# 智能外呼机器人

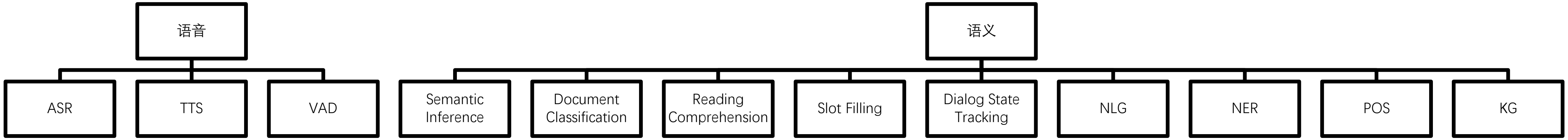
# 智能语音助手

# 校园助手

# 智能交互背后的TensorFlow

*回答你问题的不一定都是小姐姐，还有可能是 TensorFlow*

- 交互背后的技术

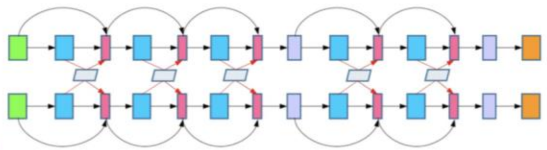

语音
ASR
TTS
VAD
语义
Semantic Inference
Document Classification
Reading Comprehension
Slot Filling
Dialog State Tracking
NLG
NER
POS
KG

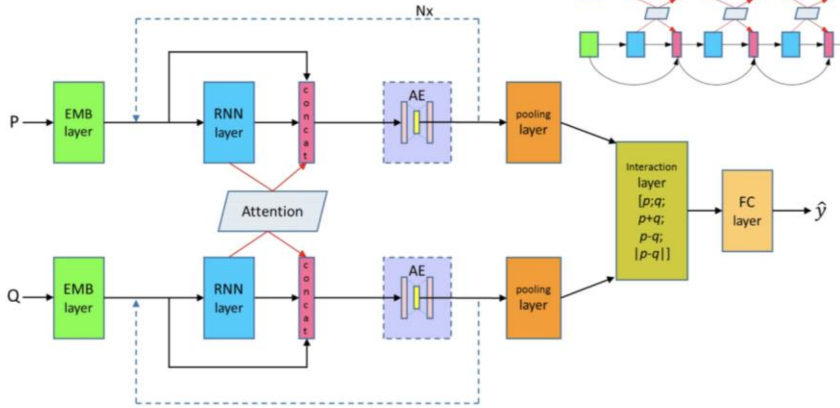
# NLI

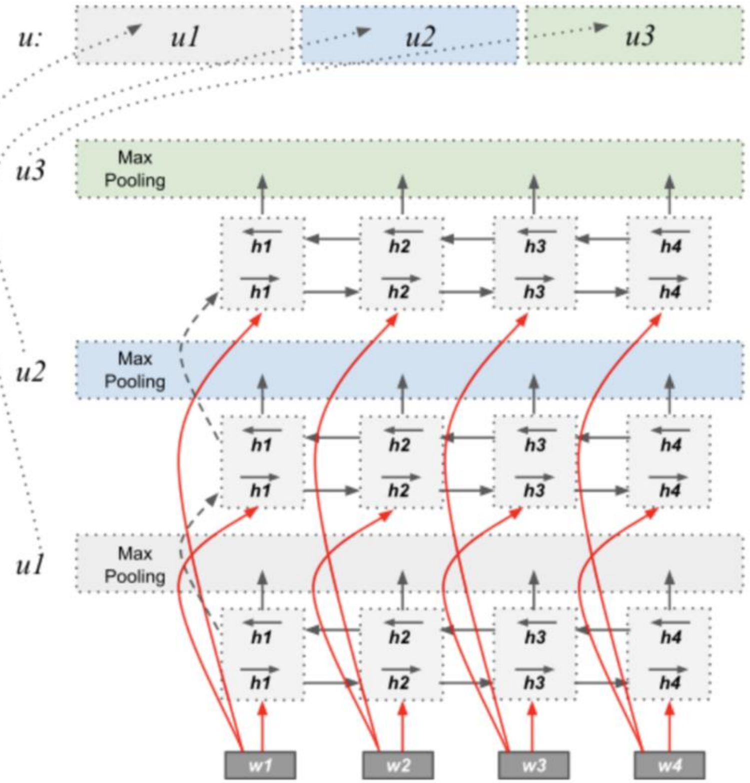
| Premise | Label | Hypothesis |
|---|---|---|
| A man inspects the uniform of a figure in some East Asian country. | contradiction | The man is sleeping. |
| An older and younger man smiling. | neutral | Two men are smiling and laughing at the cats playing on the floor. |
| A soccer game with multiple males playing. | entailment | Some men are playing a sport. |

# NLI

- Premise与Hypothesis之间的交互方式是关键
  - Tay, Y., Tuan, L. A., & Hui, S. C. (2017)采用cross attention + alignment factorization的方式;
  - Kim, S., Kang, I., & Kwak, N. (2018)使用多层RNN多次cross attention的方式，借鉴Densenet的思想尽可能不对信息作压缩，利用bottleneck结构的autoencoder保持网络大小;
  - Talman, A., Yli-Jyrä, A., & Tiedemann, J. (2018)采用多层RNN和max pooling over time的方式在SciTail上也获得了不错的效果

