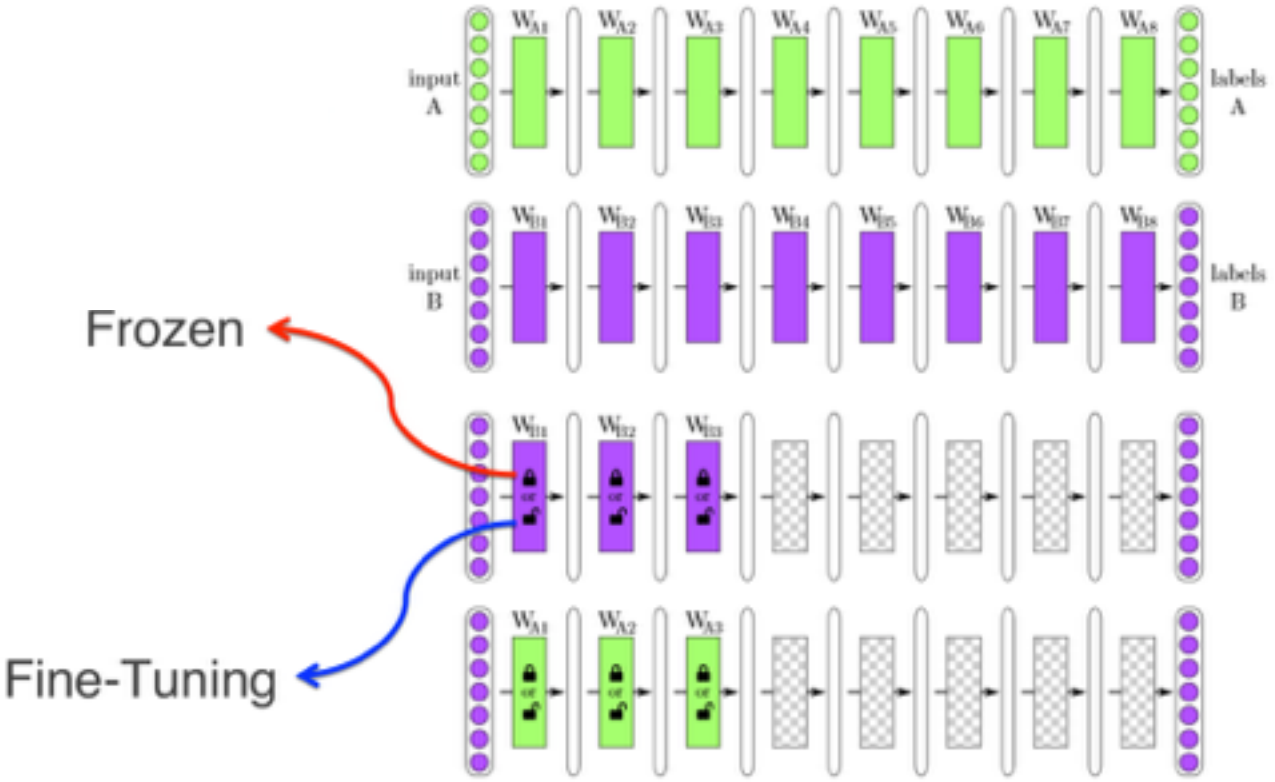
预训练 & Elmo

Syllabus

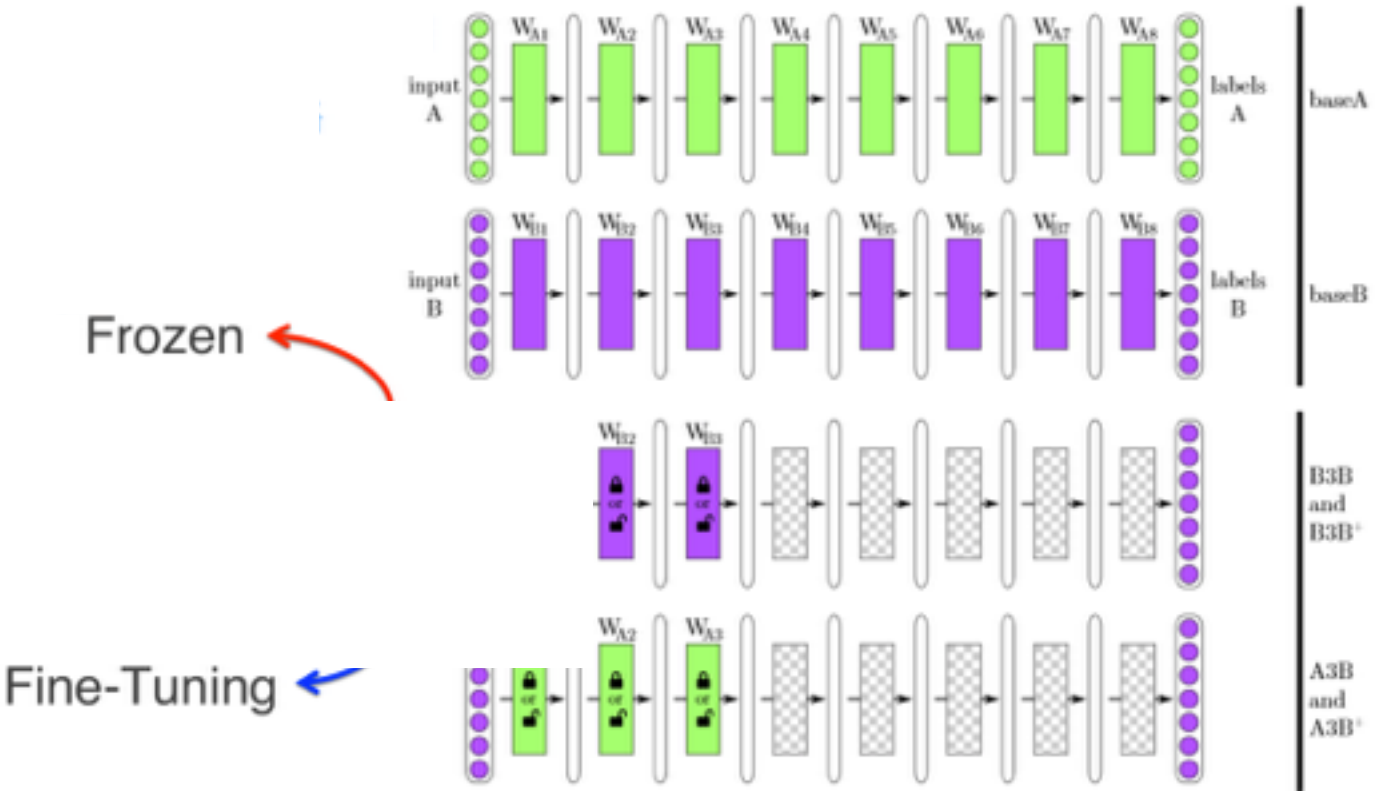
- 预训练在图像领域的应用
- ELMO：基于上下文的word-embedding

