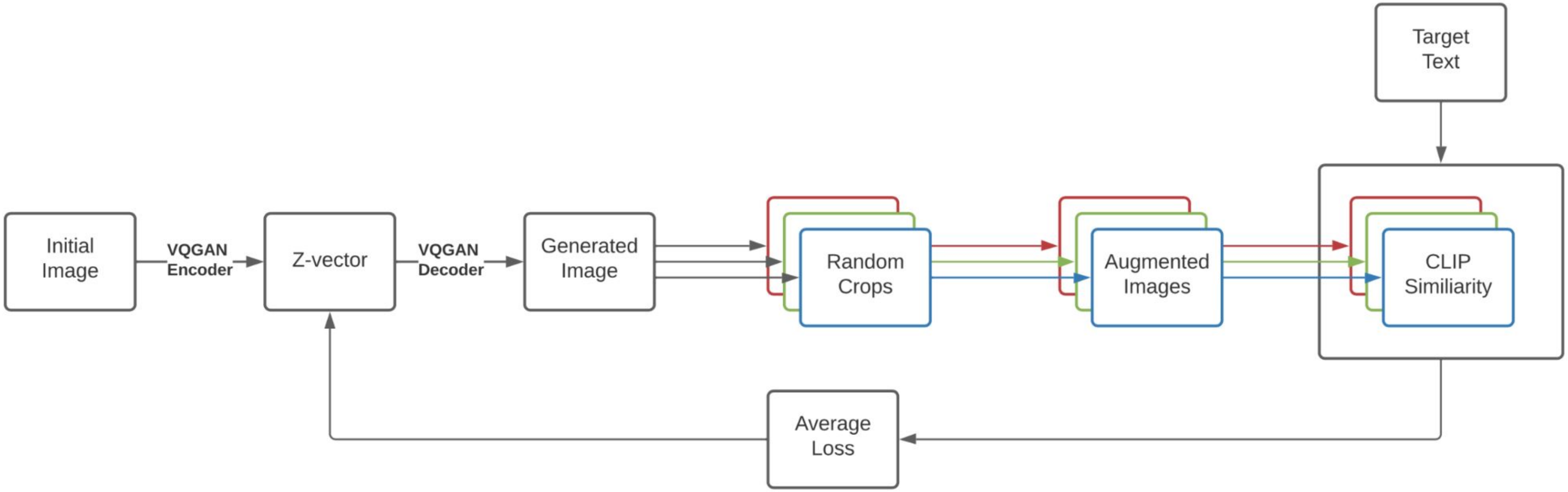
Architecture

Imagination-Augmented NLU



Imagination Generator

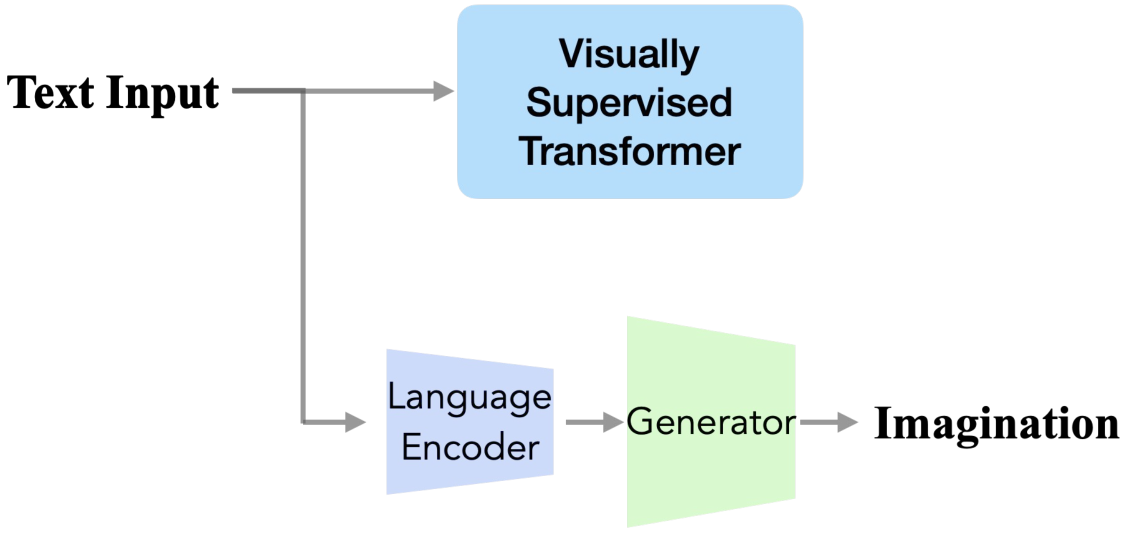
Generating Semantically Relevant Imagery



Katherine et al. 2021. VQGAN-CLIP: Open Domain Image Generation and Editing with Natural Language Guidance