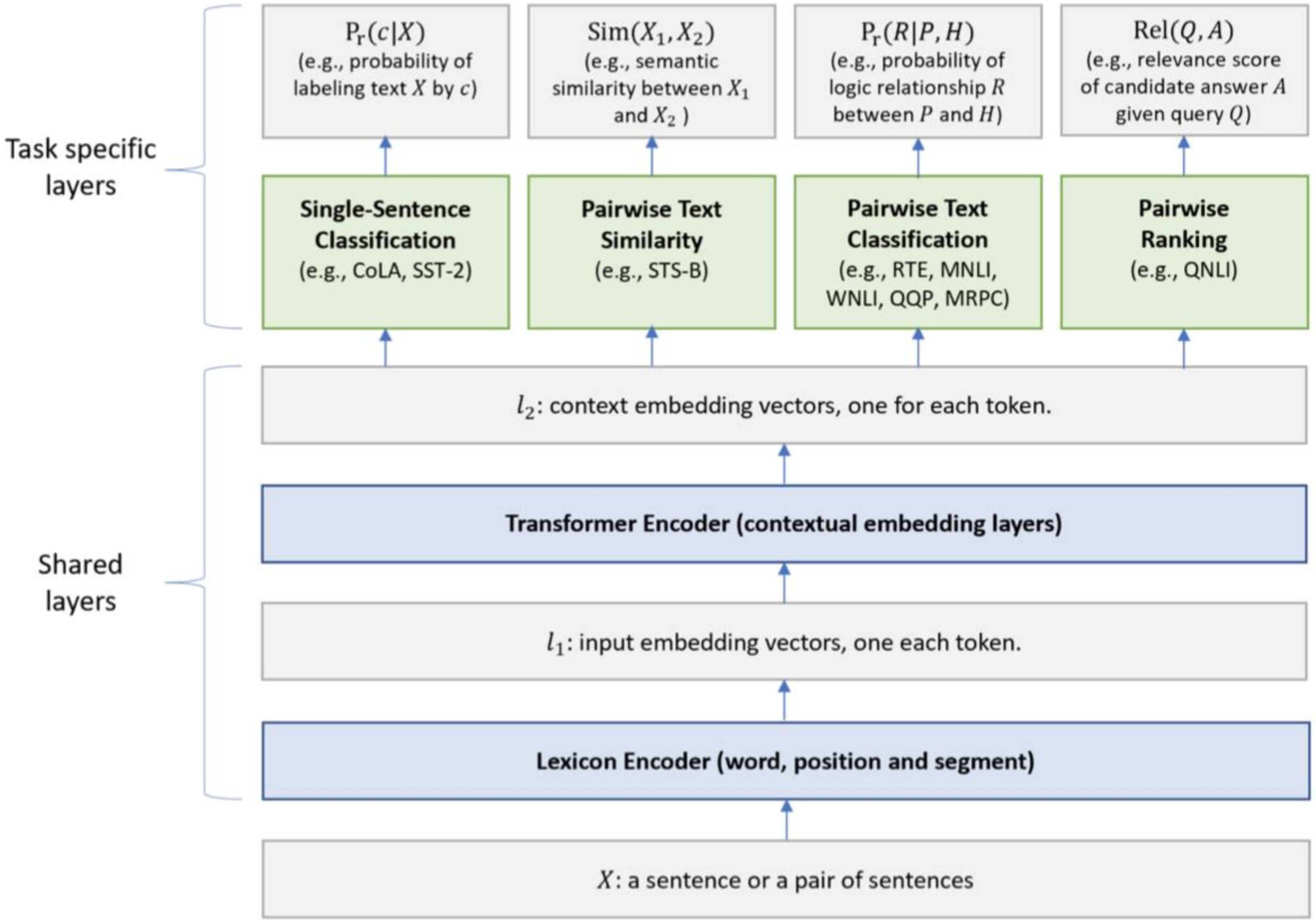




# NLI

- Pre-train model, BERT
- Multi-task, Xiaodong Liu, P. H., Weizhu Chen, Jianfeng Gao



# NLI

- 数据集中的“人为偏差”， Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., & Smith, N. A. (2018)

| **Model** | **SNLI** | **MultiNLI** | |
|---|---|---|---|
| | | **Matched** | **Mismatched** |
| majority class | 34.3 | 35.4 | 35.2 |
| `fastText` | **67.0** | **53.9** | **52.3** |

| **Premise** | A woman selling bamboo sticks talking to two men on a loading dock. |
|---|---|
| **Entailment** | There are **at least** three **people** on a loading dock. |
| **Neutral** | A woman is selling bamboo sticks **to help provide for her family.** |
| **Contradiction** | A woman is **not** taking money for any of her sticks. |

Table 1: An instance from SNLI that illustrates the artifacts that arise from the annotation protocol. A common strategy for generating entailed hypotheses is to remove gender or number information. Neutral hypotheses are often constructed by adding a purpose clause. Negations are often introduced to generate contradictions.

# NLI

- SWAG数据集， Rowan Zellers, Y. B., Roy Schwartz, Yejin Choi (2018)
- 采用简单分类器多次过滤的方法来得到对抗样本。

