OSC 源创会
Opensource Innovation Meetup

IT大咖说
知识共享平台

2017源创会年终盛典

与电子标准院共建开源标准

12月23日　北京万豪酒店

主办方 开源中国 oschina.net

CESI

视频分析的进展

张辉

李飞飞 (Fei-Fei Li)

Cloud 人工智能和机器学习首席科学家

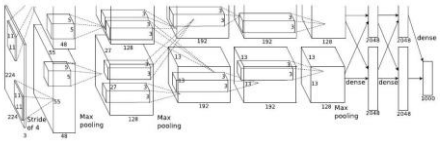
李飞飞博士是 Google Cloud 人工智能和机器学习团队的首席科学家。她还是斯坦福计算机科学系的终身教授和斯坦福人工智能实验室的主任。李飞飞博士的主要研究领域是机器学习、深度学习、计算机视觉与认知以及计算机神经科学。她曾在各类顶级期刊和会议（包括《自然》、《美国国家科学院院刊》、《神经科学期刊》(Journal of Neuroscience)、《计算机视觉与模式识别会议》(CVPR)、《国际计算机视觉大会》(ICCV)、《神经信息处理系统大会》(NIPS)、《欧洲计算机视觉大会》(ECCV)、《国际计算机视觉期刊》(IJCV)、《IEEE 模式分析与机器智能汇刊》(IEEE-PAMI) 等）中发表了 150 多篇科学论文。李飞飞博士于 1999 年以优异成绩从普林斯顿大学获得物理学士学位，2005 年从加州理工学院 (Caltech) 获得电子工程博士学位。她于 2009 年以助理教授身份加入斯坦福大学，并于 2012 年晋升为终身教授。在此之前，她曾先后在伊利诺伊大学厄巴纳-香槟分校 (2005-2006) 和普林斯顿大学担任讲师 (2007-2009)。李博士是 ImageNet 和 ImageNet Challenge 的创始人，前者是一个非常重要的大规模数据集，后者则是一个标杆分析项目，它们推动了深度学习和人工智能领域的最新发展。除了技术上的贡献之外，她还是提倡 STEM 和 AI 多元化发展的先锋。她是斯坦福大学针对高校女生的著名 SAILORS 外展计划和国家级非盈利性 AI4ALL 计划的联合发起人。在 AI 领域，李博士是 TED2015 主会场的演讲嘉宾，曾荣获 IAPR 2016 J.K. Aggarwal 奖、2016 nVidia AI 先锋奖、2014 IBM Faculty Fellow 奖、2011 Alfred Sloan Faculty 奖、2012 Yahoo Labs FREP 奖、2009 NSF CAREER 奖、2006 Microsoft Research New Faculty Fellowship 奖和多项 Google 研究奖。李博士实验室的工作成果已经刊发在多家广受欢迎的媒体杂志和新闻报纸上，包括《纽约时报》、《华尔街日报》、《财富杂志》、《科学》、《连线》杂志、《麻省理工科技评论》、《金融时报》等等。她曾被《ELLE》杂志评选为 2017 科技界女强人，曾荣获《Good Housekeeping》颁发的 2017 卓越女性奖、美国《外交政策》杂志评选的 2015 全球思想家，同时还曾入选卡内基教学促进基金会 2016 "杰出移民：美国的骄傲"，该奖项的前任获奖者当中包括阿尔伯特·爱因斯坦、马友友、谢尔盖·布林等。

