预训练思想源于图像领域（通过大量数据预训练模型，再经冻结 / 微调适配小数据任务），NLP 领域预训练技术从Word Embedding起步，历经统计语言模型（解决词概率计算，存在数据稀疏）、神经网络语言模型（引入词向量缓解稀疏）、Word2Vec（CBOW/Skip-gram 专注词向量）、RNN/LSTM（处理时序但 LSTM 仅支持 200 词长）、ELMo（双向 LSTM 实现动态词向量）、Attention（解决长距离依赖与并行计算）、Transformer（Self-Attention 叠加 + Positional Embedding）、GPT（Transformer Decoder + 单向预训练 + Fine-tuning 模式），最终发展到BERT。BERT 作为集大成者，以Transformer Encoder为特征提取器，通过语言掩码模型（MLM，15% 掩码比例） 和下句预测（NSP） 完成无监督预训练，再经微调适配句对分类、单句分类、文本问答、单句标注四类下游任务，刷新 11 项 NLP 任务性能，凭借效果与普适性成为里程碑。