Architecture

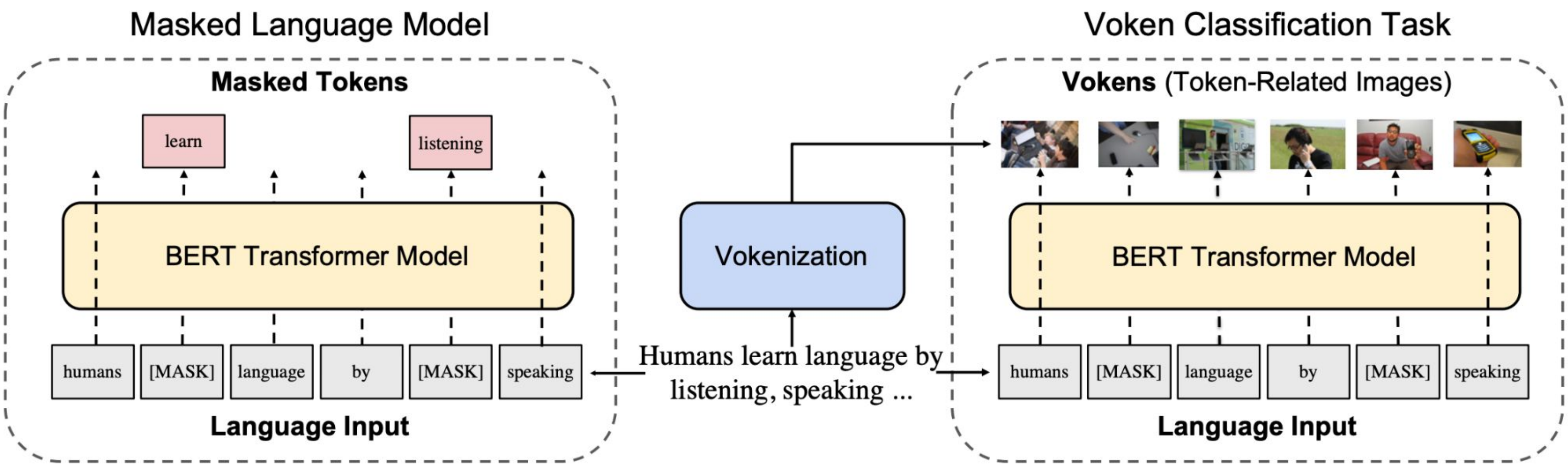
Imagination-Augmented NLU



Visually Supervised Transformer

Pre-training Language Model with Visual Supervision

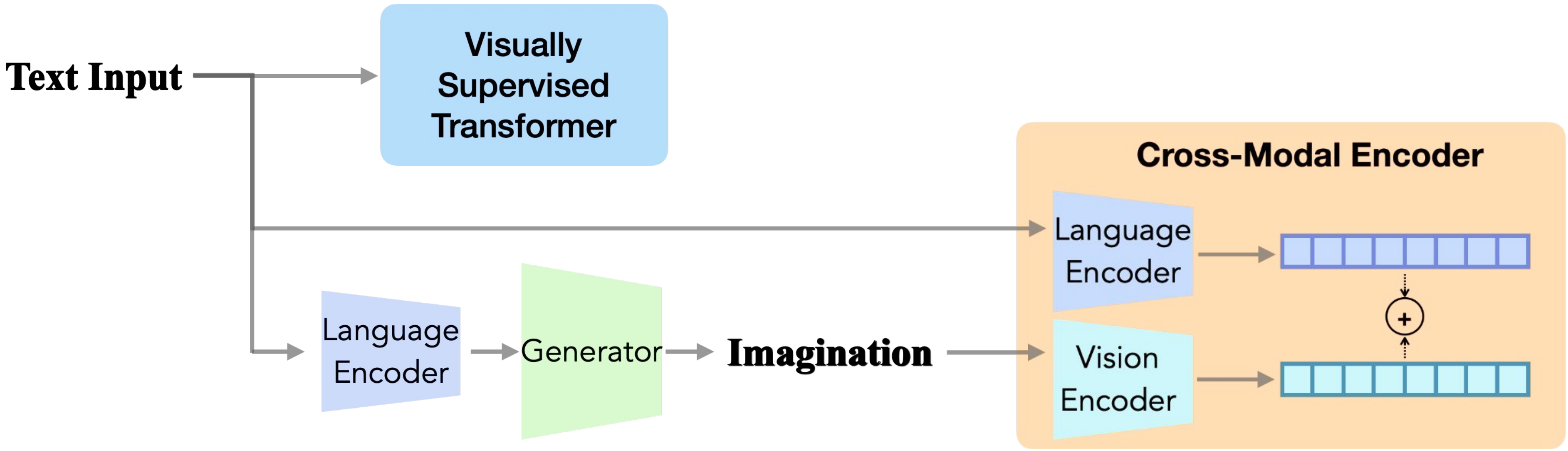
- BERT-like pure-language based masked language model



Hao et al. 2020. Vokenization: Improving Language Understanding with Contextualized, Visual-Grounded Supervision

