深度学习如何对化学反应进行分类？

深度学习已经和化学家们做了朋友。一些化学家已经开始用深度学习来预测化学反应结果，设计实验方案等。最近，IBM 和伯尔尼大学的研究人员利用一种称为 BERT 的深度神经网络，成功实现了对化学反应的分类，研究结果发表在今天（2021 年 1 月 28 日）的《自然-机器智能》杂志上[1]。

对化学反应进行分类非常重要：知道了反应的类别，化学家们就可以借鉴类似的反应来分析反应发生的优化条件，推测原子在反应中的重组方式等等。传统分类方法一般由专家们依自己的经验编写大量规则来实现；近年来，一些机器学习方法开始应用到这一领域，但都有一定的局限性。

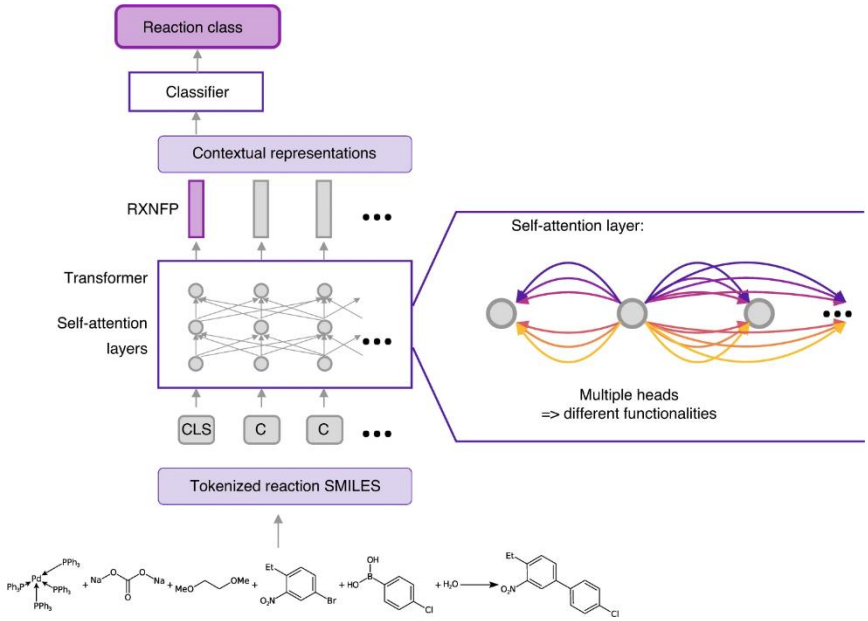

图 1. 用于化学反应分类的 BERT 模型。首先将反应方程写成 SMILES 格式，得到一个符号串，再经过若干个称为 Transformer 的神经网络结构，得到一个称为 RXNFP 的代表向量，基于该向量可实现对反应的类型进行判断。

IBM 和伯尔尼大学的研究者提出一种基于 BERT 模型的化学反应分类方法。BERT 模型在自然语言处理领域可以说是大名鼎鼎。基于一种称为“自注意力（Self Attention）“的机制，这一模型可以利用大规模文本数据学习语言的基础特性，如词与词之间的搭配关系，标点符号的作用，前后句之间的语义关联等等。

为了将这一模型应用到化学反应分类中，研究者首先将化学反应方程转化成一种称为” SMILES “的符号串格式，相当于设计了一门描述化学反应的语言，然后应用 BERT 模型来学

习这门语言，就和学习人的语言一样。学习完成之后，他们在 13.2 万个化学反应上做了测试，发现分类效果可达到 98.2%，而此前方法的性能只有 41.0%。

有趣的是，BERT 模型的注意力机制可以发现对反应的分类起关键作用的成分, 如图 2 所示。这就好比要理解一句话的意思，需要找到最有价值的单词一样。这一能力也解释了 BERT 为什么有这么好的分类效果。