12月13日，谷歌云首席科学家，斯坦福人工智能实验室主任
李飞飞博士在2017年中国谷歌开发者大会上宣布，
谷歌AI中国中心正式成立。

Google Developer Days China 2017

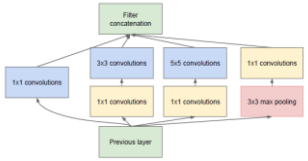
ImageNet数据库

历年ImageNet大规模视觉识别挑战赛ILSVRC冠军模型

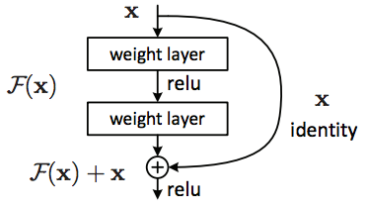


AlexNet, 2012
26.2%

ZFNet, 2013
11.2%



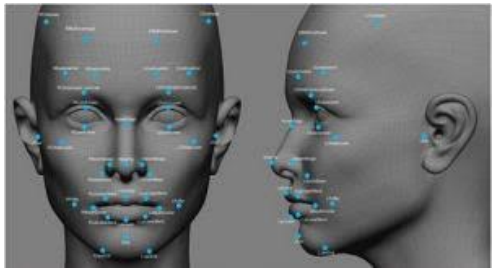
Google Inception, 2014
6.7%



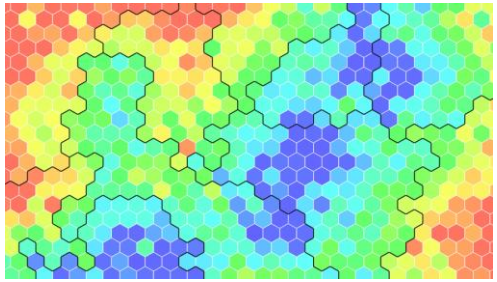
ResidualNet, 2015
3.6%

传统机器学习实现分类任务

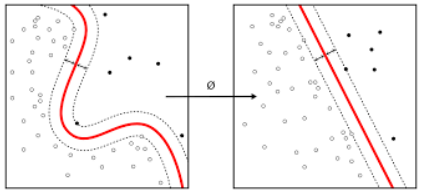
特征提取