Architecture

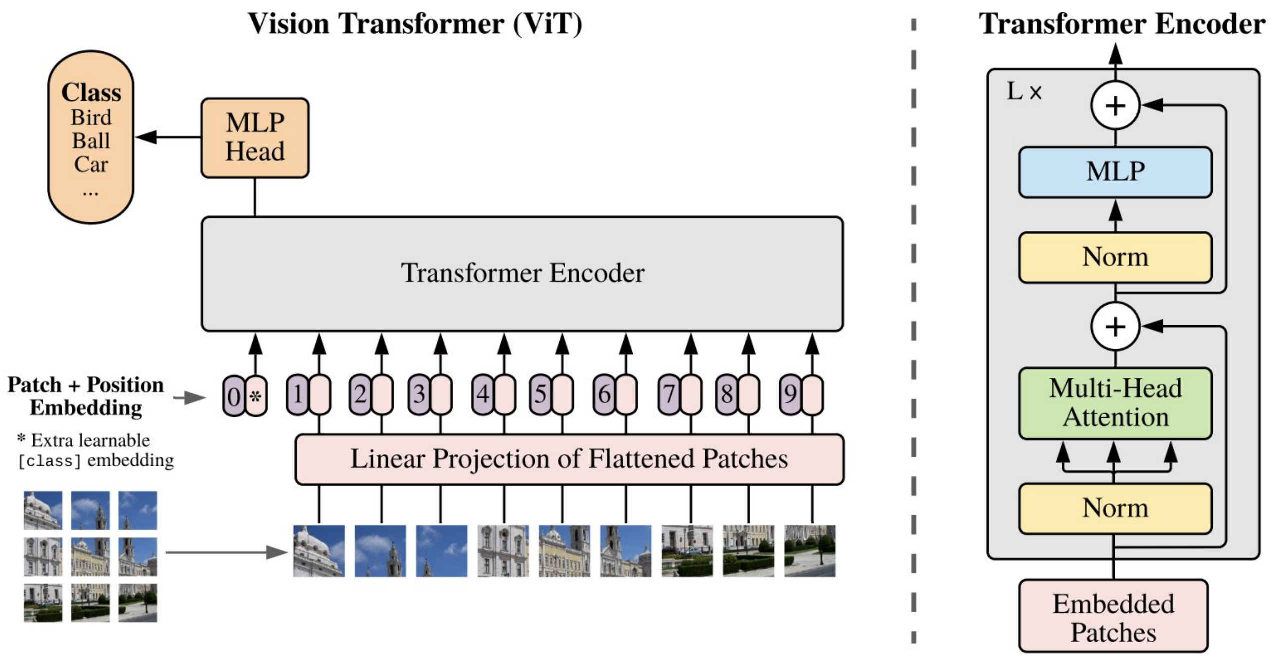
Imagination-Augmented NLU



Cross-modal Encoder

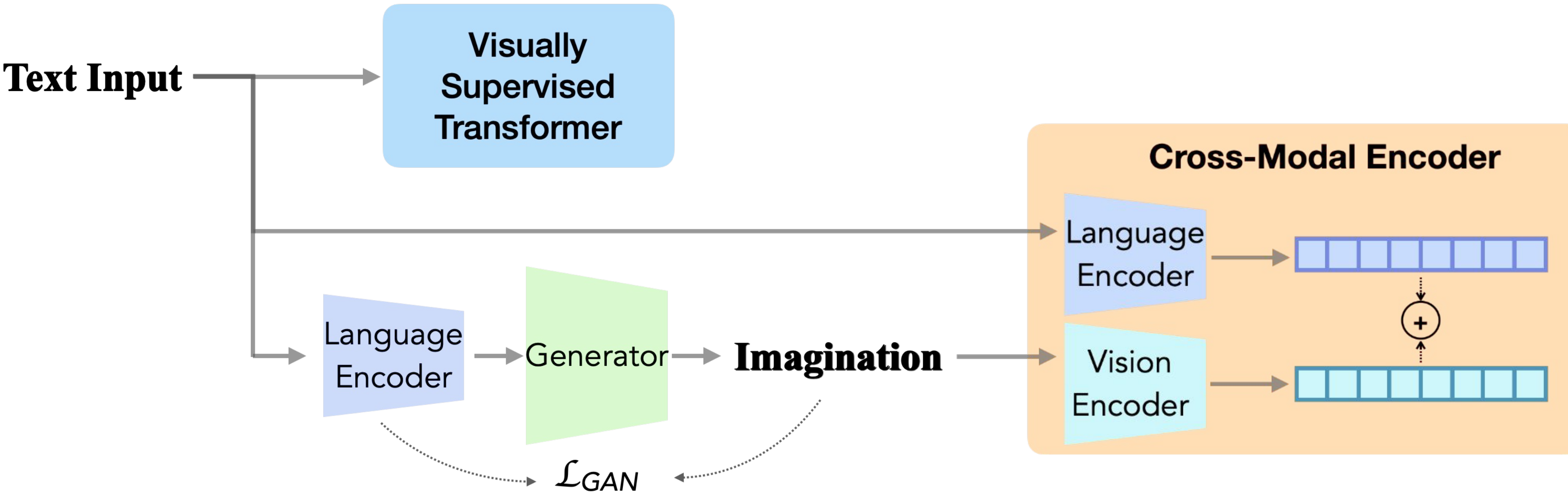
Imagination-Augmented Language Representation

- Vision Encoder: Vision Transformer (Alexey et al., 2020)
- Language Encoder: Transformer (Vaswani et al., 2017;Radford et al., 2019)



Learning Procedure

Imagination Construction



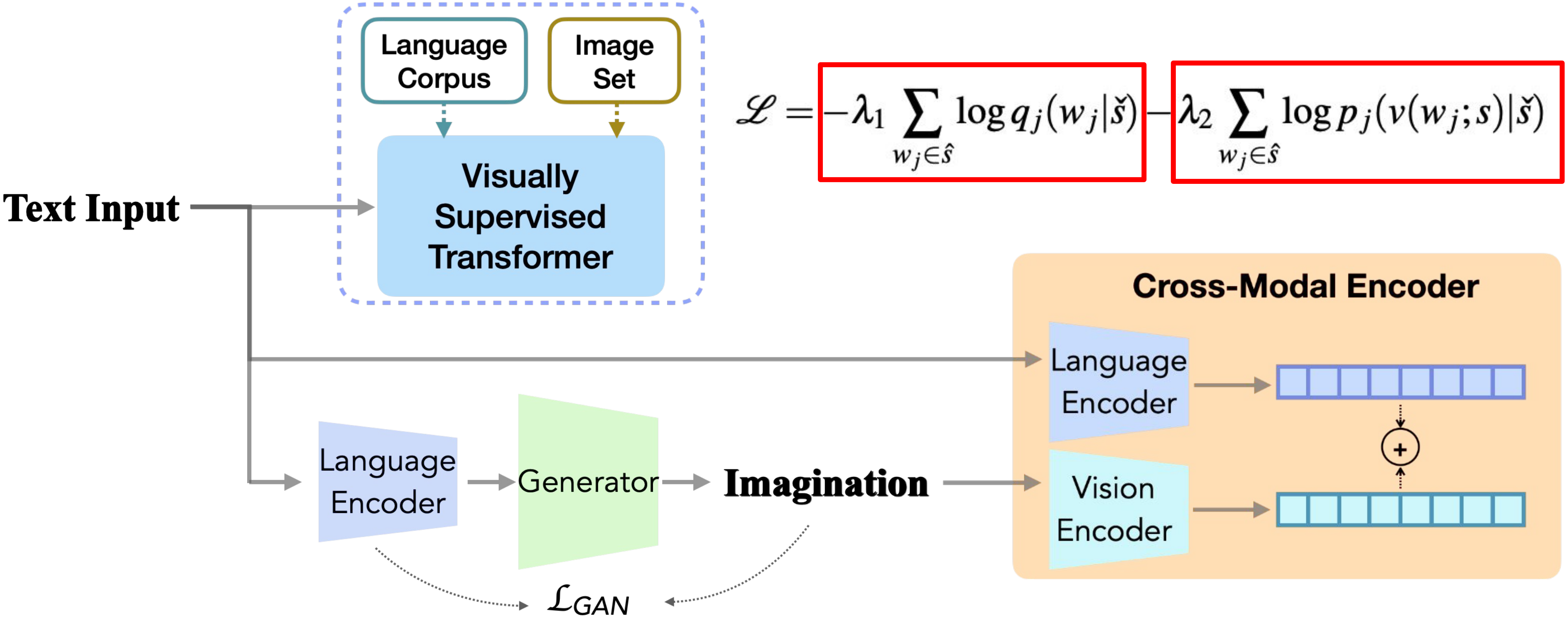
$$\boxed{\mathcal{L}_{GAN} = 2[\arcsin(\frac{1}{2}\|\boldsymbol{t}-\boldsymbol{v}\|)]^2}$$

