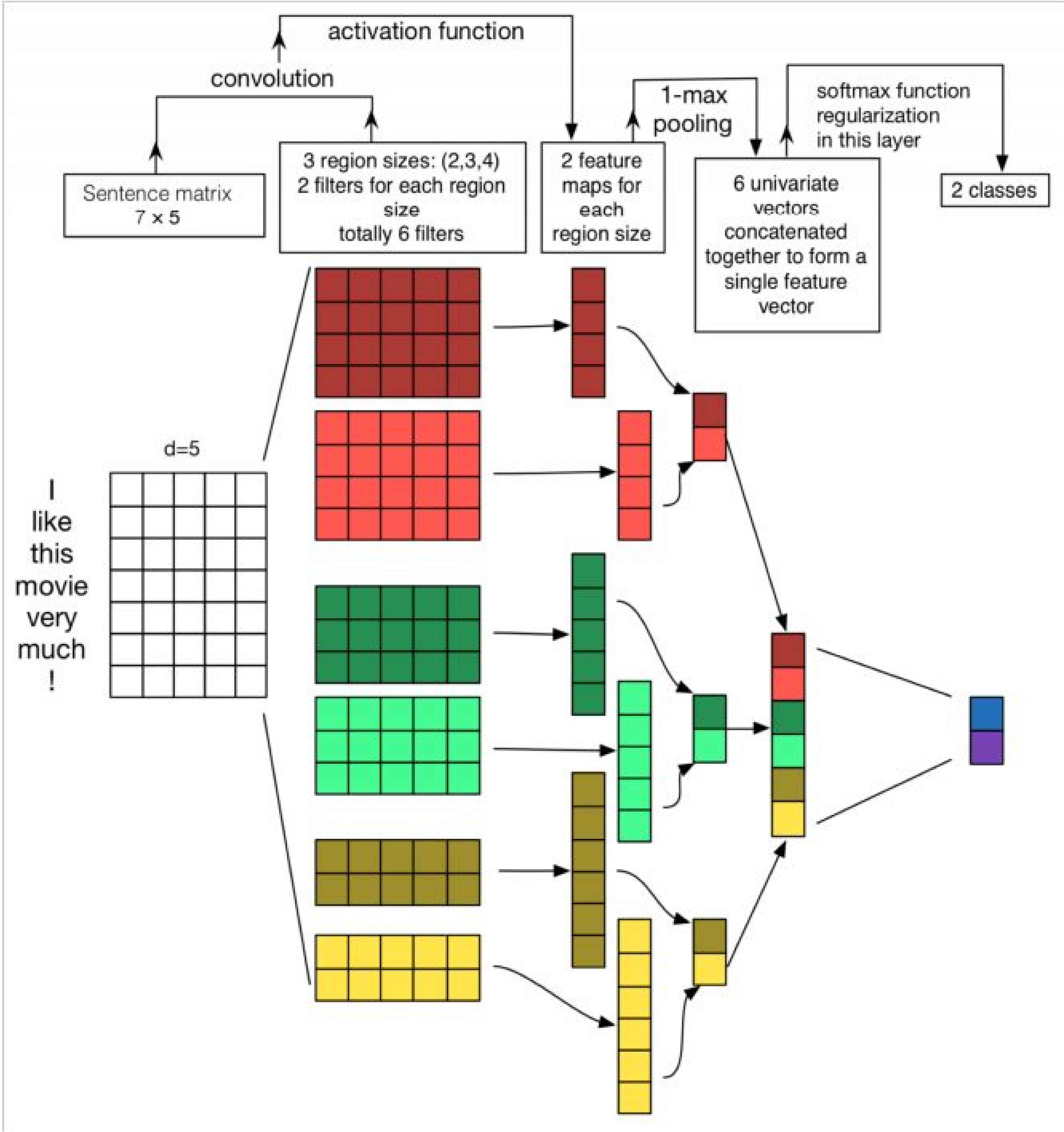
- 第11课：NLP中的卷积神经网络
- 第12课：子词（Subword）模型

第11课：NLP中的卷积神经网络

NLP卷积神经网络尝试

- [Yoon Kim (2014): Convolutional Neural Networks for Sentence Classification. EMNLP 2014.]
- [Zhang and Wallace (2015) A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification]
- 基于句子里的每个窗口词汇进行卷积
- Multi-channel input idea: two copies of pre-trained word vectors, backprop into only one set.
- 采用多个范围的filter，接上max-pooling，再拼接到一起做softmax，得到二分类。
- 可以增加Dropout
- All hyperparameters in Kim (2014)：ReLU、window h=3,4,5，Dropout p=0.5.....