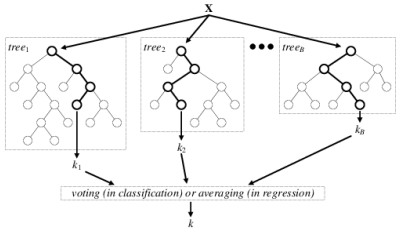


特征选择



特征分类

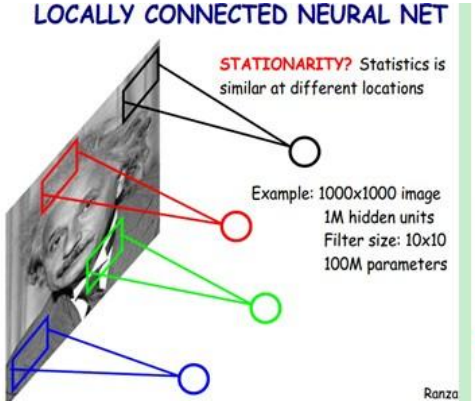


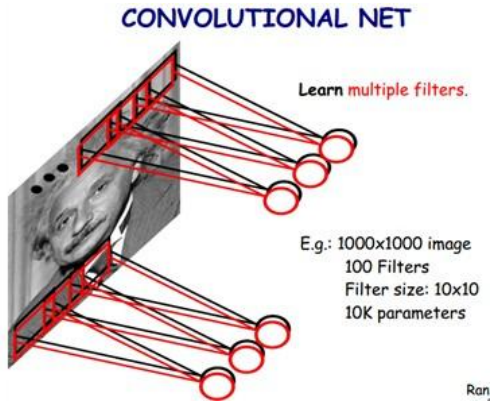


深度卷积神经网络的出现

1. 权值共享的卷积核特征提取方式

2. 反向梯度回传的最优化参数更新方式



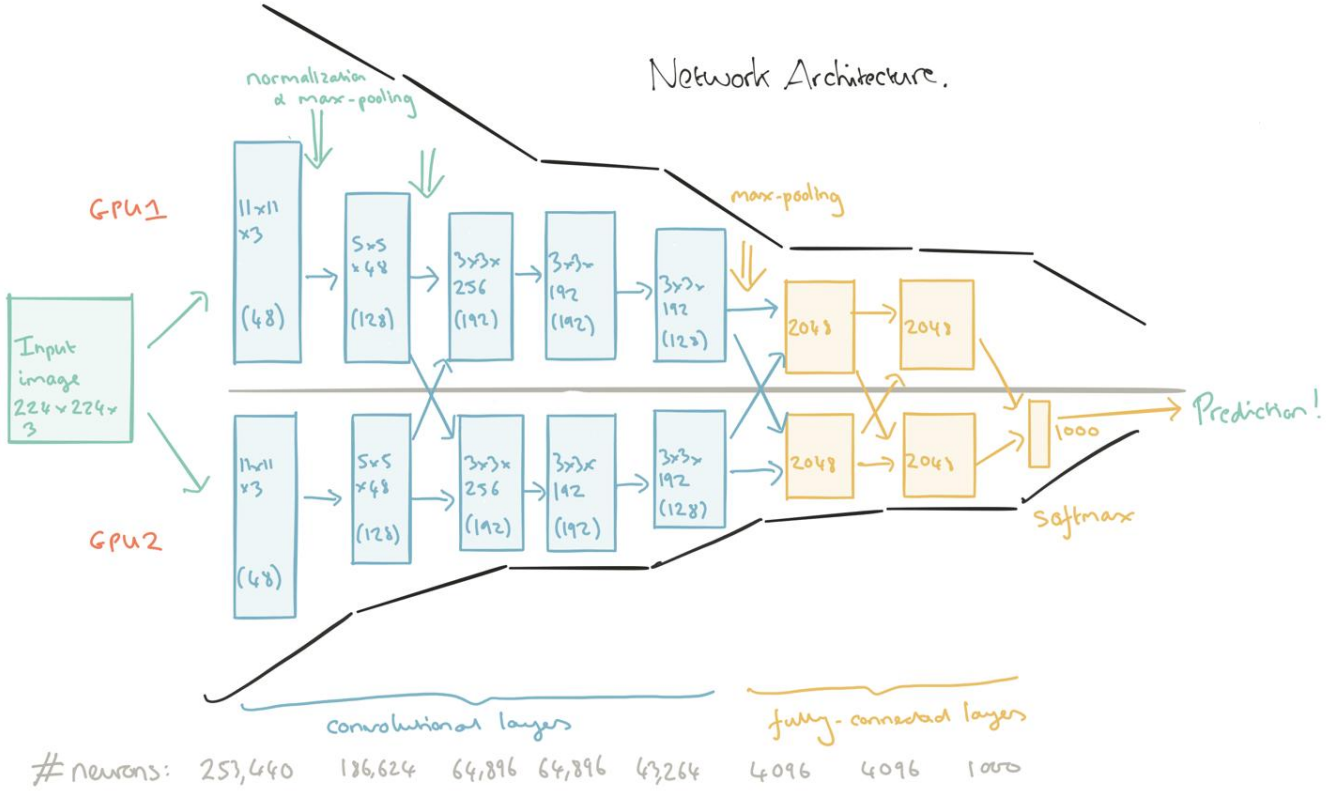


深度卷积神经网络的提出者Yan Lecun于1988年创新性提出了LeNet网络结构

[1] 对于图片不同位置的区域，采用同一卷积核进行特征提取；

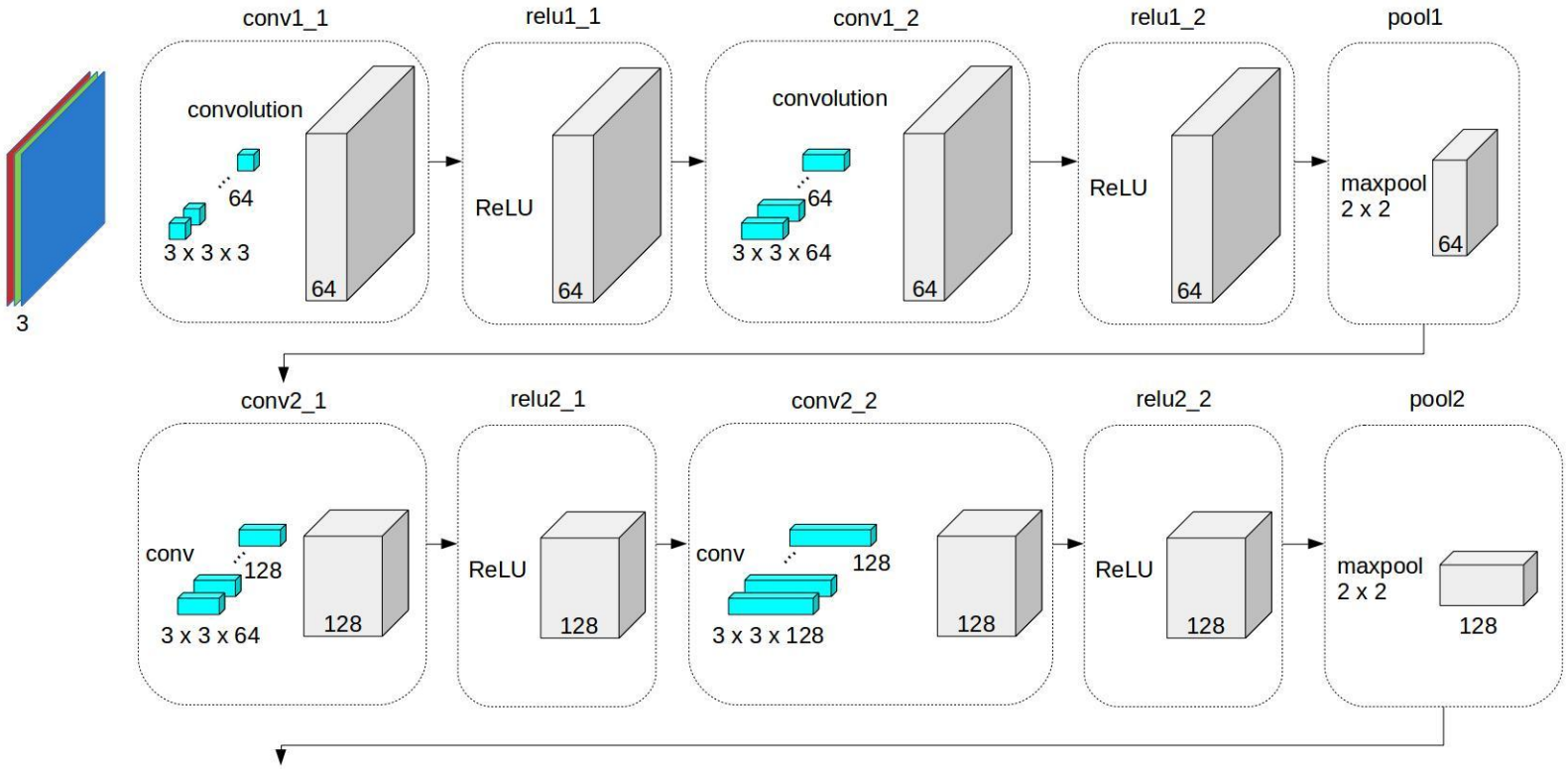
[2] 同时设计多通道的卷积核，提取图片不同维度方向的特征。

深度卷积神经网络的爆炸式发展



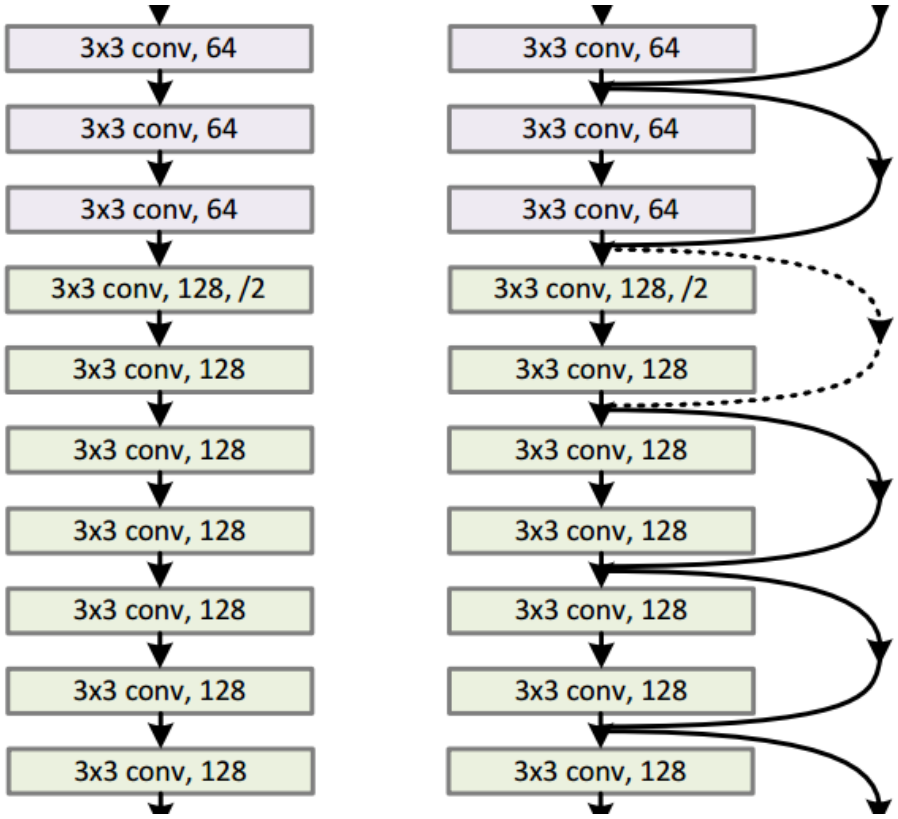
2012年由Krizhevsky提出的AlexNet实现了ImageNet数据库上26.2%错误率