Learning Procedure

Visually Supervised Pre-training

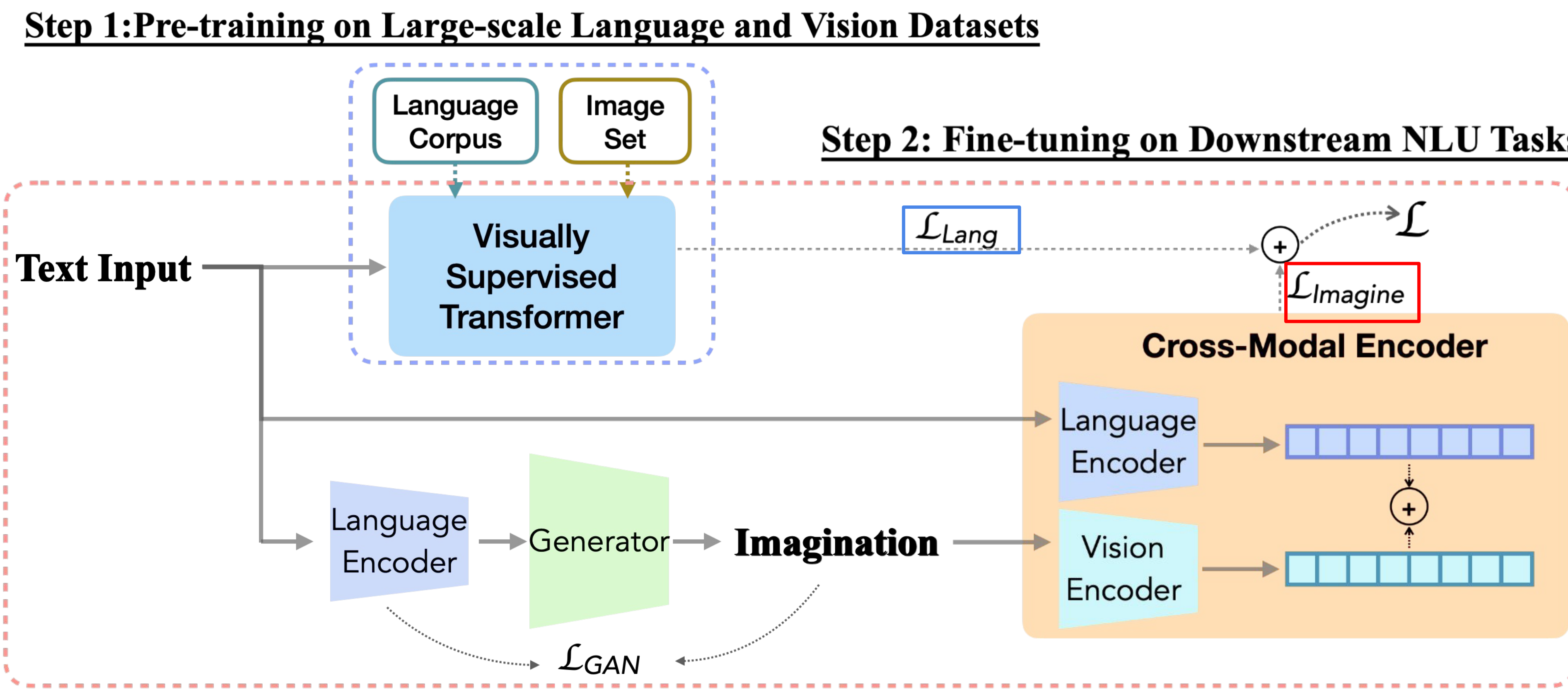
Step 1:Pre-training on Large-scale Language and Vision Datasets

$$\mathcal{L} = \boxed{-\lambda_1 \sum_{w_j \in \hat{s}} \log q_j(w_j|\check{s})}\boxed{-\lambda_2 \sum_{w_j \in \hat{s}} \log p_j(v(w_j;s)|\check{s})}$$



Learning Procedure

Incorporating Downstream Tasks with Visual Imagination.



$$\mathcal{L}_{Lang} = -\sum_{j=1}^{|D|}\sum_{k=1}^{K} y_k \log p_k(d_j(\boldsymbol{t})|D)$$

$$\mathcal{L}_{Imagine} = -\sum_{j=1}^{|D|}\sum_{k=1}^{K} y_k \log p_k(d_j(\boldsymbol{t};\boldsymbol{v})|D)$$