预训练在图像领域的应用



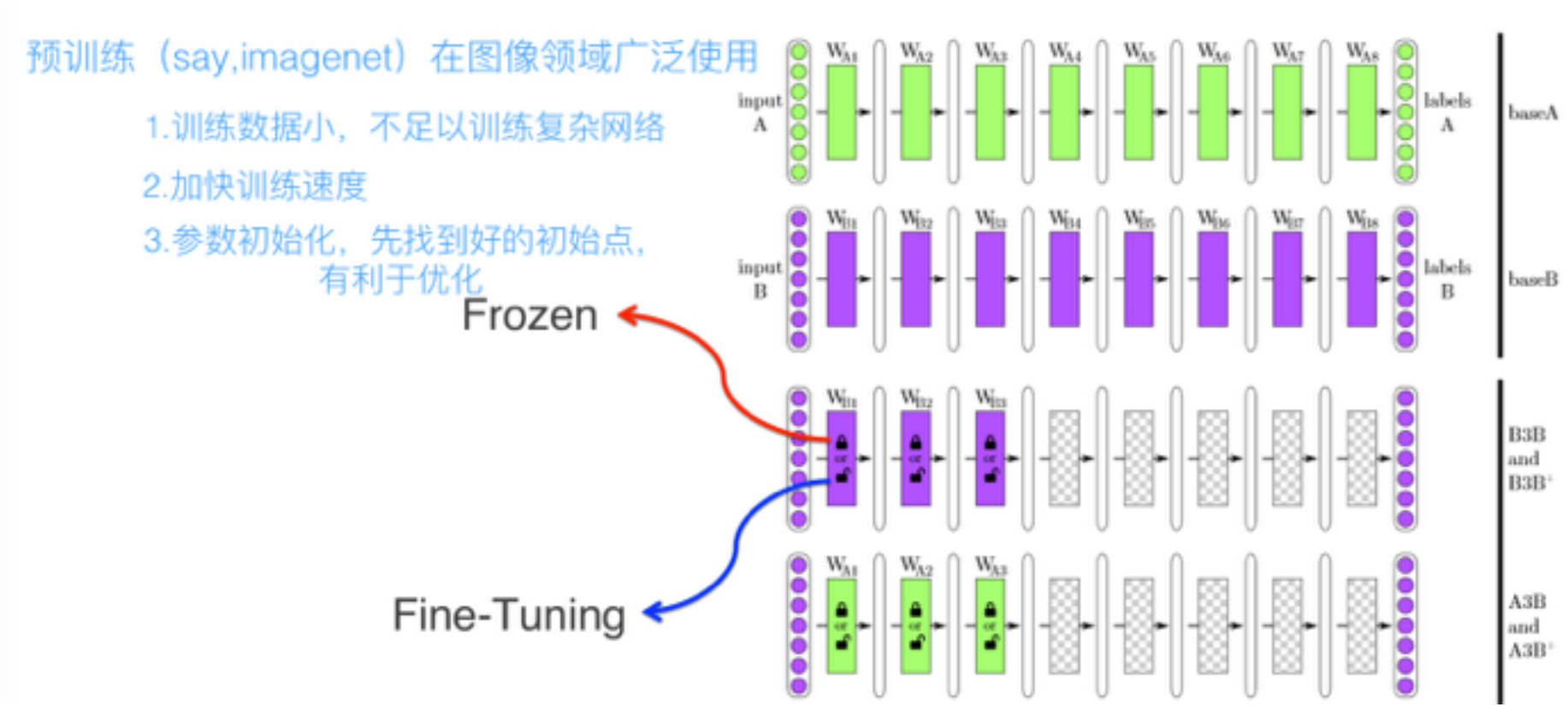
- 对于图像来说一般是CNN的多层叠加网络结构，可以先用某个训练集合比如训练集合A或者训练集合B对这个网络进行预先训练，在A任务上或者B任务上学会网络参数，然后存起来以备后用。