VGG



VGGNet论证了采用小卷积核，加深网络深度的策略增加了系统的非线性的同时，也提升了卷积神经网络对于复杂问题的拟合性能。

Residual Net



ResNet在VGG的基础上引入了Residual Block的概念，
解决了由于神经网络的加深而带来了训练饱和与性能下降的问题

视频识别- Youtube-8M视频数据库

YouTube

YouTube-8M is a large-scale labeled video dataset that consists of millions of YouTube video IDs and associated labels from a diverse vocabulary of 4700+ visual entities. It comes with precomputed state-of-the-art audio-visual features from billions of frames and audio segments, designed to fit on a single hard disk. This makes it possible to get started on this dataset by training a baseline video model in less than a day on a single machine! At the same time, the dataset's scale and diversity can enable deep exploration of complex audio-visual models that can take weeks to train even in a distributed fashion.

Our goal is to accelerate research on large-scale video understanding, representation learning, noisy data modeling, transfer learning, and domain adaptation approaches for video. More details about the dataset and initial experiments can be found in our technical report. Some statistics from the latest version of the dataset are included below.

7 Million	450,000	3.2 Billion	4716	3.4
Video URLs	Hours of Video	Audio/Visual Features	Classes	Avg. Labels / Video

The videos are sampled uniformly to preserve the diverse distribution of popular content on YouTube, subject to a few constraints selected to ensure dataset quality and stability:

- Each video must be public and have at least 1000 views
- Each video must be between 120 and 500 seconds long
- Each video must be associated with at least one entity from our target vocabulary
- Adult & sensitive content is removed (as determined by automated classifiers)