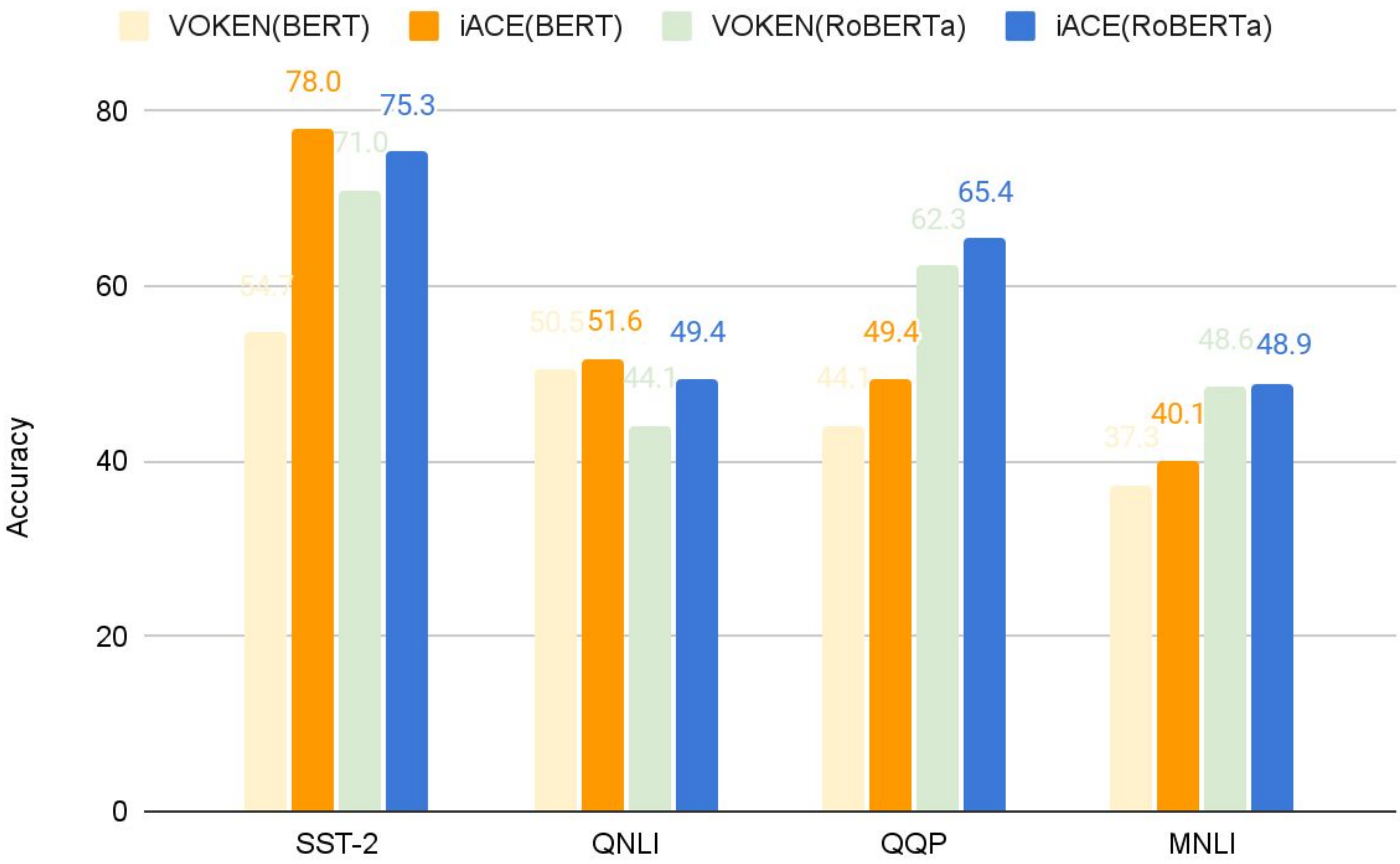
Experiment Setup

Datasets, Metrics, Baselines

- **Datasets**
  - GLUE (SST-2, QNLI, QQP, MNLI, MRPC, STS-B), SWAG
    - Sentiment Analysis
    - Paraphrase
    - Natural Language Inference
    - Commonsense Inference
  - Few-shot Setting: 0.1%, 0.3%, 0.5%, 1%, 3%, 5% of instances
- **Metrics**
  - Accuracy, F1
- **Baselines**
  - Textual-Only: BERT, RoBERTa
  - Visual-Only: CLIP
  - Visually-supervised language model: Vokenization (Tan, 2020)

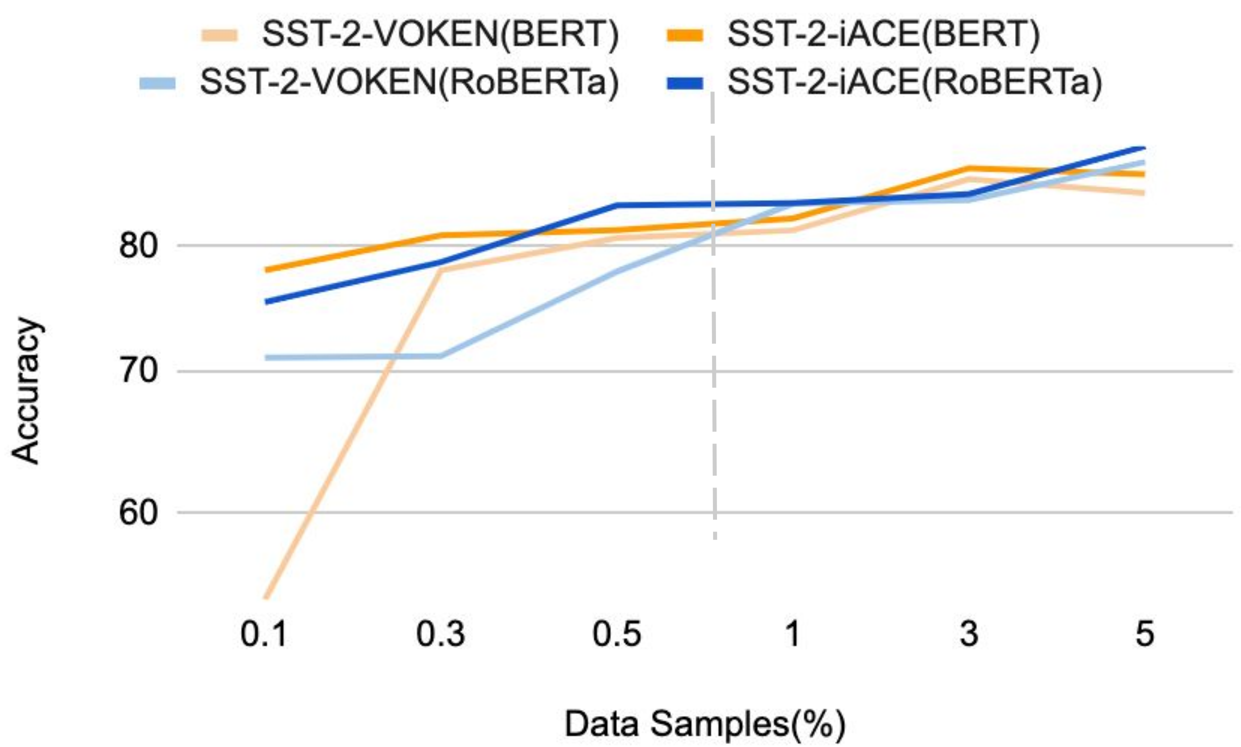
Performance with Limited Samples

How do we perform in the few-shot setting?