预训练在图像领域的应用



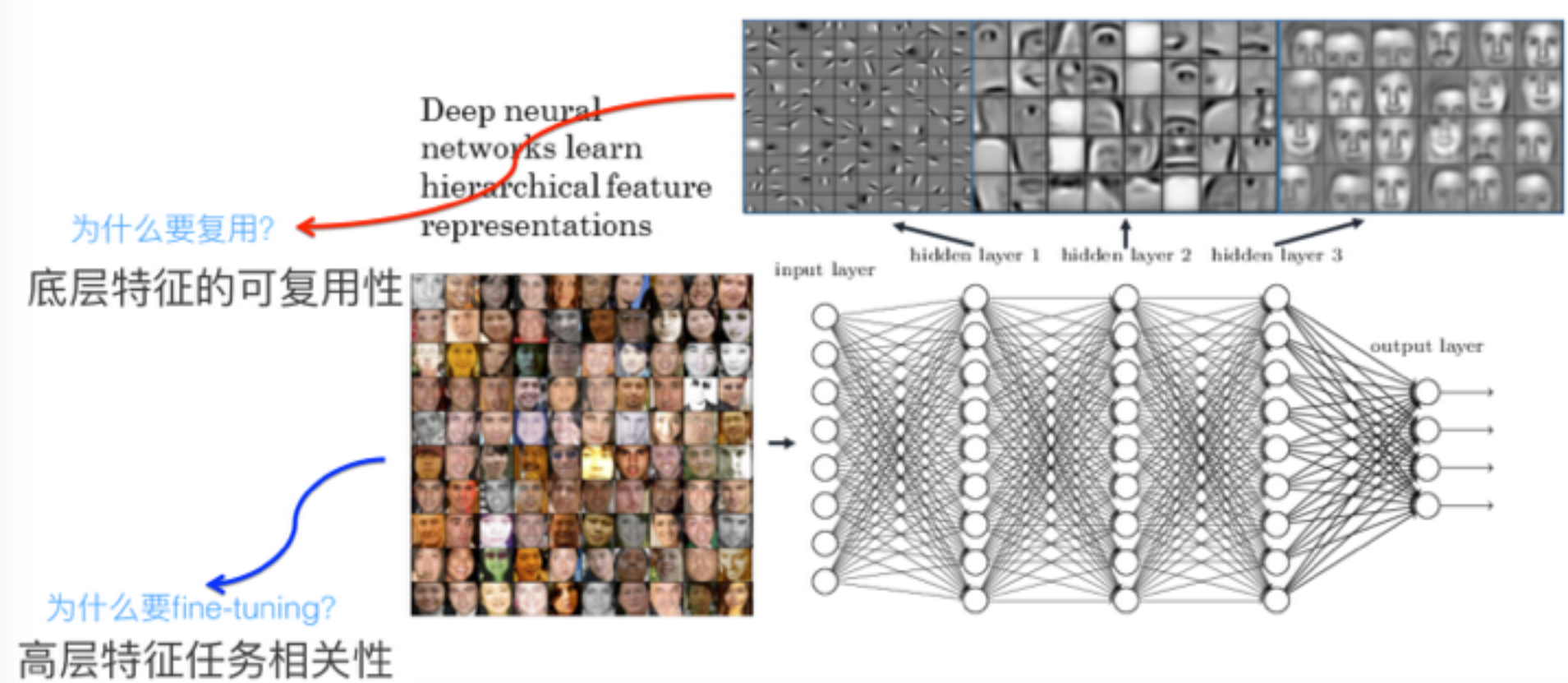
- 假设我们面临第三个任务C，网络结构采取相同的网络结构，在比较浅的几层CNN结构，网络参数初始化的时候可以加载A任务或者B任务学习好的参数，其它CNN高层参数仍然随机初始化。之后我们用C任务的训练数据来训练网络.