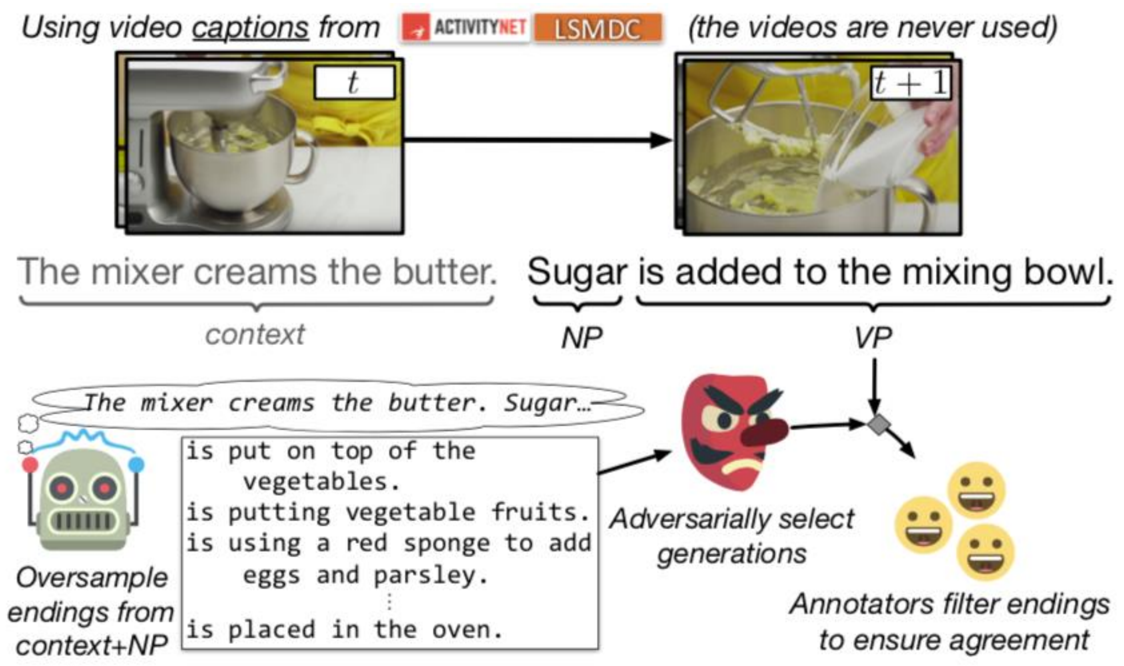


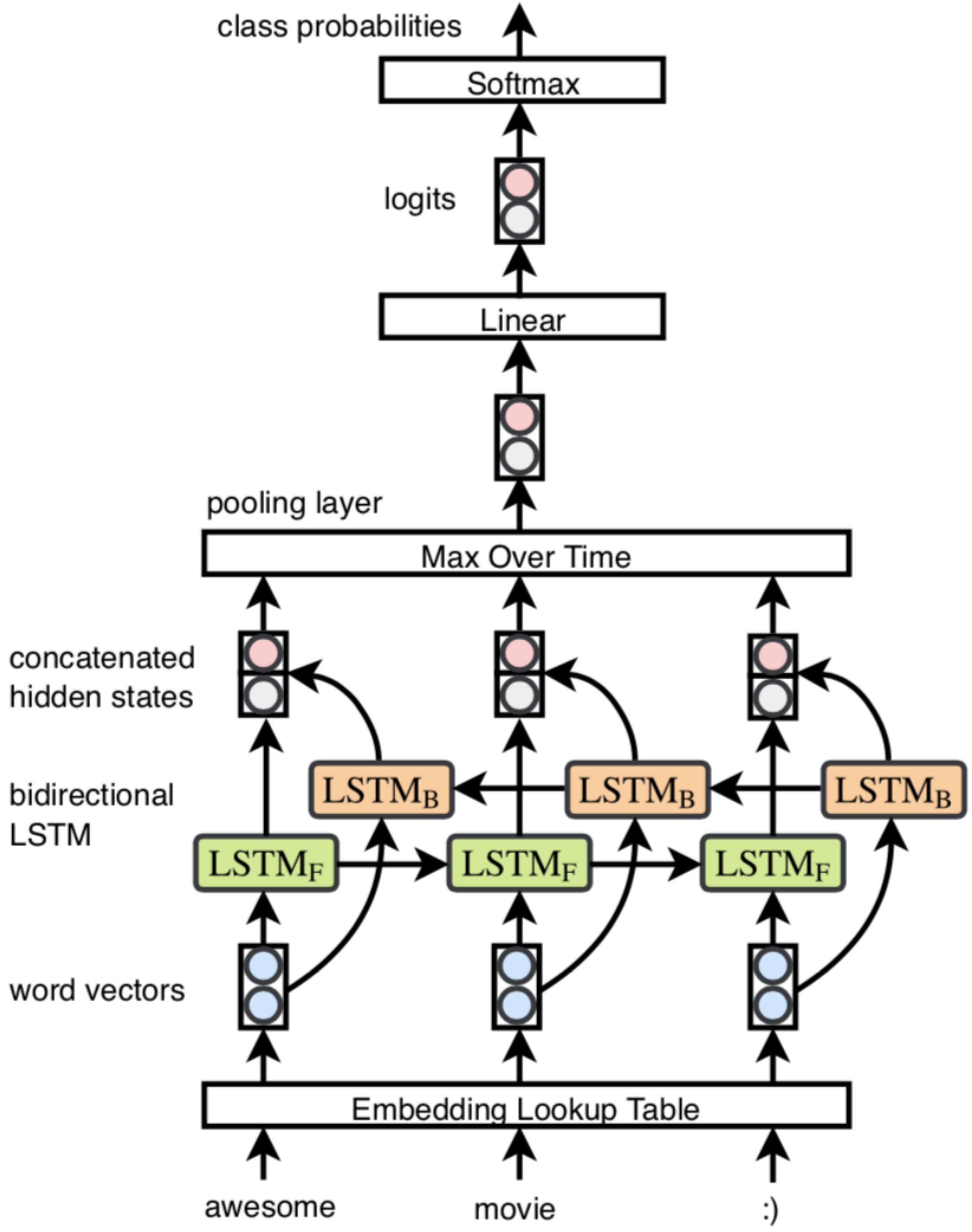

Figure 1: Overview of the data collection process. For a pair of sequential video captions, the second caption is split into noun and verb phrases. A language model generates many negative endings, of which a difficult subset are human-annotated.

