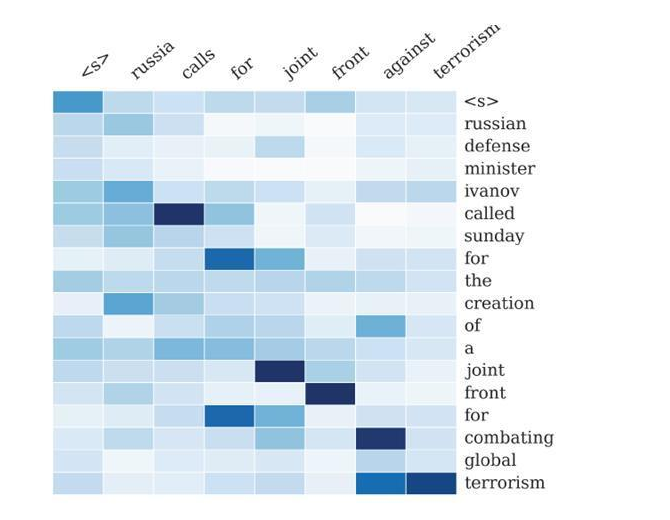
1. 简而言之，深度学习中的注意力可以广义地解释为重要性权重的向量：**为了预测一个元素，例如句子中的单词，使用注意力向量来估计它与其他元素的相关程度有多强，并将其值的总和作为目标的近似值。**
2. 而在NMT的翻译模型中经典的做法是由编码器 - 解码器架构制定(encoder-decoder)，用作encoder和decoder常用的是循环神经网络。这类模型大概过程是首先将源句子的输入序列送入到编码器中，提取最后隐藏状态的表示并用于解码器的输入，然后一个接一个地生成目标单词，这个过程广义上可以理解为不断地将前一个时刻 t-1 的输出作为后一个时刻 t 的输入，循环解码，直到输出停止符为止。

RNN是健忘的，这意味着前面的信息在经过多个时间步骤传播后会被**逐渐消弱乃至消失**。其次，在解码期间没有进行对齐操作，**因此在解码每个元素的过程中，焦点分散在整个序列中。**

Bahdanau等人提出在机器翻译任务上在 encoder–decoder 做出了如下扩展：将翻译和对齐联合学习。这个操作在生成Target序列的每个词时，用到的中间语义向量context是Source序列通过encoder的隐藏层的加权和，而传统的做法是只用encoder最后一个时刻输出作为context，这样就能保证在解码不同词的时候，Source序列对现在解码词的贡献是不一样的。

作者将Source序列的每个词(通过encoder的隐藏层输出)和Target序列 (当前要翻译的词) 的每个词巧妙的建立了联系。想一想，翻译每个词的时候，都有一个语义向量，而这个语义向量是Source序列每个词通过encoder之后的隐藏层的加权和。 由此可以得到一个Source序列和Target序列的对齐矩阵，通过可视化这个矩阵，可以看出在翻译一个词的时候，Source序列的每个词对当前要翻译词的重要性分布，这在直觉上也能给人一种可解释性的感觉。



**问题：使训练出这个矩阵吗？如何训练？如何预测？相当于每个词与vocab中所有词建立一个权重信息，如上图一列，然后把输入的一句话拼接成这样的一个矩阵，然后去翻译对应的单词？**