视频识别- Youtube-8M视频数据库



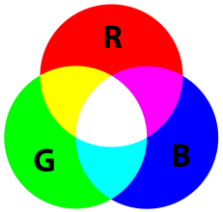





Youtube-8M数据库提供了4716种场景，为视频中的行为和事件检测提供了数据基础

视频识别

帧特征提取



RGB

Optical Flow

Audio

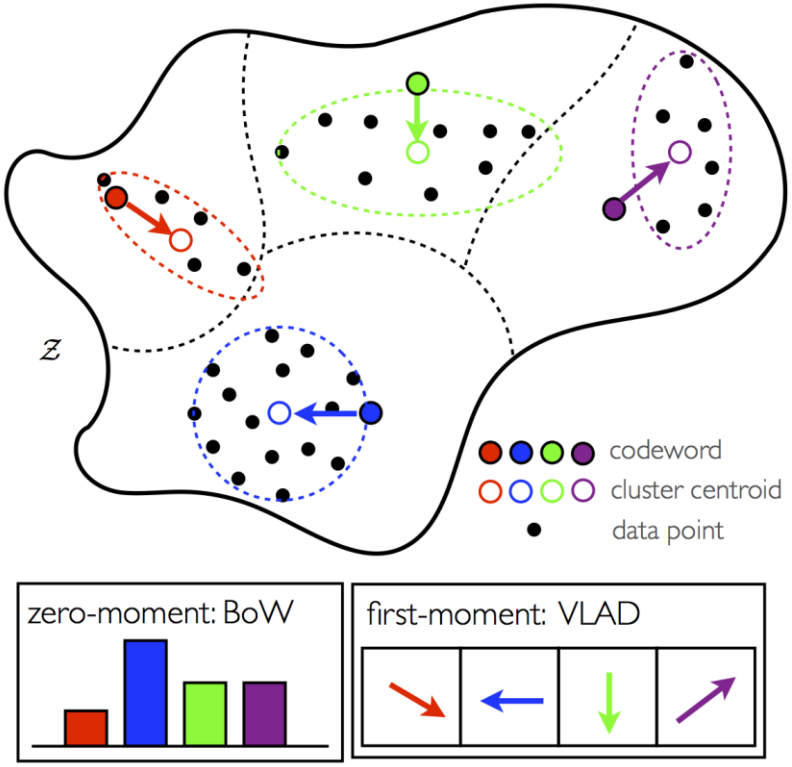
帧间关系 — Bidirectional LSTM

注意力模型 — Attention Model

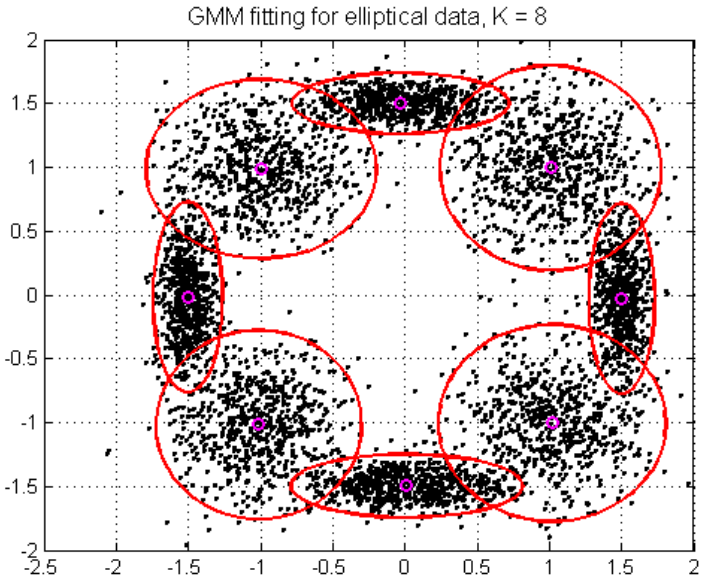
帧特征聚集 — Feature Aggregation

视频特征分类 — Sigmoid + Classification

视频识别



VLAD



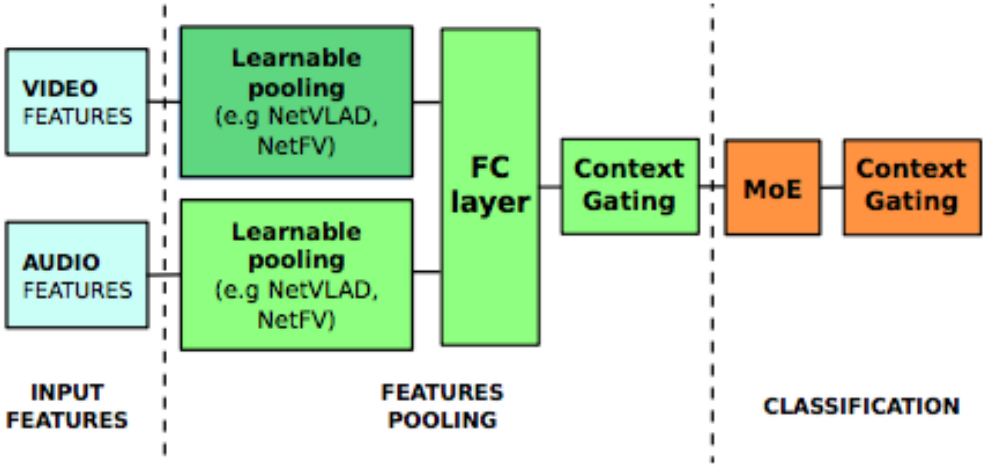
GMM-> Fisher Vector

视频识别

帧特征聚集

Context Gating

视频特征分类



最新的Youtube-8M数据库文章“Learnable Pooling with Context Gating for Video Understanding”
采用Context Gating决定哪一部分的特征
对于视频的识别分类是最有效的

深度学习文本摘要自动生成

- 文本摘要自动生成发展背景
- 基于排序算法的抽取式文本摘要
- 基于深度学习的生成式文本摘要
  - Sequence to Sequence（编解码架构）
  - Attention机制
  - Convolution Seq2seq
  - All Attention Seq2seq
  - 提升文本摘要性能的一些Tricks
- 文本摘要自动生成展望

文本摘要自动生成发展背景

- **标题党：** 14亿人都不知道的真相，历史的血泪…”，” 删前速看！，XXX视频流出”等，往往点进去的跟想像的大失所望。
- **新闻、媒体工作者：** 每天都要面对几十、上百篇新闻稿，通读文章，人工提炼关键短语，反复思考，才能形成一个能反应全文思想的标题。
- **如果机器可以先帮我们阅读一遍文章，生成合适的标题，无疑将会带来巨大的便利！！**

源文：

海湾报刊对美国新当选总统克林顿，能否帮助振兴中东和平进程感到怀疑，但也确实看到了一丝希望。

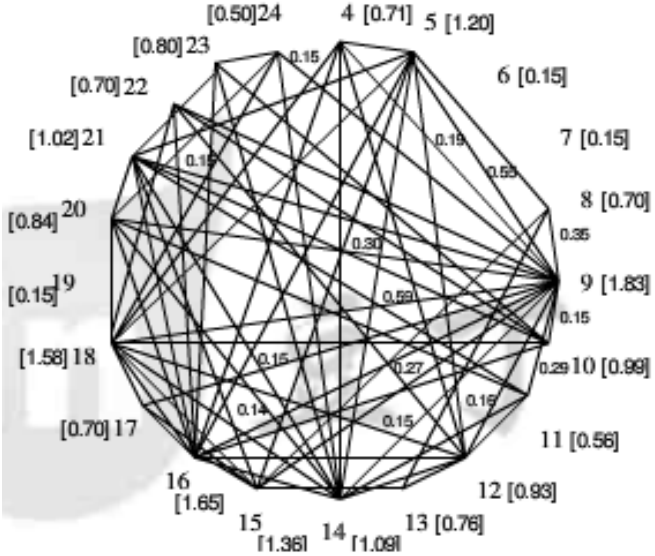
==========

人类：海湾报刊对克林顿是否会恢复和平进程，持怀疑态度

机器：海湾新闻界对克林顿恢复和平进程的前景，持怀疑态度

基于排序算法的抽取式文本摘要

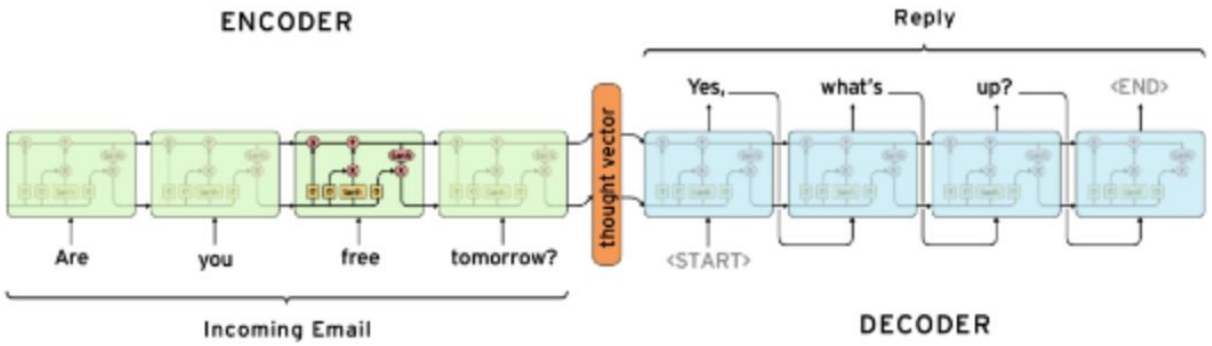
抽取式：顾名思义，就是从原文中抽取一些句子，来代表中心思想。通常把每个句子视为下图中的一个结点。迭代计算某一句子与其他句子的相似程度，相似度越高，则权重越高，最终抽取出那些权重排名较高的句子。代表算法：Text rank。