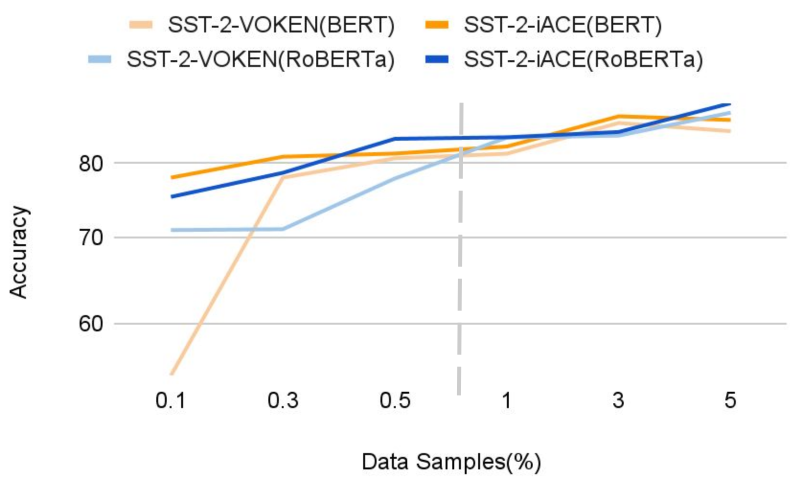
Data Samples

SST-2

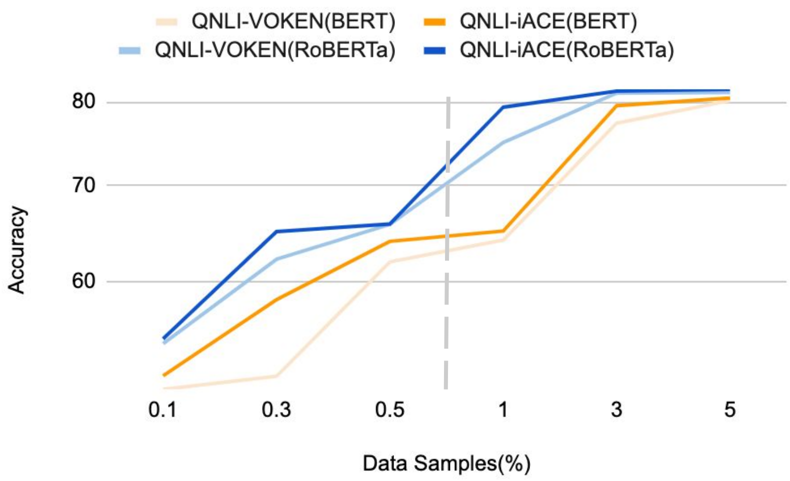


Data Samples

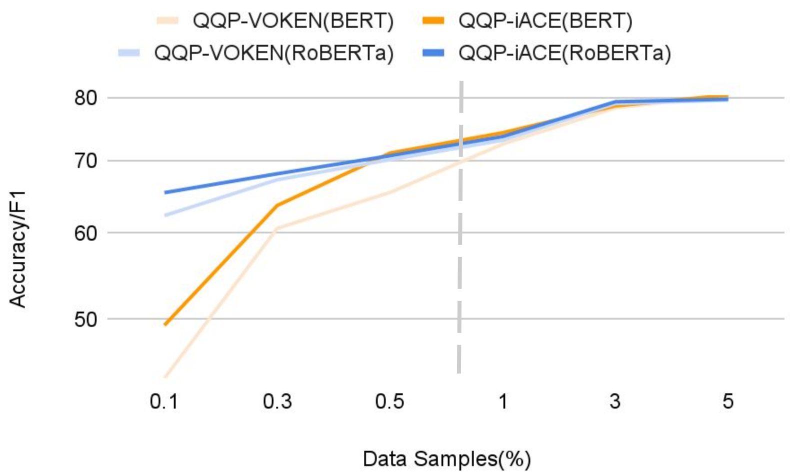
SST-2