预训练在图像领域的应用

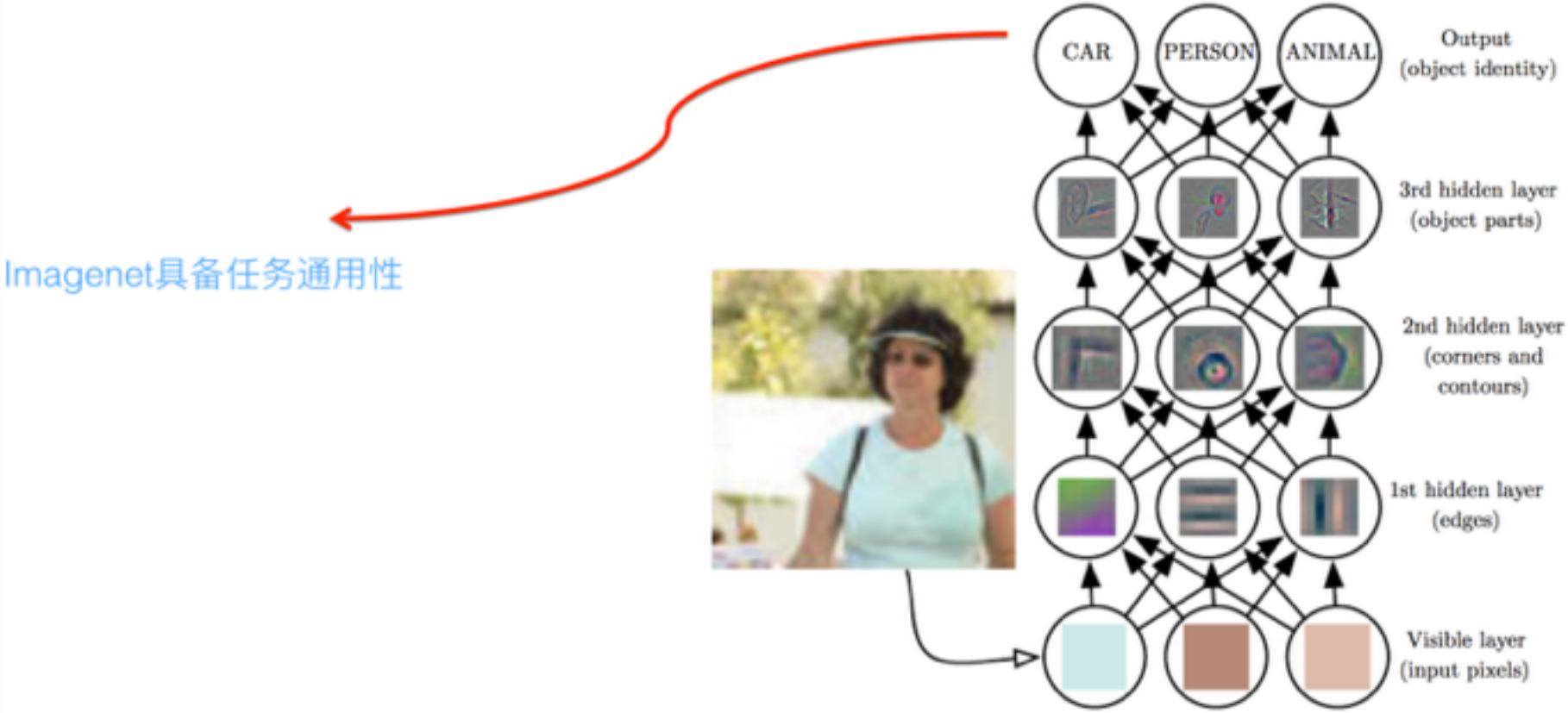


- **Frozen**：浅层加载的参数在训练C任务过程中不动
- **Fine-Tuning:** 底层网络参数尽管被初始化了，在C任务训练过程中仍然随着训练的进程不断改变，更好地把参数进行调整使得更适应当前的C任务

为什么预训练可行

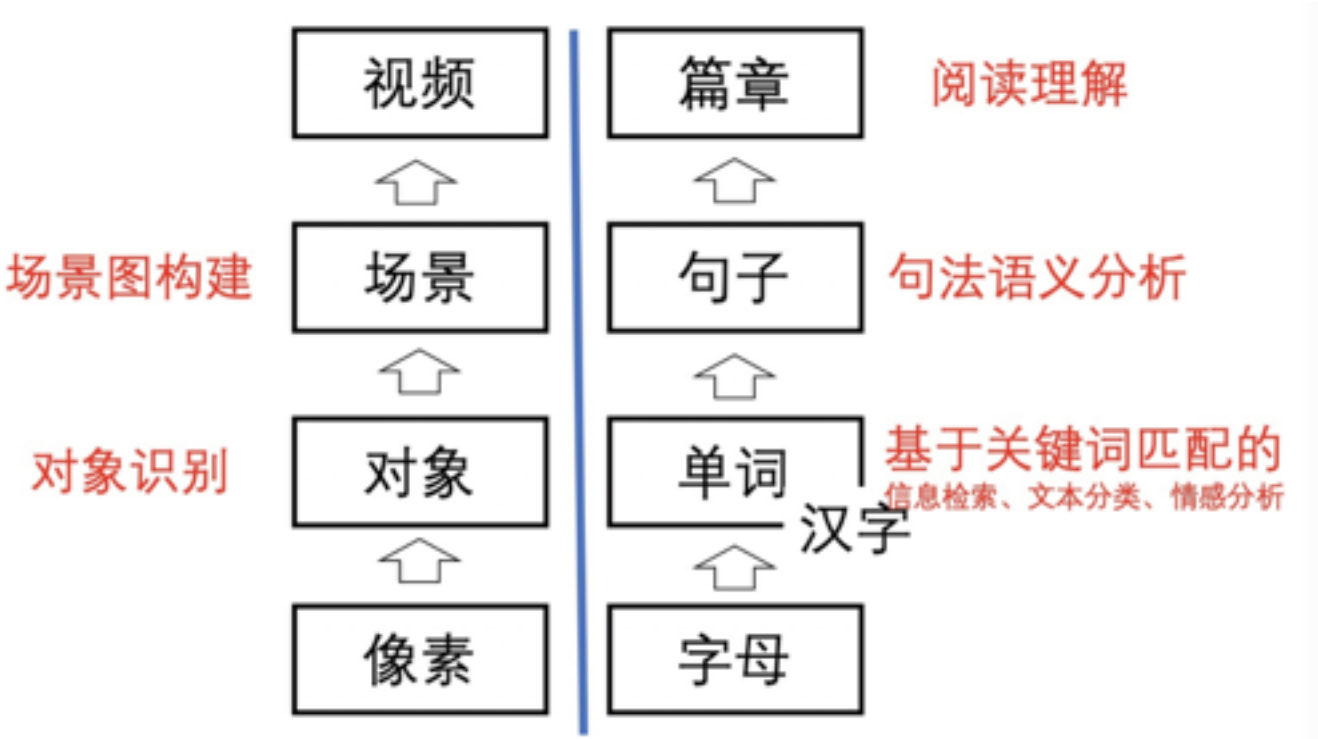

Deep neural networks learn hierarchical feature representations
为什么要复用?
底层特征的可复用性
为什么要fine-tuning?
高层特征任务相关性
input layer
hidden layer 1
hidden layer 2
hidden layer 3
output layer