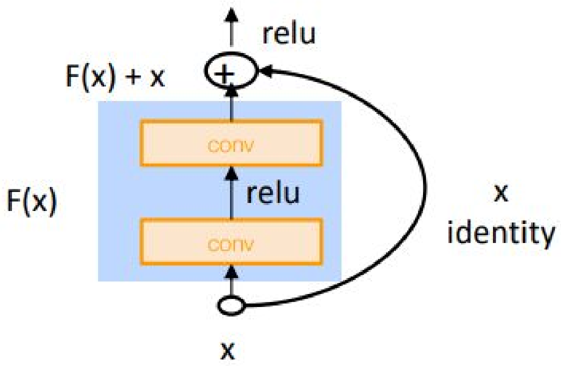


前面学过的各类模型对比：

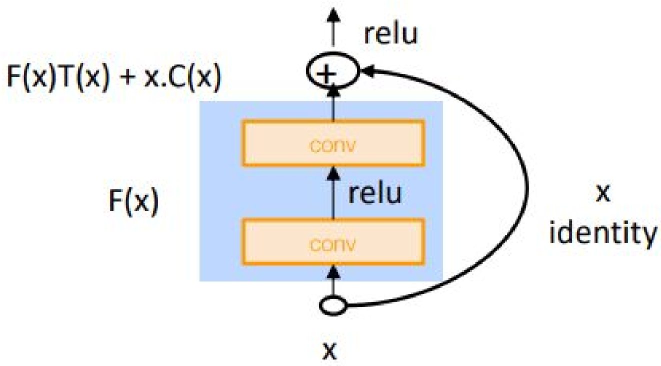
- Bag of Vectors：对简单分类任务效果不错，特别是后续带上ReLU。
- Window Model：单个词的分类不错，但长距离上下文无法处理。
- CNNs：分类问题不错，容易用GPU并行计算。
- RNNs：符合认知，仅采用最终state则对分类并不够好，计算慢，适合序列标签和分类、语言模型，加上Attention非常好。

CNN方法优化：

- **门控制单元**竖向使用。



Residual block
(He et al. ECCV 2016)

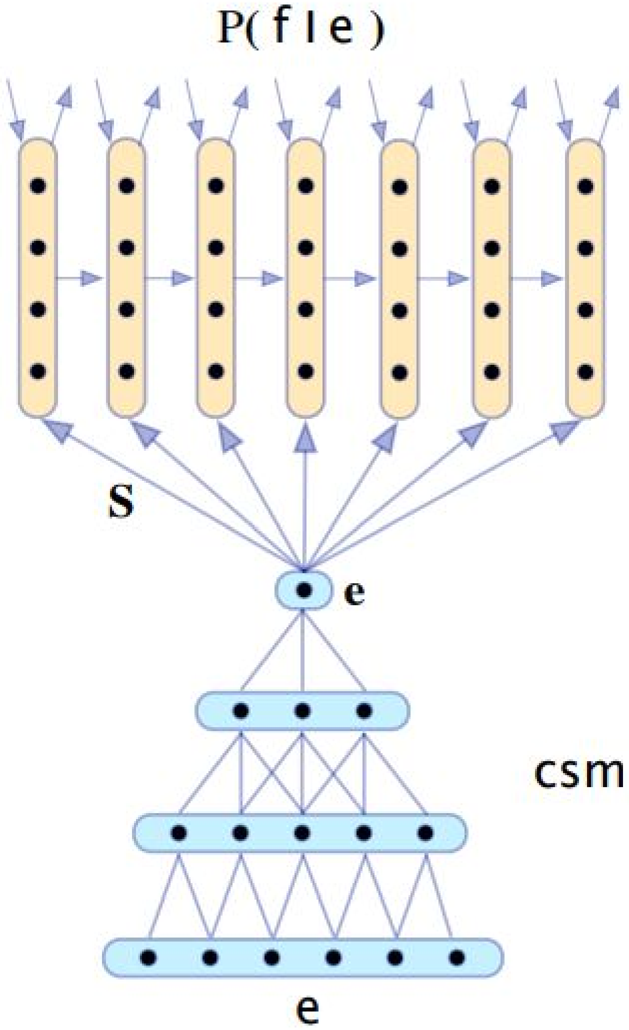


Highway block
(Srivistava et al. NeurIPS 2015)

- **Batch Normalization (BatchNorm)**：对分布进行规整，并携带当前偏移量和缩放量，便于后续训练更容易学习
  [Ioffe and Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift]
- **1 x 1 Convolutions**：仅对channel进行转换，相当于channel的fc
  [Lin, Chen, and Yan. 2013. Network in network. arXiv:1312.4400.]