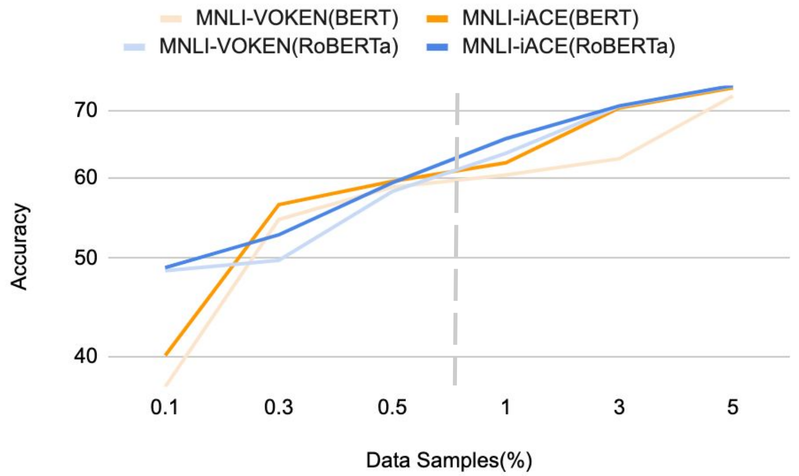
QNLI



QQP

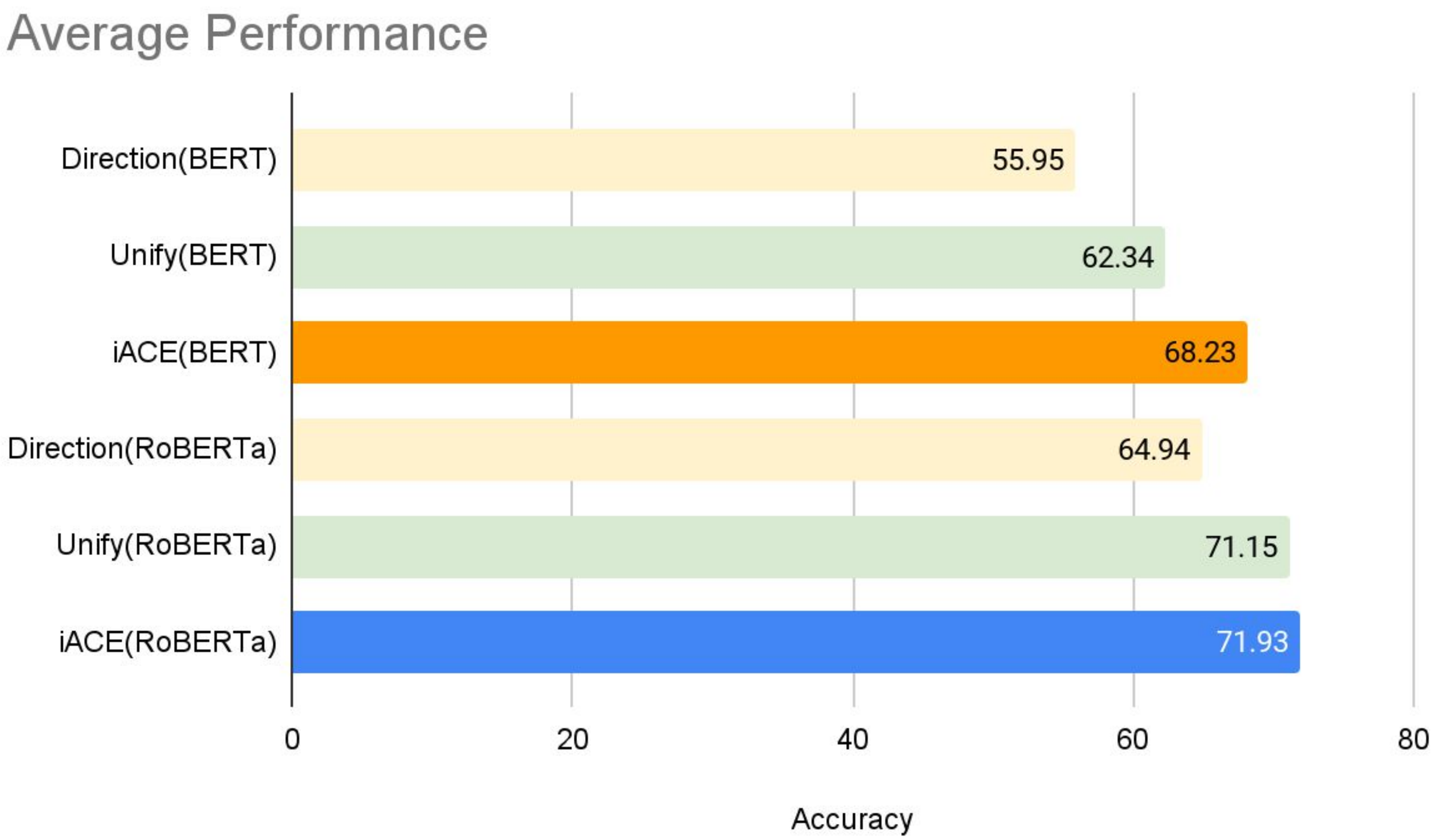


MNLI



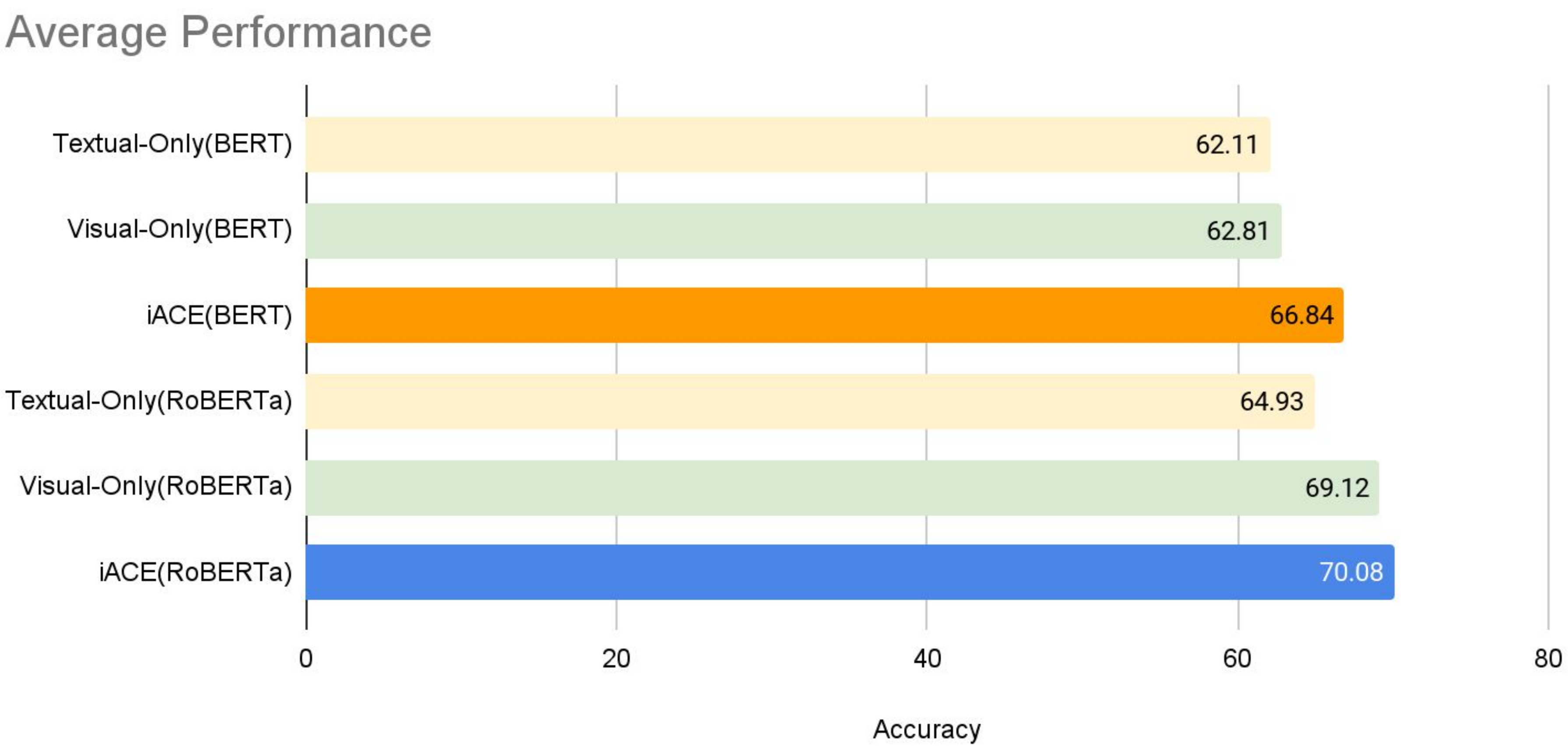
Method Ablation

Is the imagination incorporated correctly?