ImageNet



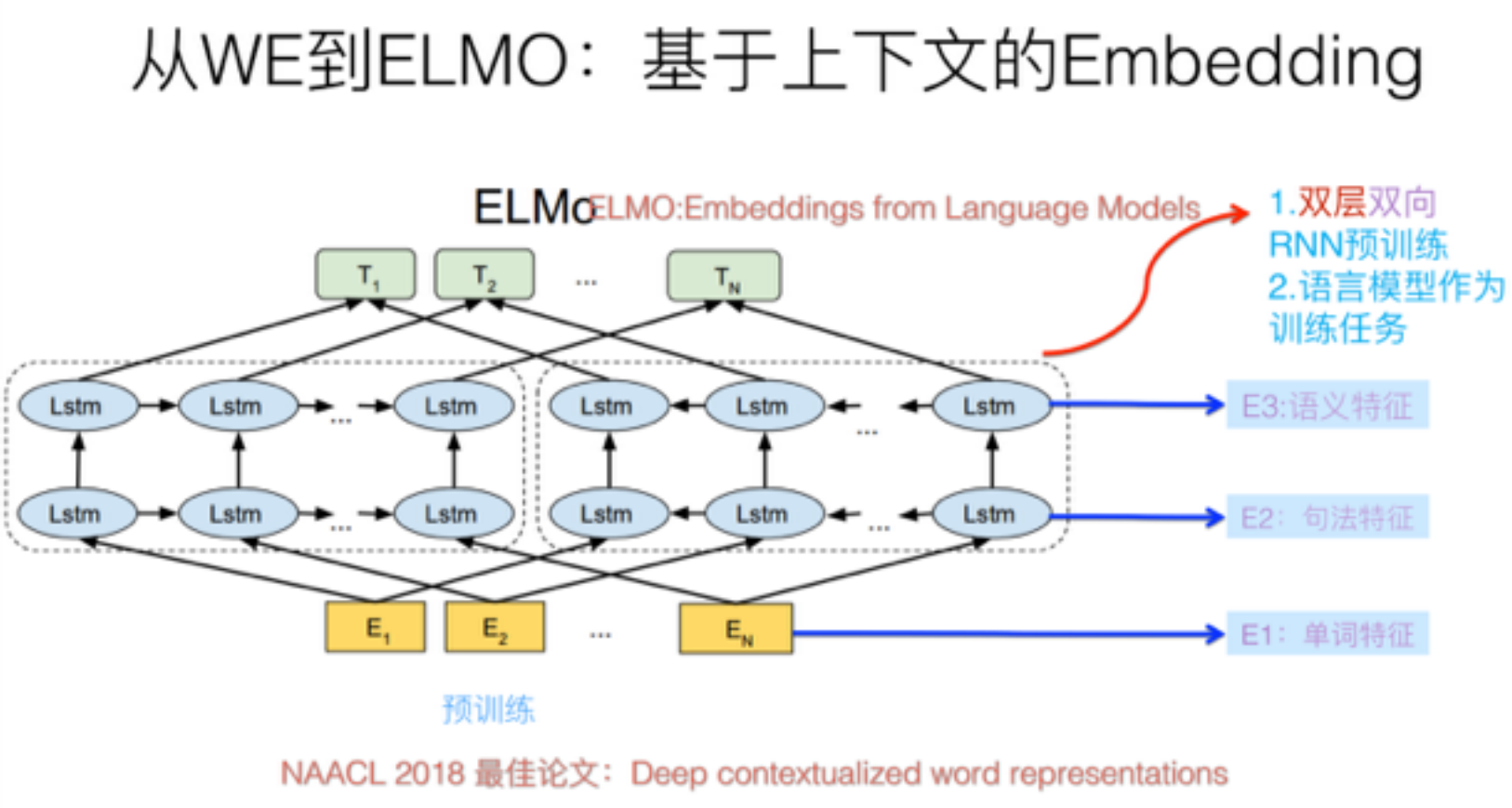
- ImageNet http://www.image-net.org/ 是图像领域里有超多事先标注好训练数据的数据集合
- ImageNet有1000类图像数据始于2009年，当时李飞飞教授等在CVPR2009上发表了一篇名为《ImageNet: A Large-Scale Hierarchica Image Database》的论文，之后就是基于ImageNet数据集的7届ImageNet挑战赛. ImageNet是根据WordNet层次结构组织的图像数据集。