# NLI

| | | Model | Ending only | | | | 2nd sentence only | | | | Context+2nd sentence | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | found only | | found+gen | | found only | | found+gen | | found only | | found+gen | |
| | | | Val | Test | Val | Test | Val | Test | Val | Test | Val | Test | Val | Test |
| | misc | Random | 25.0 | 25.0 | 25.0 | 25.0 | 25.0 | 25.0 | 25.0 | 25.0 | 25.0 | 25.0 | 25.0 | 25.0 |
| | misc | Length | 26.7 | 27.0 | 26.7 | 27.0 | | | | | | | | |
| | misc | ConceptNet | | | | | 26.0 | 26.0 | 26.0 | 26.0 | | | | |
| Unary models | | fastText | 27.5 | 26.9 | 29.9 | 29.0 | 29.2 | 27.8 | 29.8 | 29.0 | 29.4 | 28.0 | 30.3 | 29.8 |
| Unary models | Sentence encoders | SkipThoughts | 32.4 | 32.1 | 32.2 | 31.8 | 33.0 | 32.4 | 32.8 | 32.3 | | | | |
| Unary models | Sentence encoders | InferSent | 30.6 | 30.2 | 32.0 | 31.9 | 33.2 | 32.0 | 34.0 | 32.6 | | | | |
| Unary models | LSTM sequence model | LSTM+GloVe | 31.9 | 31.8 | 32.9 | 32.4 | 32.7 | 32.4 | 34.3 | 33.5 | 43.1 | 43.6 | 45.6 | 45.7 |
| Unary models | LSTM sequence model | LSTM+Numberbatch | 32.4 | 32.6 | 32.3 | 31.9 | 31.9 | 31.9 | 34.1 | 32.8 | 39.9 | 40.2 | 41.2 | 40.5 |
| Unary models | LSTM sequence model | LSTM+ELMo | **43.6** | **42.9** | **43.3** | **42.3** | **47.4** | **46.7** | **46.3** | **46.0** | 51.4 | 50.6 | 51.3 | 50.4 |
| Binary models | DualBoW | DualBoW+GloVe | | | | | 31.3 | 31.3 | 31.9 | 31.2 | 34.5 | 34.7 | 32.9 | 33.1 |
| Binary models | DualBoW | DualBoW+Numberbatch | | | | | 31.9 | 31.4 | 31.6 | 31.3 | 35.1 | 35.1 | 34.2 | 34.1 |
| Binary models | Dual sentence encoders | SkipThoughts-MLP | | | | | 34.6 | 33.9 | 36.2 | 35.5 | 33.4 | 32.3 | 37.4 | 36.4 |
| Binary models | Dual sentence encoders | SkipThoughts-Bilinear | | | | | 36.0 | 35.7 | 34.7 | 34.5 | 36.5 | 35.6 | 35.3 | 34.9 |
| Binary models | Dual sentence encoders | InferSent-MLP | | | | | 32.9 | 32.1 | 32.8 | 32.7 | 35.9 | 36.2 | 39.5 | 39.4 |
| Binary models | Dual sentence encoders | InferSent-Bilinear | | | | | 32.0 | 31.3 | 31.6 | 31.3 | 40.5 | 40.3 | 39.0 | 38.4 |
| Binary models | SNLI inference | SNLI-ESIM | | | | | | | | | 36.4 | 36.1 | 36.2 | 36.0 |
| Binary models | SNLI inference | SNLI-DecompAttn | | | | | | | | | 35.8 | 35.8 | 35.8 | 35.7 |
| Binary models | SNLI models (retrained) | DecompAttn+GloVe | | | | | 29.8 | 30.3 | 31.1 | 31.7 | 47.4 | 47.6 | 48.5 | 48.6 |
| Binary models | SNLI models (retrained) | DecompAttn+Numberbatch | | | | | 32.4 | 31.7 | 32.5 | 31.9 | 47.4 | 48.0 | 48.0 | 48.3 |
| Binary models | SNLI models (retrained) | DecompAttn+ELMo | | | | | 43.4 | 43.4 | 40.6 | 40.3 | 47.7 | 47.3 | 46.0 | 45.4 |
| Binary models | SNLI models (retrained) | ESIM+GloVe | | | | | 34.8 | 35.1 | 36.3 | 36.7 | 51.9 | 52.7 | 52.5 | 52.5 |
| Binary models | SNLI models (retrained) | ESIM+Numberbatch | | | | | 33.1 | 32.6 | 33.0 | 32.4 | 46.5 | 46.4 | 44.0 | 44.6 |
| Binary models | SNLI models (retrained) | ESIM+ELMo | | | | | 46.0 | 45.7 | 45.9 | 44.8 | **59.1** | **59.2** | **58.7** | **58.5** |
| | Human | 1 turker | | | | | | | | | 82.8 | | | |
| | Human | 3 turkers | | | | | | | | | 85.1 | | | |
| | Human | 5 turkers | | | | | | | | | **88.0** | | | |
| | Human | Expert | | | | | | | | | 85.0 | | | |

# Text Classification

- 除了Pre-train之外还能怎么玩?
- Devendra Singh Sachan, M. Z., Ruslan Salakhutdinov. (2019)采用最简单的结构, 但在loss上做文章, 采用监督与半监督训练、对抗训练一起进行的方式来达到不错的效果。