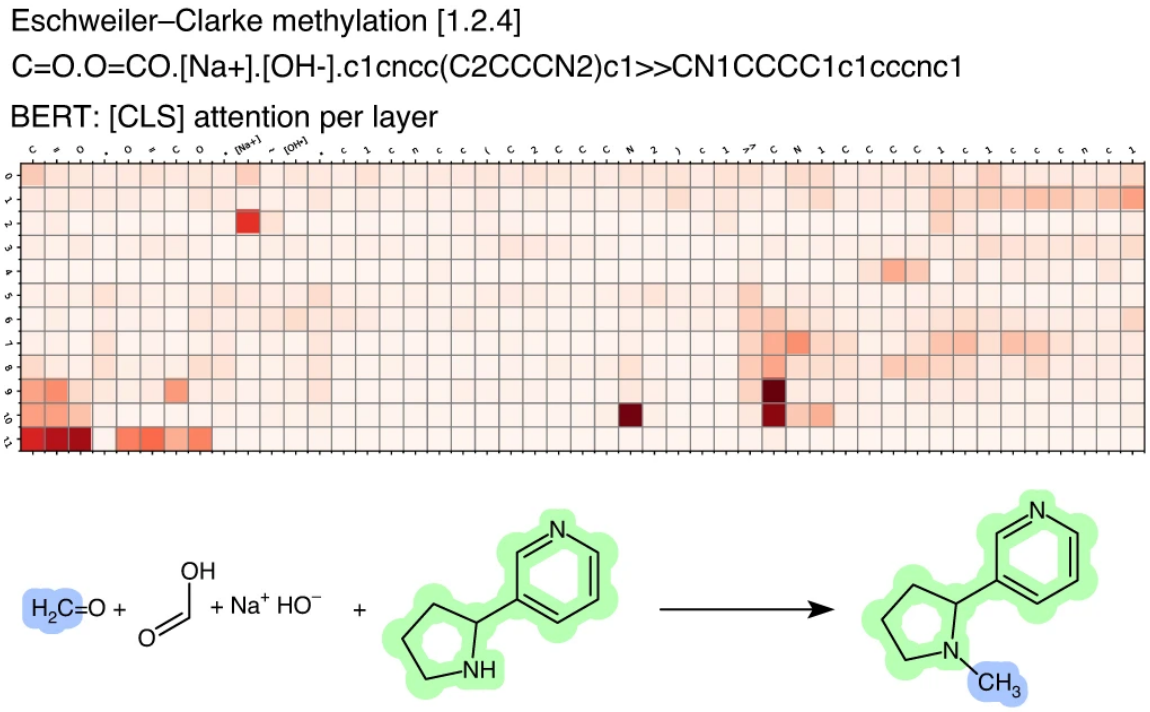

图 2. BERT 可以发现对分类最有价值的成分。第一行是反应类型。第二行是反应的 SMILES 格式。中间方格图是 BERT 模型的不同层（纵轴）对每一个成分（横轴）的关注程度，颜色越深表示关注度越高。下图是实际参与化学反应的成分。

最后，研究者发现利用 BERT，可以把一个化学反应表示成一个固定维度的向量，从而把各种化学反应和他们之间的关系画在一张图上，如图 3 所示。有了这张图，化学家们对各类化学反应的相关性就更清晰了，也许还可以发现很多以前没有关注到的秘密。

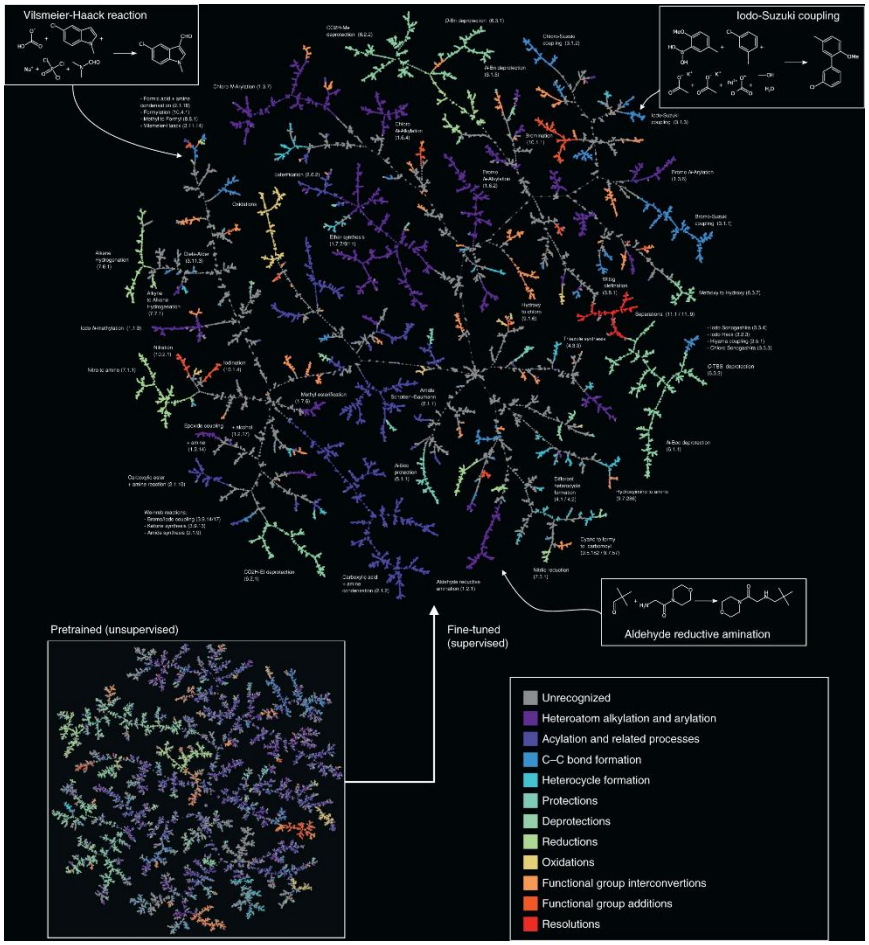

图 3. 利用 BERT 可以把各种化学反应表示在一幅图上。每一种有颜色代表一类化学反应。左下角图是基于预训练网络得到的结果，中间大图是基于训练后的网络得到的结果。

参考文献：

[1] Schwaller, P., Probst, D., Vaucher, A.C. *et al.* Mapping the space of chemical reactions using attention-based neural networks. *Nat Mach Intell* (2021).