预训练对应NLP



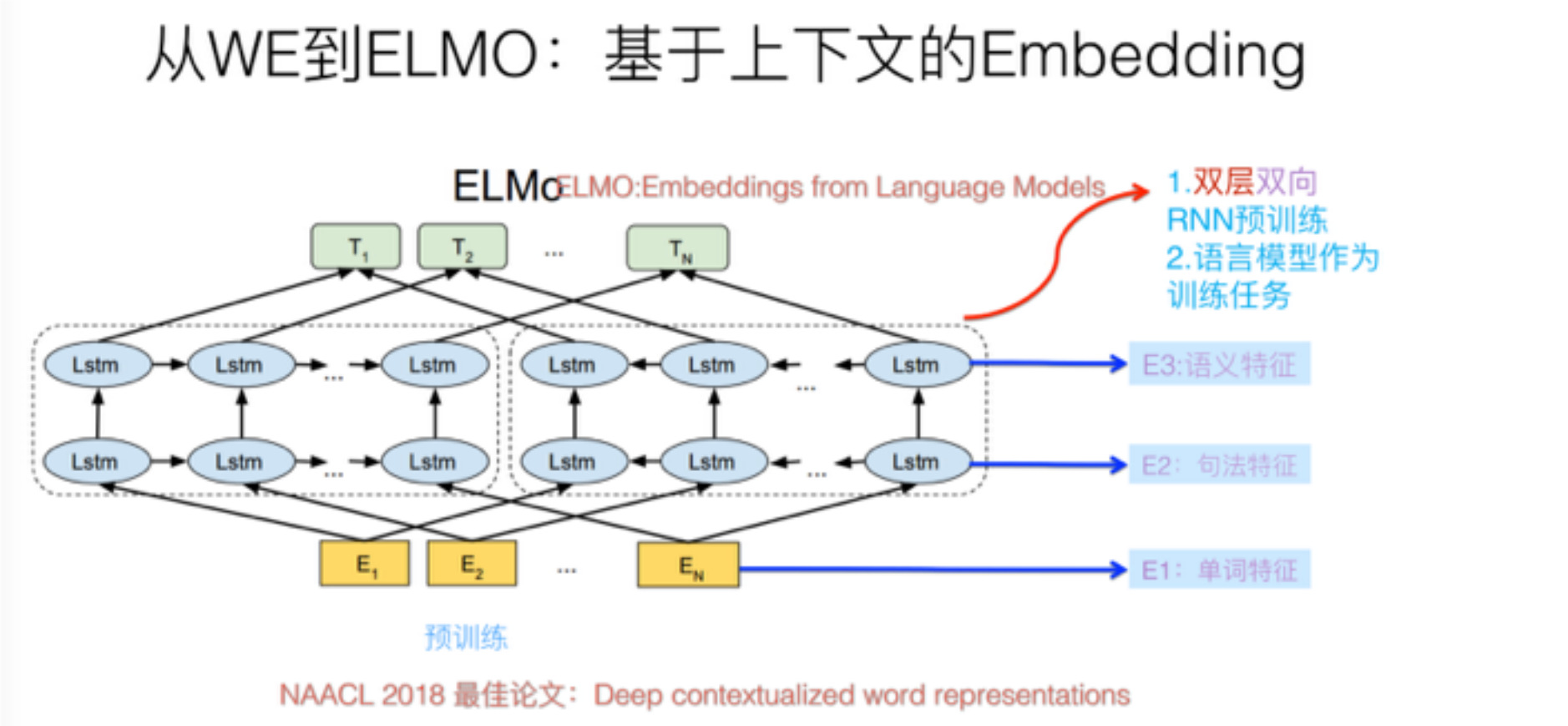
词作为NLP的基本要素，比像素的抽象程度更高，已经加入了人类数万年进化而来的抽象经验。所以之前的NLP预训练工作主要集中于对词的表示。

ELMO：Pre-train



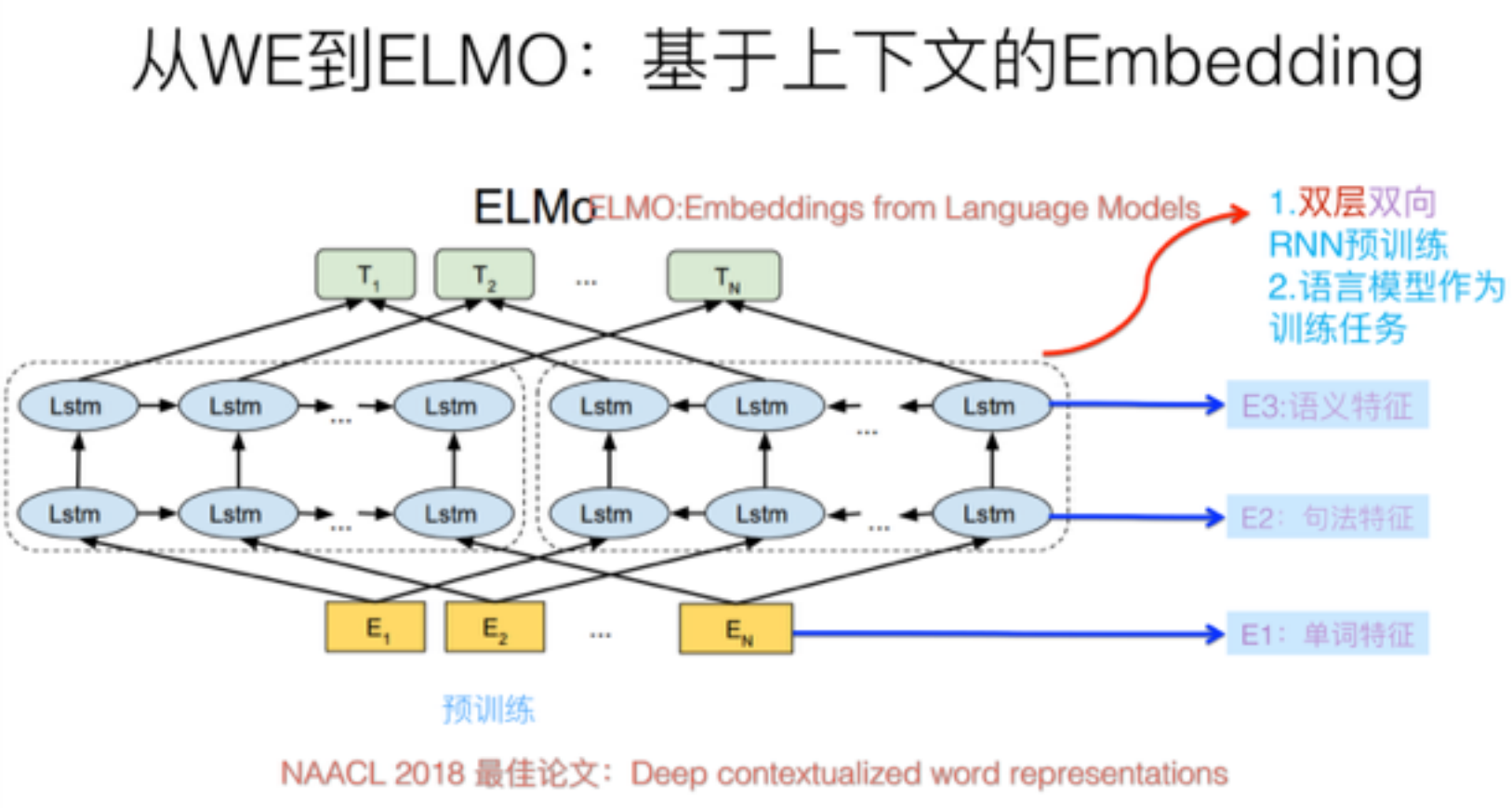
- Deep and Contextualised: 网络结构采用了双层双向LSTM，目前**语言模型**训练的任务目标是**根据单词 W_i 的上下文去正确预测单词 W_i**， W_i 之前的单词序列Context-before称为上文，之后的单词序列Context-after称为下文