CNN应用：机器翻译

- 采用CNN作为encoder，采用RNN作为decoder
- [“Recurrent Continuous Translation Models” , Kalchbrenner and Blunsom (2013)]



Learning Character-level Representations for Part-of-Speech Tagging [Dos Santos and Zadrozny (2014)]

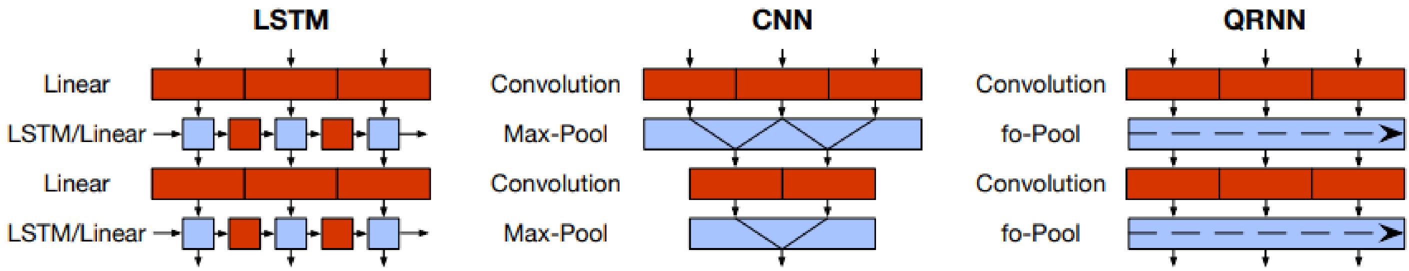
- 字符级别的CNN，做词性标注
- 字符上卷积形成word embedding

Character-Aware Neural Language Models (Kim, Jernite, Sontag, and Rush 2015) ——下一节课详细讲

Very Deep Convolutional Networks for Text Classification (**VD-CNN**) [Conneau, Schwenk, Lecun, Barrault. EACL 2017]

Quasi-Recurrent Neural Networks (**QRNN**) [James Bradbury, Stephen Merity, Caiming Xiong & Richard Socher. ICLR 2017]

- RNN训练无法并行化而且非常慢，于是试图组合RNN和CNN的一部分
- 把CNN的最大池化部分，更改为带有Forget、Output门的fo-Pool
- 在Language Modeling和Sentiment Analysis方面效果效果更好、速度更快、具备更好解释性
- 缺点：无法像LSTM一样处理基于字符的Language Modeling、同等效果通常比LSTM需要更深