# NLG

- Gu, J., Liu, Q., & Cho, K. (2019)，用插入法生成，采用同时预测相对位置以及词来决定生成顺序

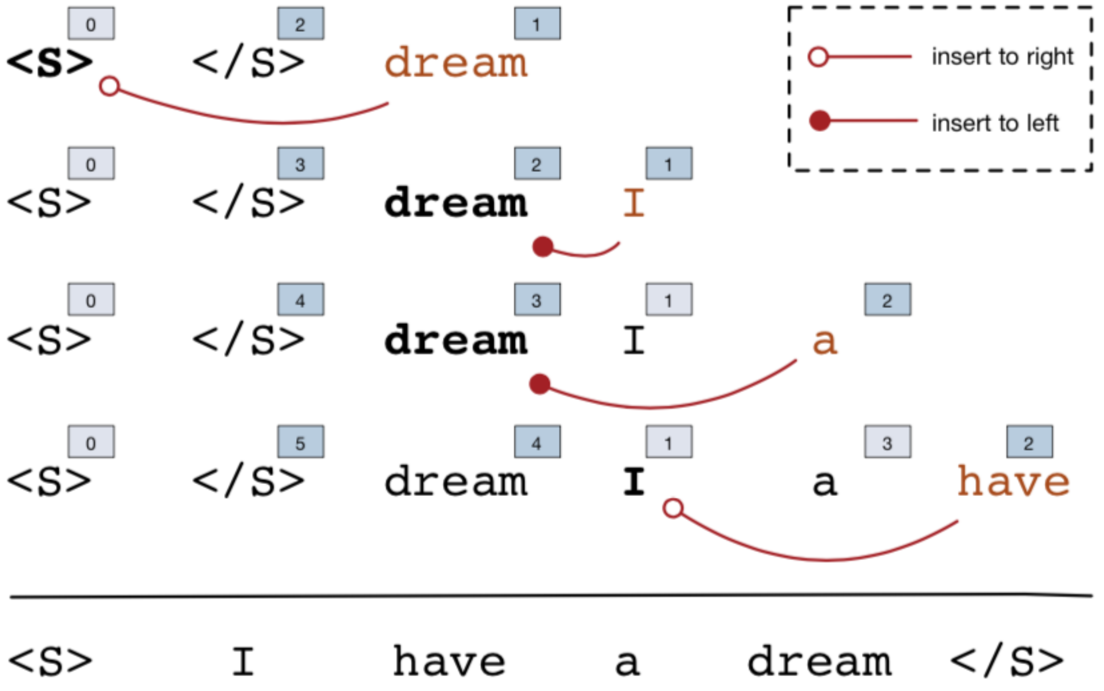
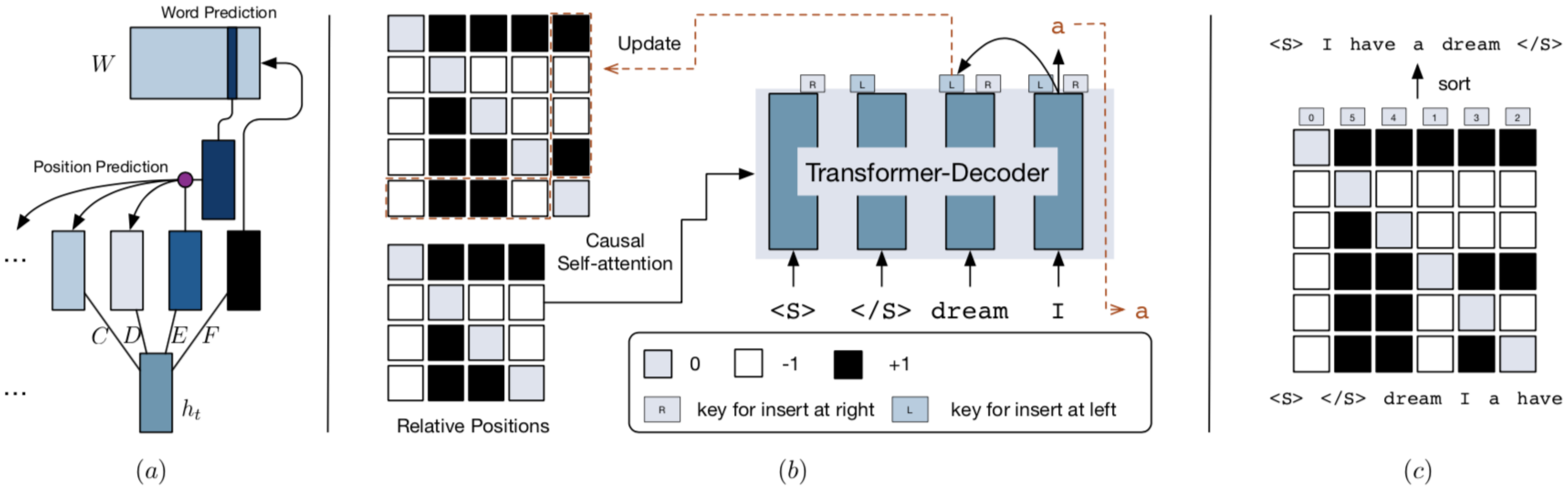



Figure 1: An example of InDIGO. At each step, we simultaneously predict the next token and its (relative) position to be inserted. The final output sequence is obtained by mapping the words based on their positions.

# NLG



Word Prediction
W
Position Prediction
...
C D E F
$h_t$
(a)
Update
a
Transformer-Decoder
Causal
Self-attention
Relative Positions
<S> </S> dream I
0 -1 +1
key for insert at right
key for insert at left
(b)
<S> I have a dream </S>
sort
0 5 4 1 3 2
<S> </S> dream I a have
(c)