ELMO：Pre-train



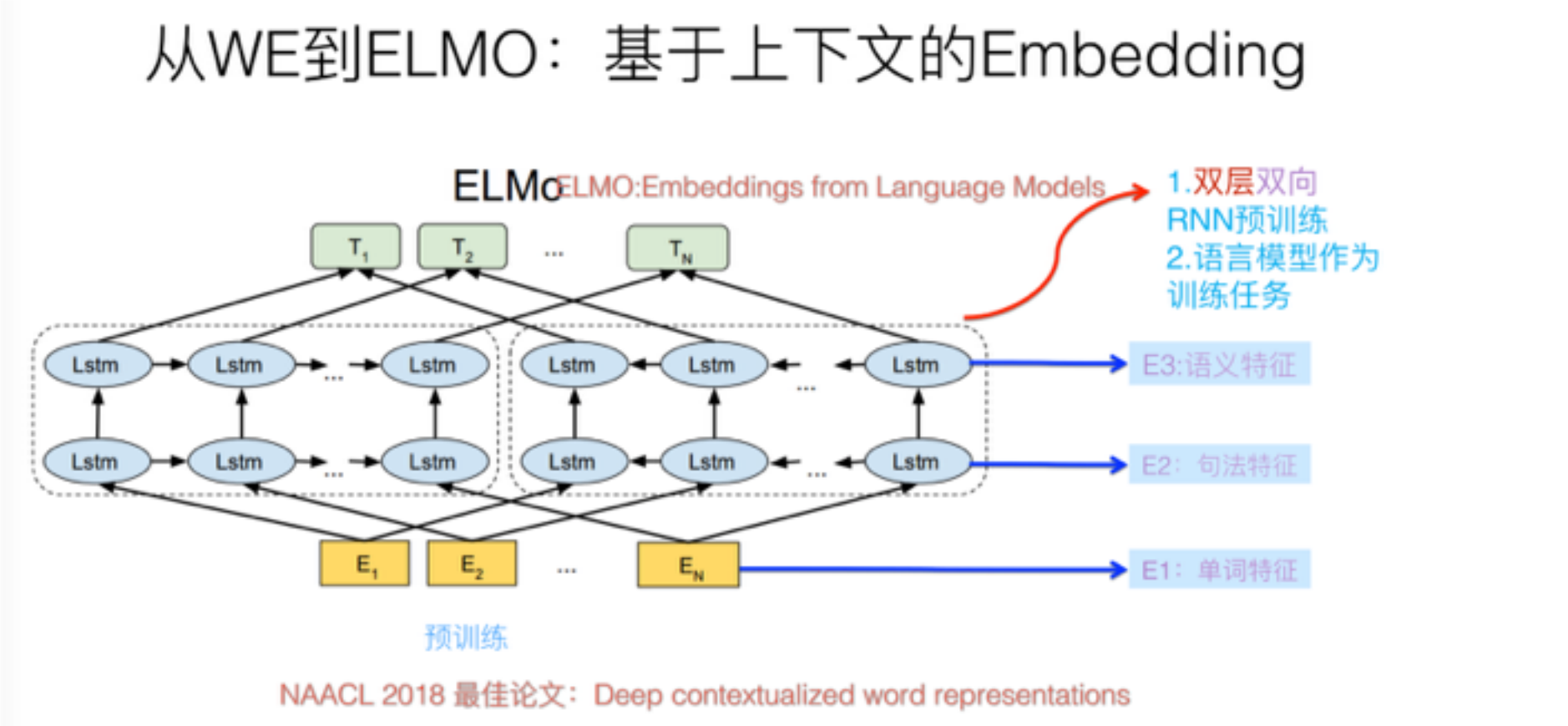
- 左端的前向双层LSTM代表正方向编码器，输入的是从左到右顺序的除了预测单词外 W_i 的上文Context-before；右端的逆向双层LSTM代表反方向编码器，输入的是从右到左的逆序的句子下文Context-after；每个编码器的深度都是两层LSTM叠加。