缺点：语义衔接不连贯，无法跟人类总结的句子相提并论

基于深度学习的生成式文本摘要

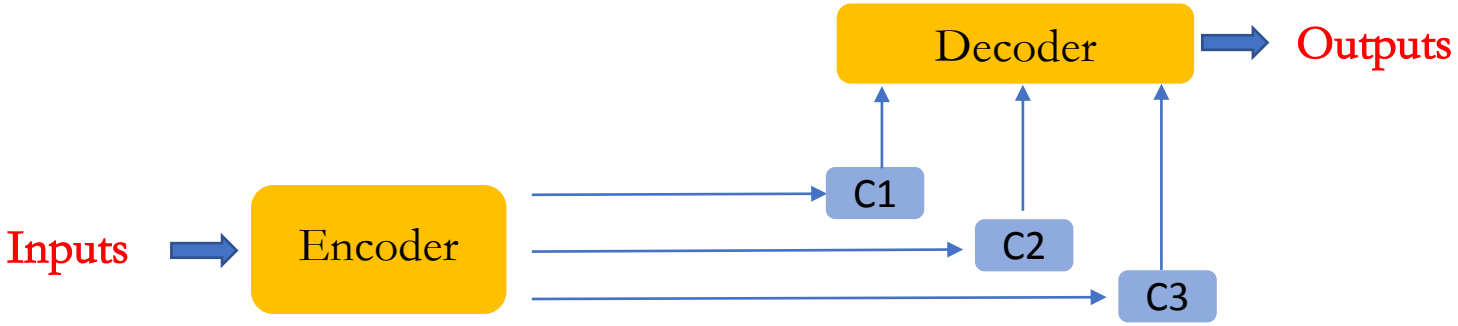
生成式：是计算机通读原文，在理解整篇文章句子意思的基础上，按自己的话，重新生成流畅的翻译。伴随着深度学习的研究，生成式摘要对质量和流畅度都有很大的提升。

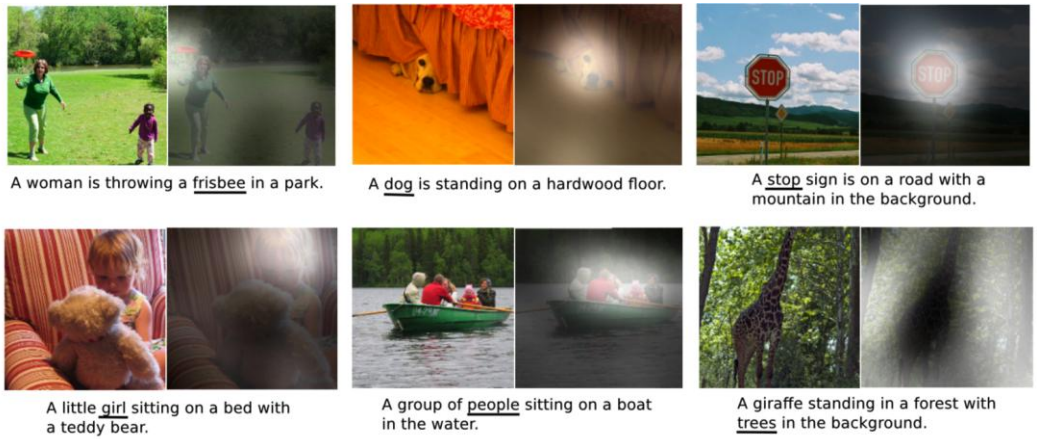


生成式：2014年，Google Brain团队提出的Sequence-to-Sequence序列，开启了NLP中端到端的火热研究。其中，编解码器的神经网络单元为RNN、LSTM

Attention机制

定义：Attention机制是一种注意力（资源）分配机制，在某个特定时刻，总是特地关注跟它相关的内容，其他内容则进行选择性忽视。

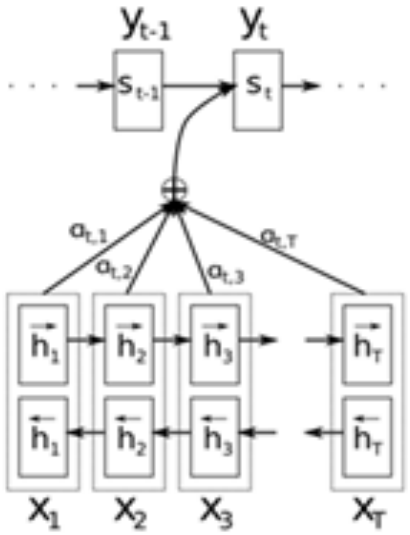




2014年，Bahdanau等人提出的《Neural Machine Translation by Jointly Learning to Align and Translate》

Attention机制

Learning to Align and Translate



$$p(y_i|y_1,\ldots,y_{i-1},\mathbf{x}) = g(y_{i-1}, s_i, c_i)$$

$$s_i = f(s_{i-1}, y_{i-1}, c_i)$$

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j.$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})},$$