# 从算法到工具

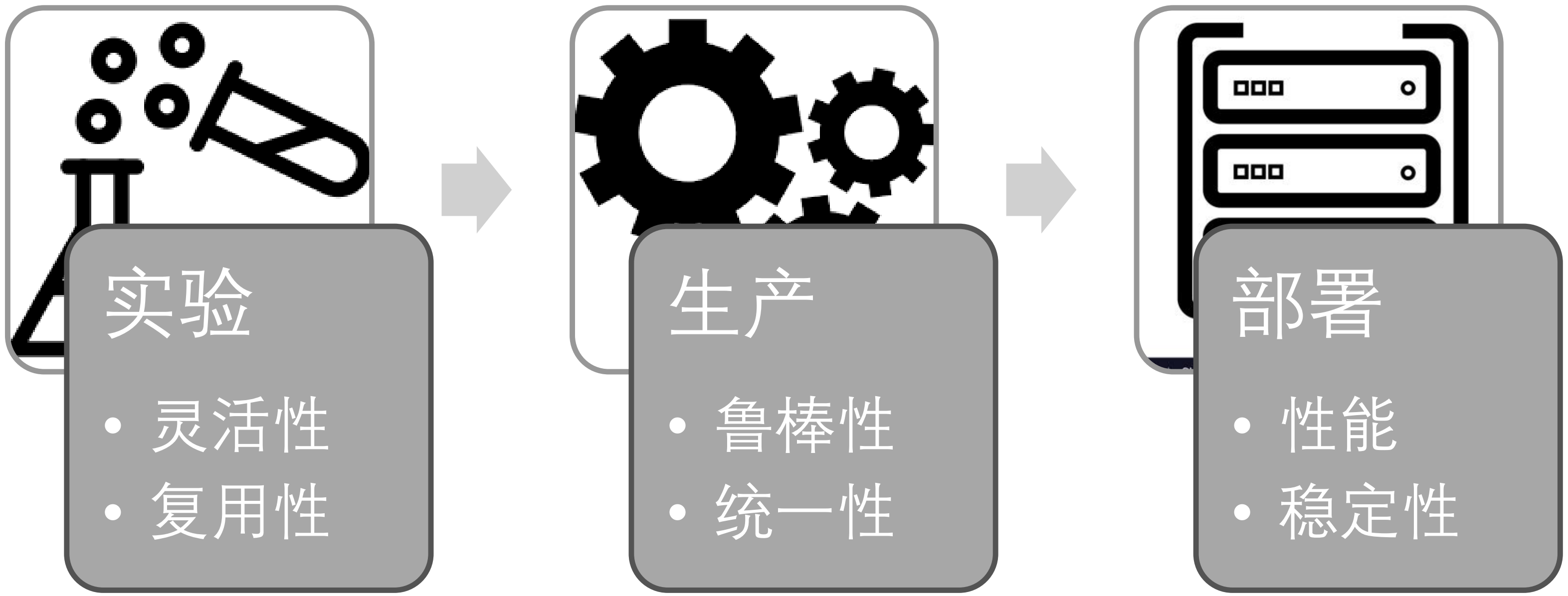
我们来谈谈TensorFlow

# 从算法到工具

现在我们来谈谈TensorFlow

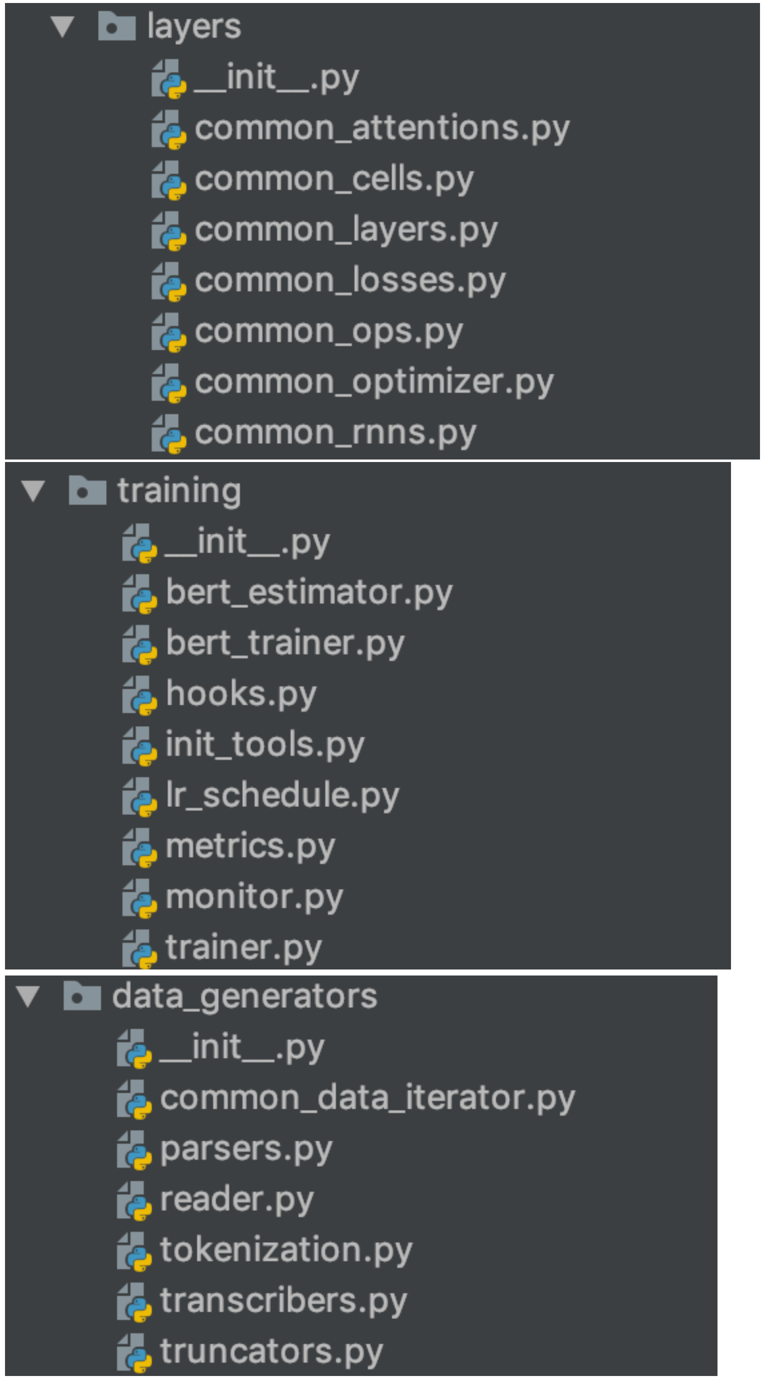
# 为什么用TensorFlow

- 历史因素
  - 当年……
  - 工业部署的支持
  - 相对齐全的文档和相对活跃的社区
  - 对分布式计算的支持
- 不友善的地方
  - 冗余的API
  - 静态图，Eager真香？


实验
灵活性
复用性
生产
鲁棒性
统一性
部署
性能
稳定性

# 设计实验代码

- 灵活性与复用性
  - 模块化成熟的部分
  - 不成熟的部分使用函数（T2T）
  - 构建自己的训练流程框架 (Estimator is good, but…)
  - 构建自己的数据流（Dataset is good, but…)
  - 单元测试



# 实验转为生产

- 鲁棒性
  - 生产环境的数据多变
  - 小心你的预处理，请考虑尽可能多的Corner case
- 统一性
  - 为与其他模块的交互设计统一的接口
  - 记得预留好额外的接口

# 服务

- 服务中的训练与推断
  - 设计合理的网络结构;
  - 合理设计训练日程，active learning;
  - 多卡方案，all reduce is good;
  - 使用cuda_rnn;