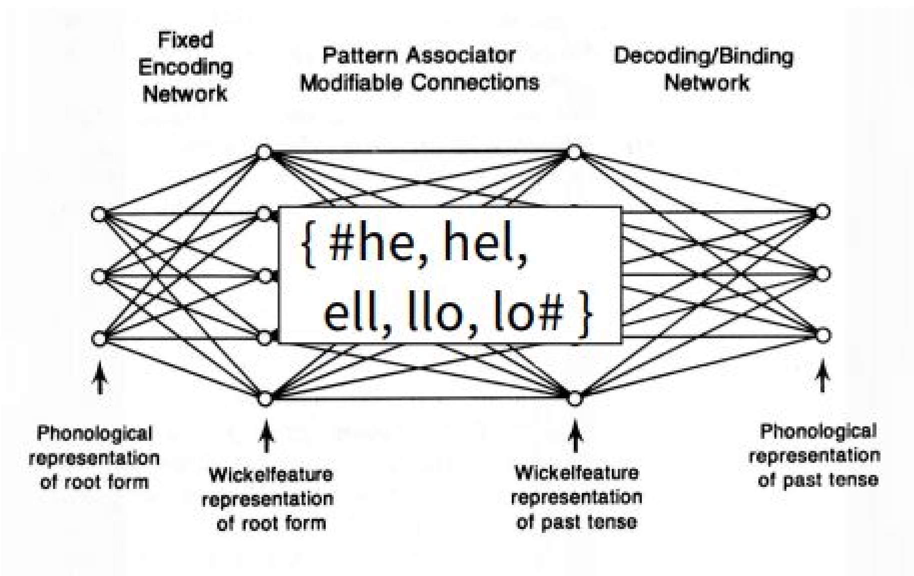


第12课：子词（Subword）模型

人类语言是有发音（phonetics）和音韵（phonology）的。

词的语义组成部分：Deep learning: Morphology little studied; one attempt with recursive neural networks is (Luong, Socher, & Manning 2013)

采用基于字符的n-grams：Wickelphones， 微软DSSM



wickelphones

基于字符的embedding可以用于组成word embedding

- 可以解决未登录词问题
- 读音类似的词具有相似的embedding

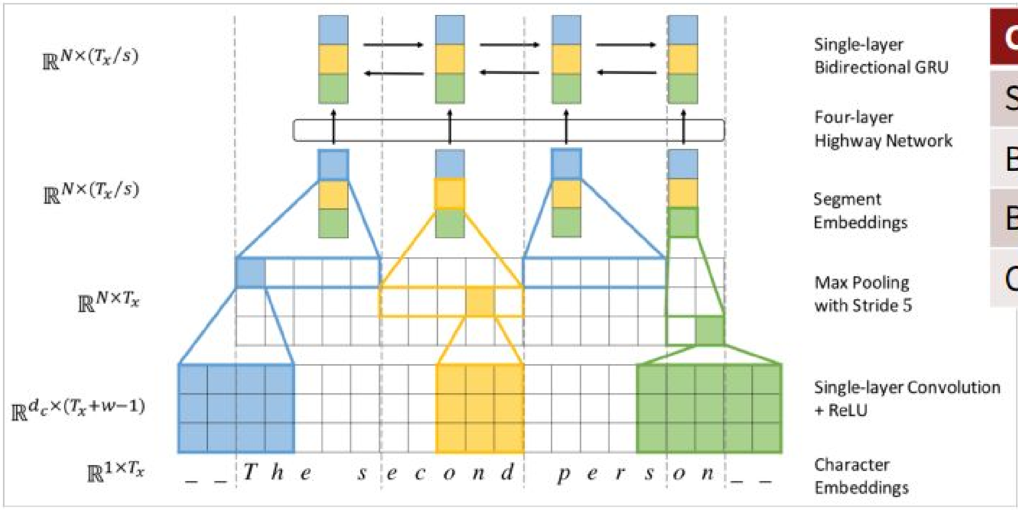
Purely character-level NMT models

- (Wang Ling, Isabel Trancoso, Chris Dyer, Alan Black, arXiv 2015)

- (Thang Luong, Christopher Manning, ACL 2016)
- (Marta R. Costa-Jussà, José A. R. Fonollosa, ACL 2016)

Fully Character-Level Neural Machine Translation without Explicit Segmentation [Jason Lee, Kyunghyun Cho, Thomas Hoffmann. 2017]

- Encoder为character embedding=>卷积+ReLU=>最大池化=>四层Highway=>单层双向GRU
- Decoder为char-level GRU



CS-En	WMT 15	Test
Source	Target	BLEU
Bpe	Bpe	20.3
Bpe	Char	22.4
Char	Char	22.5

Revisiting Character-Based Neural Machine Translation with Capacity and Compression. [2018. Cherry, Foster, Bapna, Firat, Macherey, Google AI]