ELMO：Pre-train



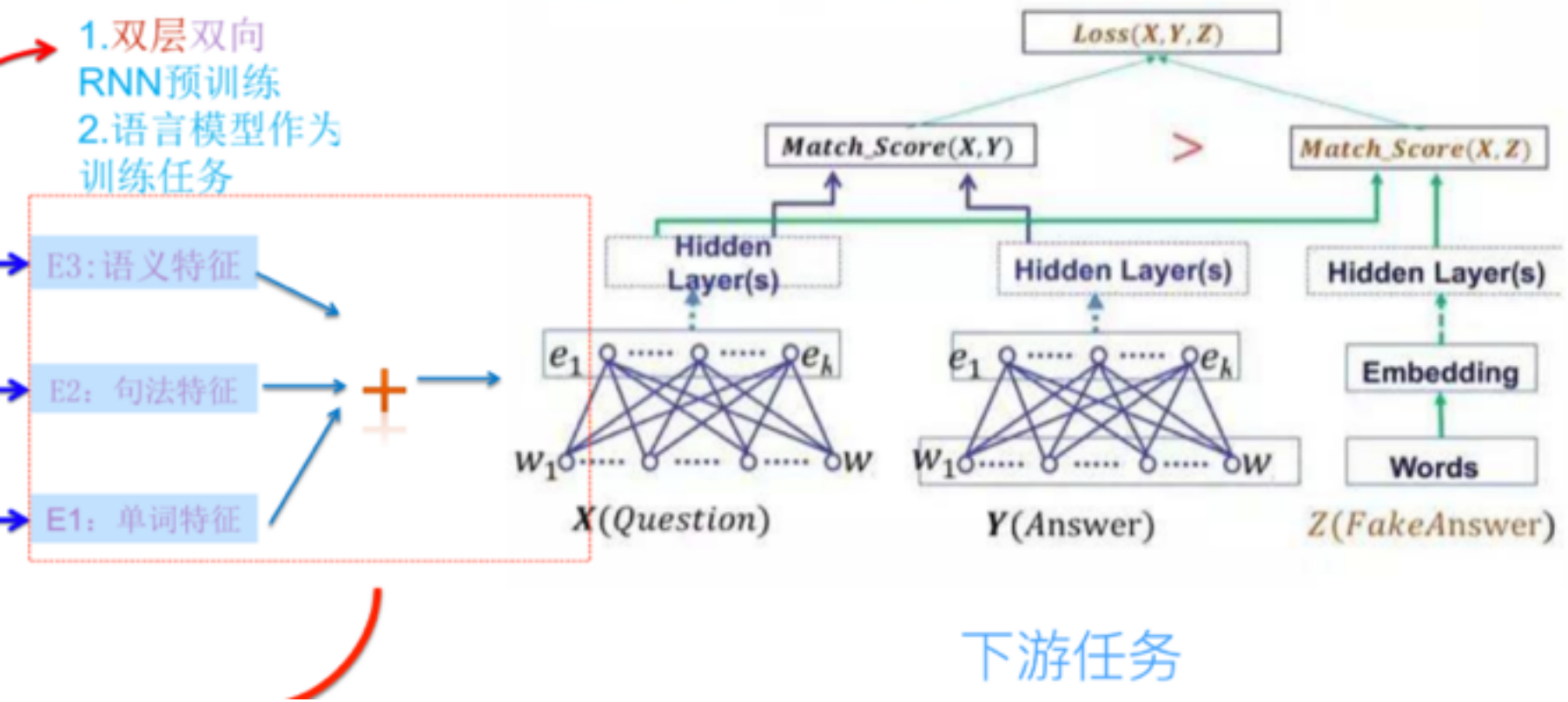
- 如果训练好这个网络后，输入一个新句子 Snew ， 句子中每个单词都能得到对应的**三个Embedding,** ELMO的预训练过程不仅仅学会单词的Word Embedding，还学会了一个双层双向的LSTM网络结构，而这两者后面都有用。

ELMO: Downstream task



比如我们的下游任务仍然是QA问题，此时对于问句X可以先**将句子X作为预训练好的ELMO网络的输入**，这样句子X中每个单词在**ELMO网络中都能获得对应的三个Embedding**，之后给予这三个Embedding中的每一个Embedding一个权重a，这个权重可以学习得来，根据各自权重累加求和，**将三个Embedding整合成一个**。然后将整合后的这个Embedding作为X句在自己任务的那个网络结构中对应单词的输入，**以此作为补充的新特征给下游任务使用**。对于上图所示下游任务QA中的回答句子Y来说也是如此处理。因为ELMO给下游提供的是每个单词的特征形式，所以这一类预训练的方法被称为“Feature-based Pre-Training”。

ELMO: Downstream task



ELMO：多义词问题解决了吗?

(多义词)Play:
1.运动
2.音乐

	Source	Nearest Neighbors
GloVe	play	playing, game, games, played, players, plays, player, Play, football, multiplayer
biLM	Chico Ruiz made a spectacular play on Alusik 's grounder {...}	Kieffer , the only junior in the group , was commended for his ability to hit in the clutch , as well as his all-round excellent play .
	Olivia De Havilland signed to do a Broadway play for Garson {...}	{...} they were actors who had been handed fat roles in a successful play , and had talent enough to fill the roles competently , with nice understatement .

Table 4: Nearest neighbors to "play" using GloVe and the context embeddings from a biLM.

不仅解决了，词性也能对应起来!

使用ELMO，根据上下文动态调整后的embedding不仅能够找出对应的“演出”的相同语义的句子，而且还可以保证找出的句子中的play对应的词性也是相同的，第一层LSTM编码了很多句法信息，这在这里起到了重要作用。

ELMO：效果如何?

Task	Previous SOTA		Our baseline	ELMo + baseline	Increase (absolute/ relative)
SQuAD	Liu et al. (2017)	84.4	81.1	85.8	4.7 / 24.9%
SNLI	Chen et al. (2017)	88.6	88.0	88.7 ± 0.17	0.7 / 5.8%
SRL	He et al. (2017)	81.7	81.4	84.6	3.2 / 17.2%
Coref	Lee et al. (2017)	67.2	67.2	70.4	3.2 / 9.8%
NER	Peters et al. (2017)	91.93 ± 0.19	90.15	92.22 ± 0.10	2.06 / 21%
SST-5	McCann et al. (2017)	53.7	51.4	54.7 ± 0.5	3.3 / 6.8%

6个NLP任务：5%到25%的提高!