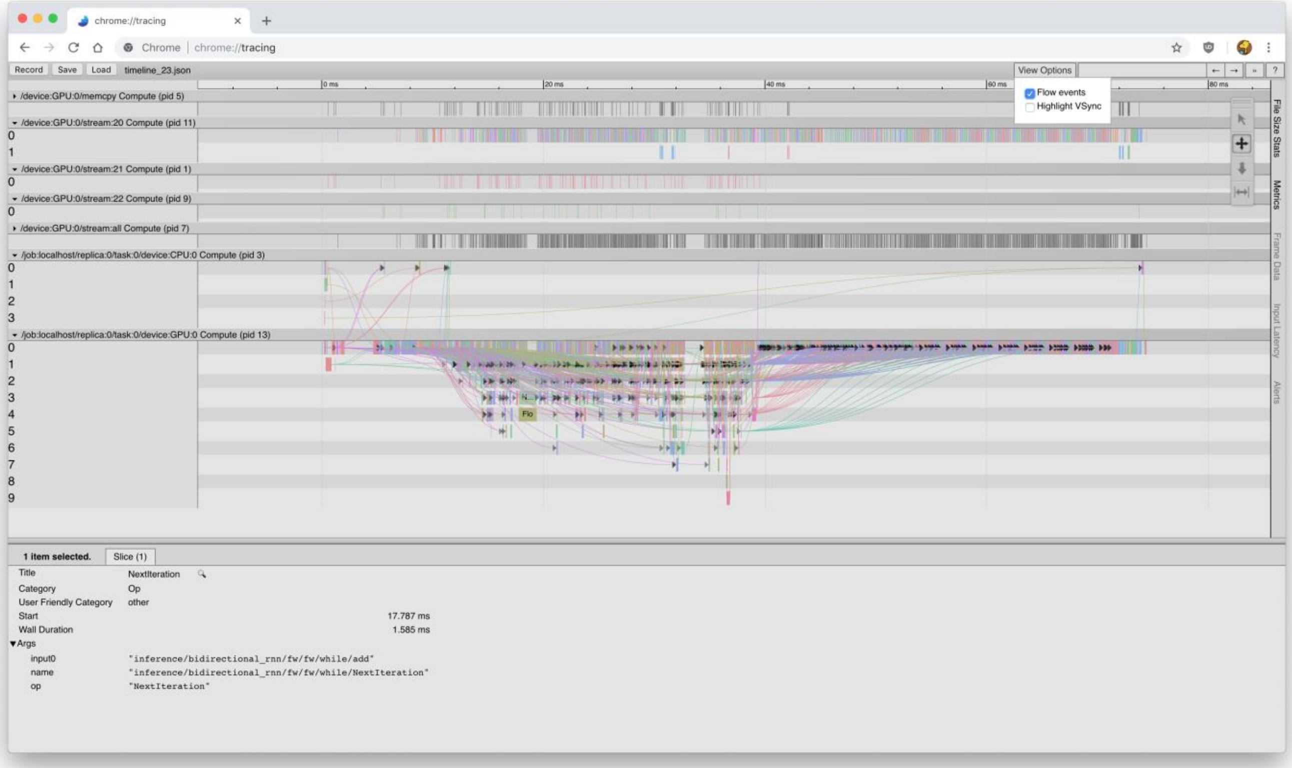
# 我们的迭代

- 从placeholder到dataset
  - 使用tf-record
  - 简单的前处理，如词转id，bucketing，padding；
  - 复杂的前处理可以用from_generator
  - pre-fetch to device

# 我们的迭代

- 从单卡到多卡；
- 卡的使用率瓶颈；
- 使用profiling来找到耗时长的节点；

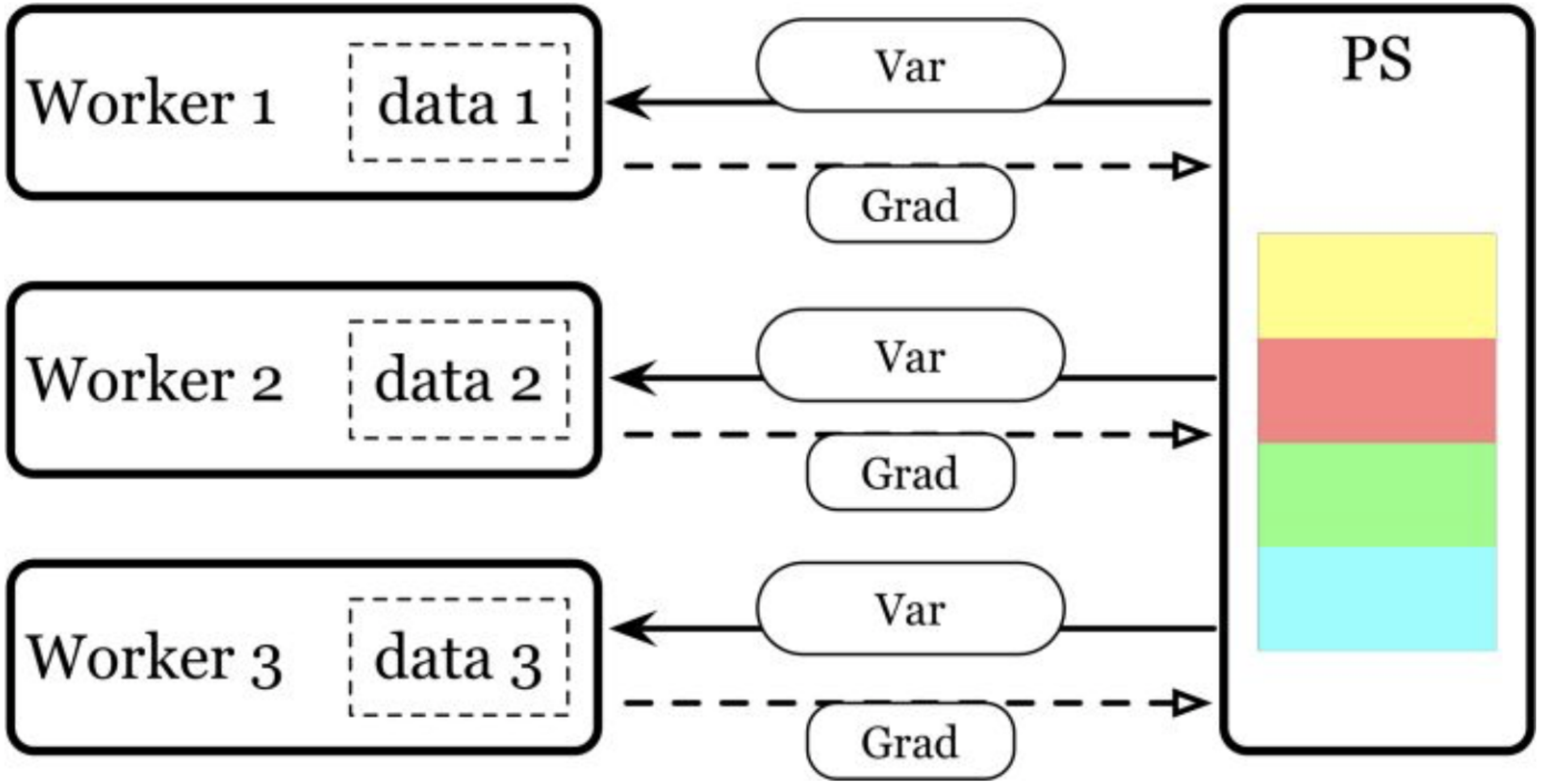
```python
import tensorflow as tf
from tensorflow.python.client import timeline
def profile(fetch_keys, sess, step, output_dir):
    run_metadata = tf.RunMetadata()
    options = tf.RunOptions(trace_level=tf.RunOptions.FULL_TRACE)
    sess.run(fetch_keys,
             options=options,
             run_metadata=run_metadata)
    trace = timeline.Timeline(step_stats=run_metadata.step_stats)
    trace_file = open(os.path.join(output_dir, 'timeline_%d.json' % step), 'w')
    trace_file.write(trace.generate_chrome_trace_format())
```