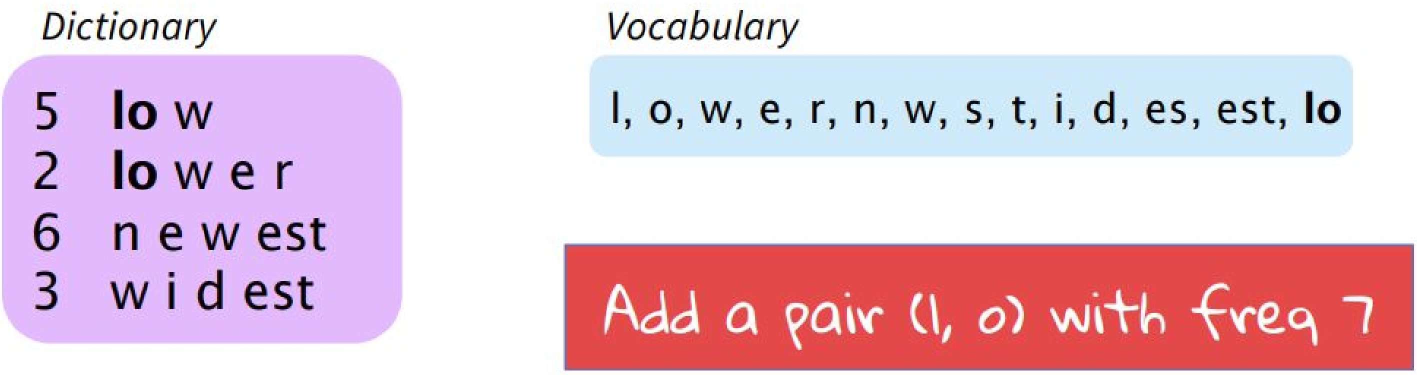
- 更强的基于character的LSTM seq2seq模型

Sub-word Model 子词模型趋势

- 与word-level模型采用同样架构
- 主模型是word-level，但一部分用于character

Byte Pair Encoding [Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural Machine Translation of Rare Words with Subword Units. ACL 2016.]

- 从单字符出发，逐步归纳最常见的字符串补充入词表



Wordpiece/Sentencepiece model

- Google NMT(GNMT)
- V1: wordpiece model
- V2: sentencepiece model
- BERT就是使用了wordpiece model的变种

Character-level to build word-level

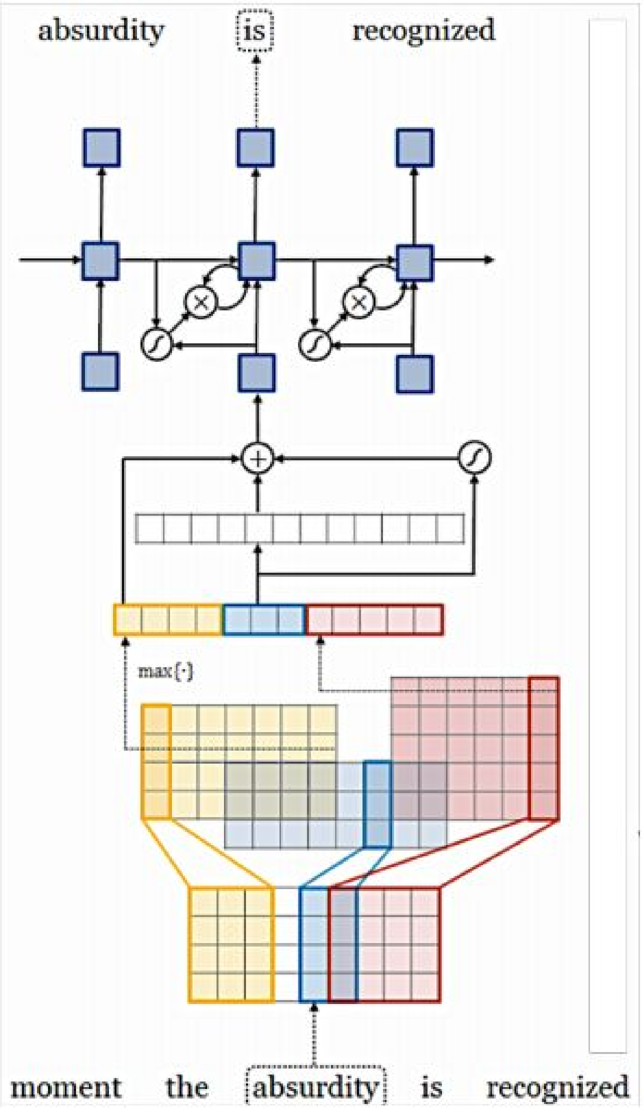
- [Learning Character-level Representations for Part-ofSpeech Tagging (Dos Santos and Zadrozny 2014)]
- 针对character进行卷积，生成word embedding
- 固定窗口的word embedding可以用于词性标注PoS tagging