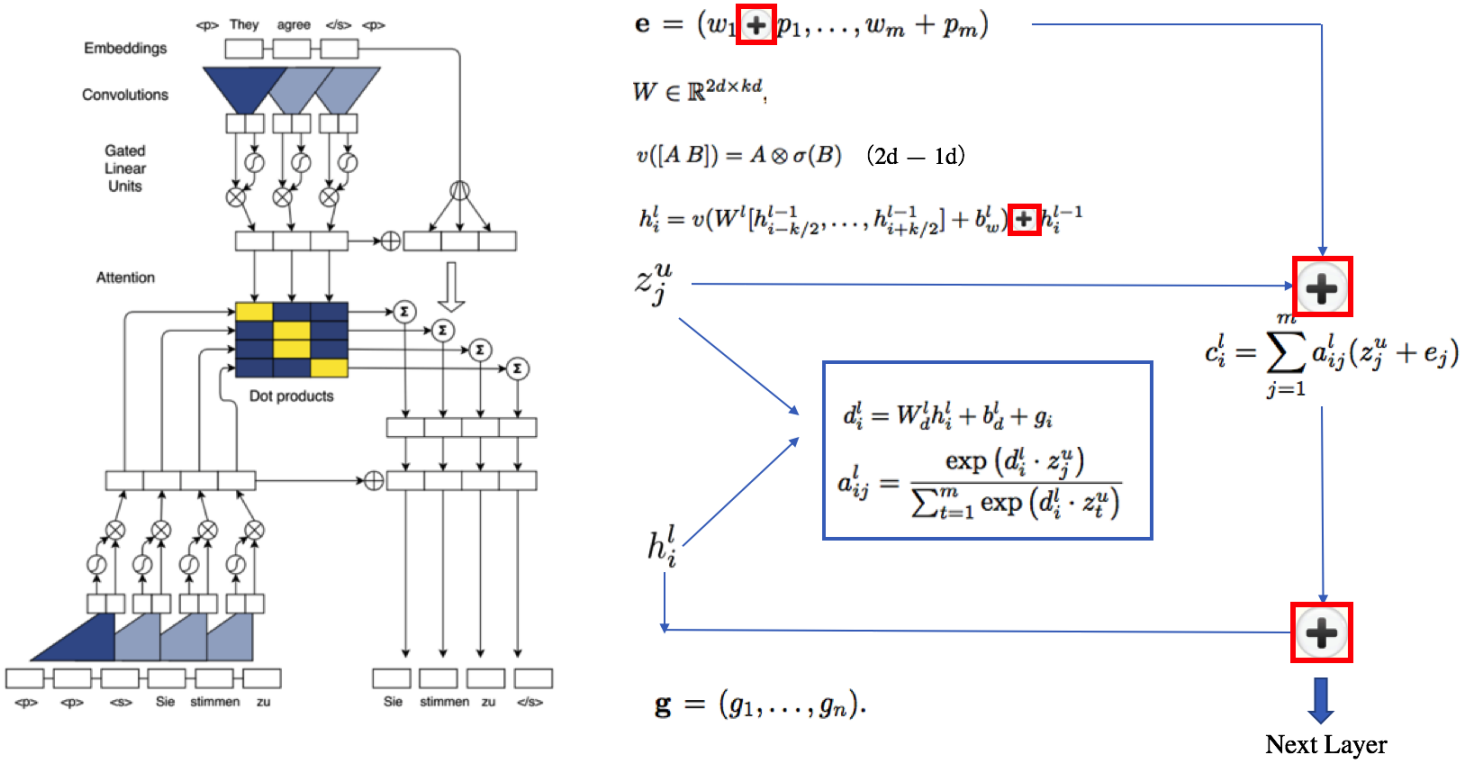
$$e_{ij} = a(s_{i-1}, h_j) = v_a^\top \tanh(W_a s_{i-1} + U_a h_j)$$

$$a_t = align(m_t, m_s) = \frac{\exp(f(m_t, m_s))}{\sum_{s'} \exp(f(m_t, m_{s'}))}$$

$$f(m_t, m_s) = \begin{cases} m_t^\top m_s & \text{dot} \\ m_t^\top W_a m_s & \text{general} \\ W_a[m_t; m_s] & \text{concat} \\ v_a^\top \tanh(W_a m_t + U_a m_s) & \text{perceptron} \end{cases}$$