# 我们的迭代

- Parameter Server

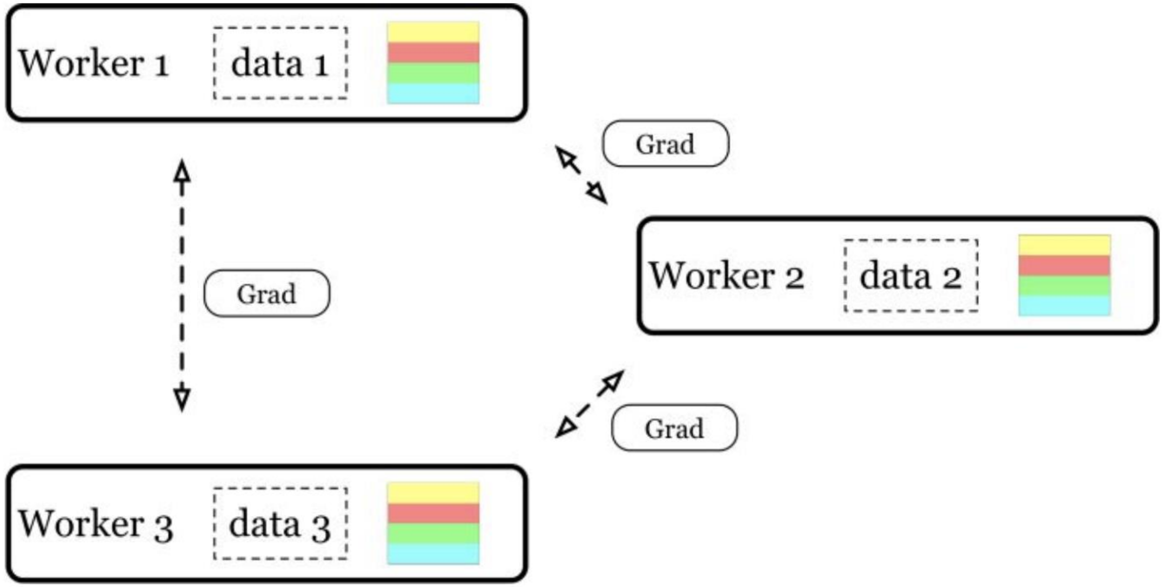
1. 从worker上收集梯度
2. 在Host上计算平均
3. 将平均梯度回传给worker



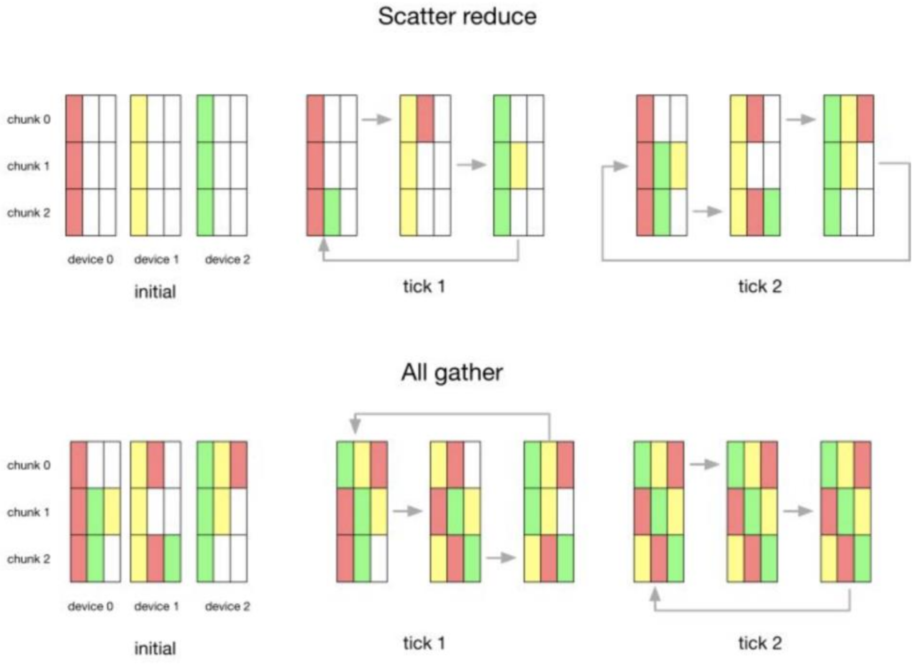

parameter server 的图示

# 我们的迭代

- All Reduce方式
- 平均梯度在worker上计算并更新;
- 核心思想：切成小块相互传




allreduce 模式的图示




ring-allreduce 的简易图示


更 美 好 AI 世 界
谢谢
ZHUIYI