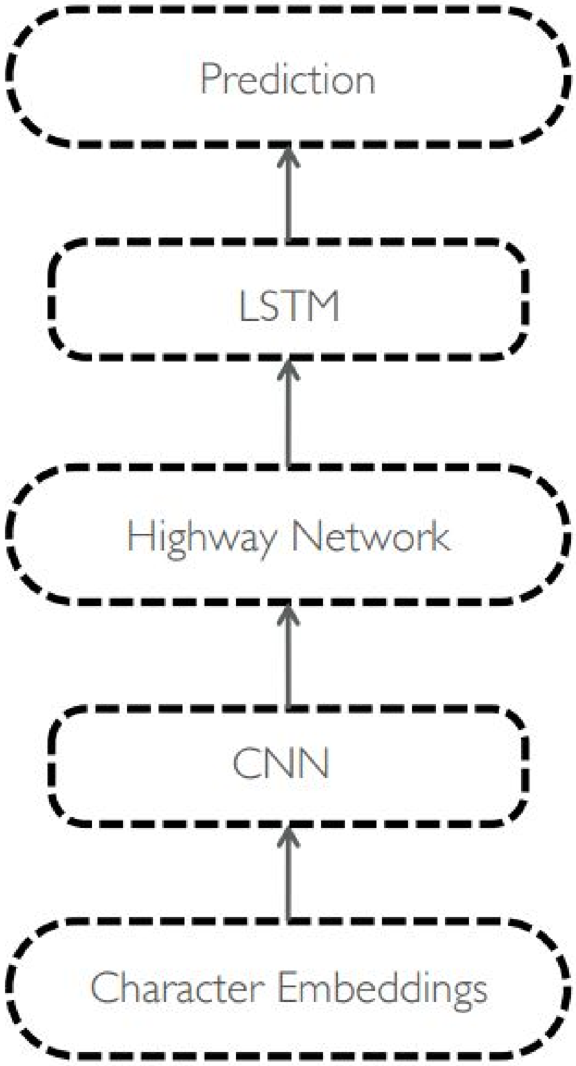
Character-based LSTM to build word representations

- Ling, Luís, Marujo, Astudillo, Amir, Dyer, Black, Trancoso. Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation. EMNLP’15
- 基于character的双向LSTM，双向的最终状态拼接，作为word representations
- 然后将word representations放入RNN作为LM

Character-Aware Neural Language Models [Yoon Kim, Yacine Jernite, David Sontag, Alexander M. Rush. 2015]

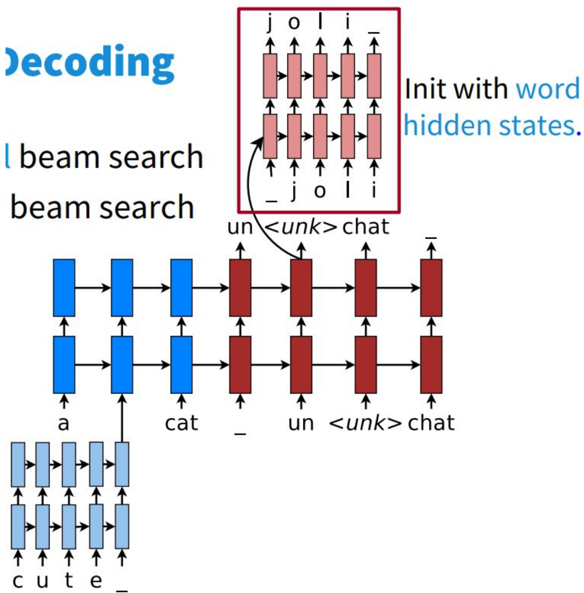
- 字符embedding连接=>用多个大小的卷积核进行卷积=>Max-over-time pooling=>Highway network=>LSTM=>Softmax





Hybrid HMT [Thang Luong and Chris Manning. Achieving Open Vocabulary Neural Machine Translation with Hybrid Word-Character Models. ACL 2016.]

- 大部分情况基于word，当处理未收录词时进入character层面



Chars for word embeddings [A Joint Model for Word Embedding and Word Morphology (Cao and Rei 2016)]

- 类似于word2vec，使用双向LSTM计算embedding

FastText embeddings

- [Enriching Word Vectors with Subword Information Bojanowski, Grave, Joulin and Mikolov. FAIR. 2016.]
- 类似word2vec，但采用了morphology，对于罕见词更好
- 针对word2vec的skip-gram模型，增加了character n-grams
- 例如where = <wh, whe, her, ere, re>, <where>，采用各部分加和作为word representations