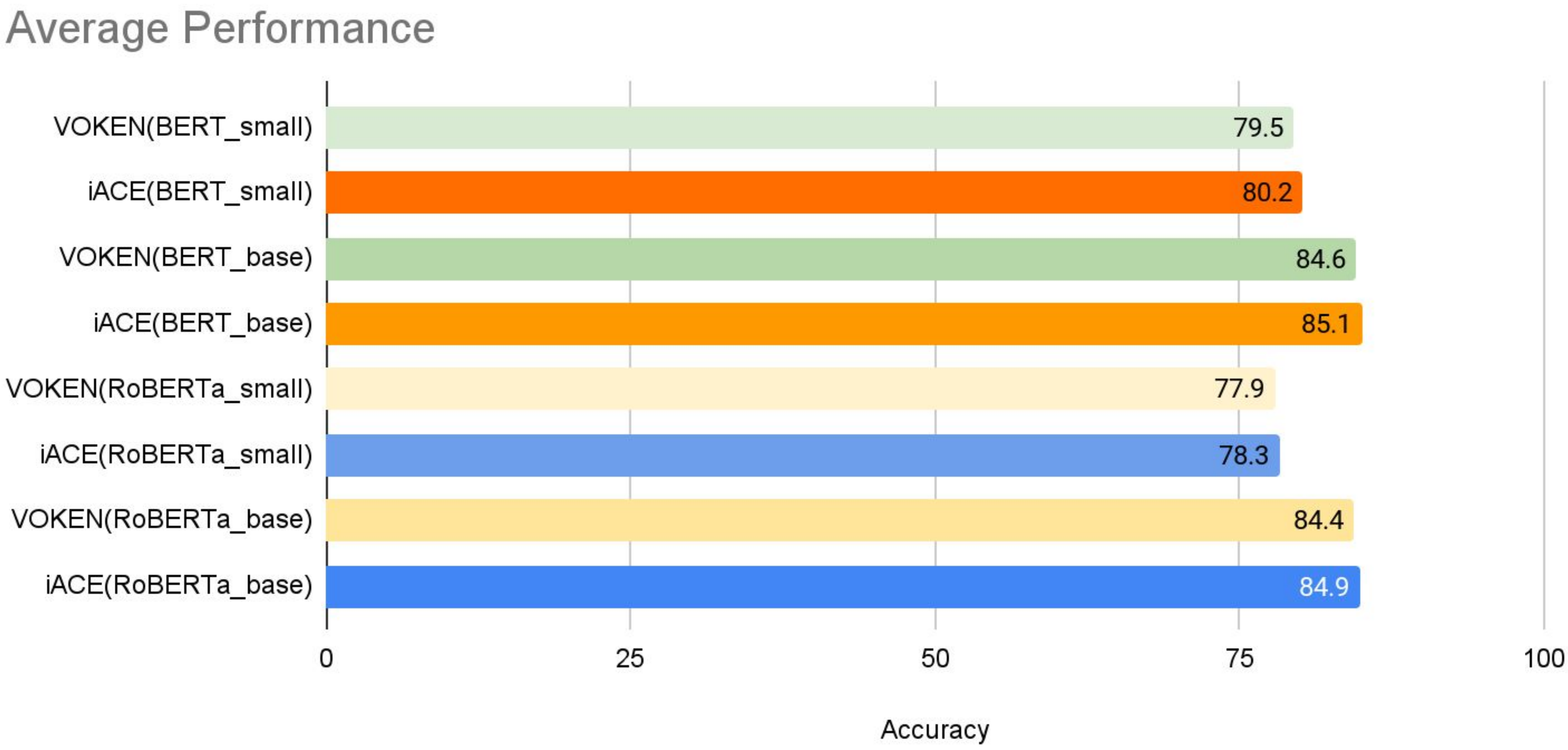
Composition Ablation

Is the imagination modality helpful?



Performance on Full Data

How do we perform in the full data setting?



Case Study

In what cases do visual modality help?

Text 1: The lady cracked an egg for the mixer.
Text 2: The lady sliced up the meat.

Ground Truth: 3.75 / 5.00

Img 1

Img 2

Baseline: 4.14 / 5.00 ❌

Ours: 3.62 / 5.00 ✅

Text 1: A woman is riding a horse along a perimeter.
Text 2: A dog is riding a skateboard.

Ground Truth: 0.60 / 5.00

Img 1

Img 2

Baseline: 0.31 / 5.00 ❌

Ours: 0.58 / 5.00 ✅

Limitation: Abstract-level language understanding

Conclusion

- Bridging the gap between human and model in natural language understanding by leveraging visual imagination.
- Eliciting visual supervision from the pre-trained generative and the vision-language models in downstream tasks.
- Achieving consistent performance boost in general NLU, especially in low-resource situations.

Paper: https://arxiv.org/abs/2204.08535

Repo: https://github.com/YujieLu10/IACE-NLU

THANK YOU

Q & A

Contact

Twitter: @yujielu_10

Email: yujielu@ucsb.edu

TODO

- ~~Split architecture in animation~~
- ~~Add references~~
- **Add detail of ablated method**
- **Replace Module Slides with our own contribution**
- ~~Add animation to all tables~~
- **Using Widescreen?**
  - **Figure it out on Sunday**