Convolution Seq2seq

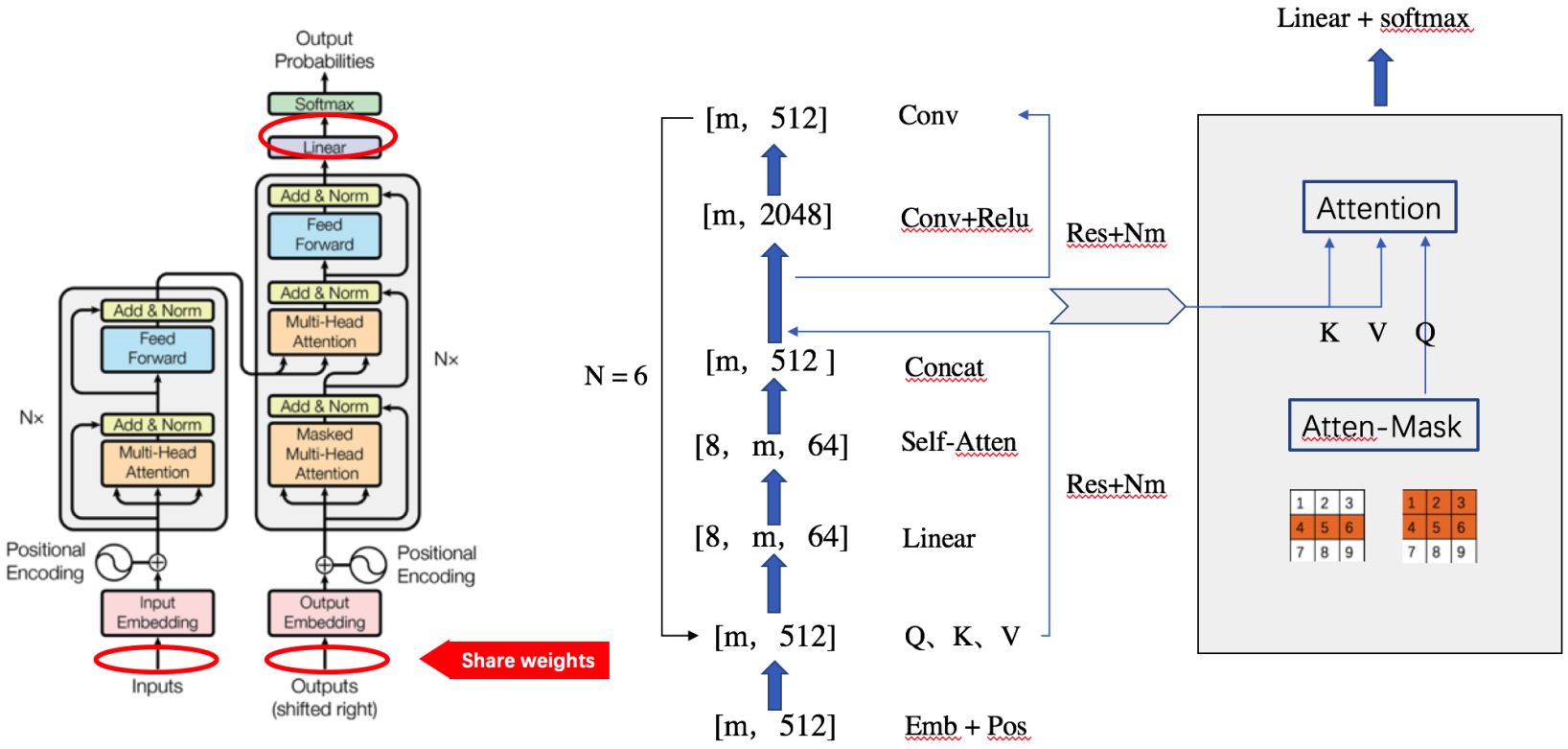
定义：由于RNN/LSTM单元，由于每个词是按顺序输入网络的，所以会记录文章的序列信息。但同时无法进行并行计算，这也限制了网络训练及摘要生成的速度。2017年5月，Facebook实现了Encoder、Decoder都采用CNN单元，使得网络在训练阶段，可以并行计算，效率进一步提升。



17年5月，Facebook团队提出的《 Convolutional Sequence to Sequence Learning 》

Attention Seq2seq

定义：17年6月，Google团队只用Self-Attention和Encoder-Decoder Attention，就完全实现了端到端的翻译任务，并且也在WMT-14英法翻译任务中，BLEU值达到了41.0的高分，因为同样可以并行计算，模型的训练及生成速度也有所提升。



17年6月，Google团队提出的《Attention Is All You Need》

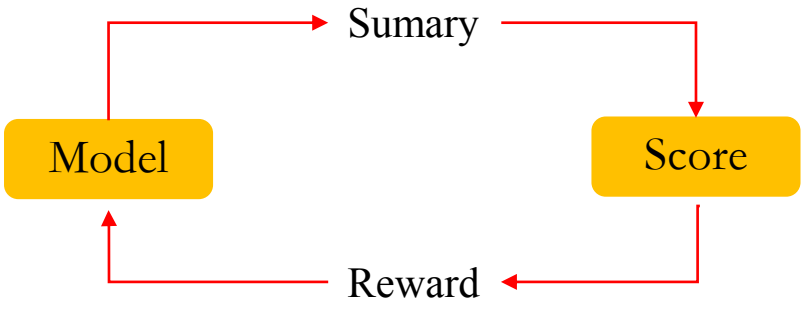
提升文本摘要性能的Tricks

BPE编码： NLP任务中，由于vocab数量是有限的，对于模型未见过的词，便会生成像UNK这样无法解析的词汇。BPE压缩技术可以对英文字母、中文字重新进行排列组合，以生成新的词汇。

lo w → low
e r· → er·

Reranking： 模型在训练阶段，以最大似然函数(MLE)作为学习目标，但这样概率式的学习方式缺乏明确的语义、句法等信息，通常会造成生成的摘要重复词较多的现象。通过人为限制重复词的数量N，或者结合N-Gram语言信息进行校验，重新输出正确的摘要

强化学习： 对于一篇文本，往往可以有语义相似的不同摘要，而训练时的MLE原则要求太过绝对，摘要评测时采用的ROUGE评测指标，却能考虑到这一灵活性。但由于ROUGE不可导，便想到可以采用强化学习对生成的摘要进行干预。



文本摘要自动生成展望

小结： 从传统的Textrank抽取式，到深度学习中采用RNN、CNN单元处理，再引入Attention，Self-Attention，机器生成摘要的方式，跟人类思维越来越像，都建立在对整段句子的理解之上。生成摘要的效果，也常常让我们拍案叫绝。接下来，我们将会在下面三个方面展开工作。

- **长文本摘要：** 改进句子级别的Attention，增强神经网络对长文本的记忆能力。
- **多文档摘要：** 从多篇文章、网页中提取出关键信息，并进行归纳总结
- **无监督生成摘要：** 无需借助大量有标签样本，适合预料资源匮乏的场景