APRIL 2021, 伍元凯

甜、攻击样本

1 什么是甜

人类作为一种哺乳动物，至少可以区分甜 (sweet)，酸 (sour)，咸 (salty)，苦 (bitter) 和鲜 (umami) 五种基本味道。从进化论角度来讲，人对甜味的喜爱主要源于甜味与高能量食物的关联性，含糖高的食物可直接提供给我们能量，为在自然界中生存提供帮助。实验表明单细胞的细菌都对含糖较高的食物具有倾向性，对于能快速使用的能量的感知能力可能很早就根植于生物系统之中。如果将生物系统视为一个基于表演者-评论家（Actor-Critic）的强化学习系统，我们可以简单的将奖励 (Reward) 和评论家估计的 Q-value 表示为：

$$R_t = E_t, \quad Q^c = \sum_{i=0}^{T} \gamma^i S_i. \tag{1}$$

其中 E_t 为生物在 t 时刻获得的快速可分解能量，S_i 为 i 时刻生物体感知的甜度。这当然只是一个简单的单目标的系统，人类已经是十分复杂的哺乳动物，对于与能量相关的糖，人只对个别单糖会有甜感，一些低聚糖也会有甜味，低聚糖能分解为葡萄糖，为新陈代谢提供能量，所以甜味与能量没有绝对的成正比的关系。

本人十分喜爱喝 0 度和健宜可乐 (图1)，这些无糖饮料里使用的是阿斯巴甜，一种能量极低的甜味剂。阿斯巴甜的发现过程颇有意思，James M. Schlatter1965 年在合成制作抑制溃疡药物时，他无意间舔到手指，**偶然**发现到阿斯巴甜具有甜味。阿斯巴甜的甜度比一般的糖甜约 200 倍，又比一般蔗糖含更少的热量；一克的阿斯巴甜约有 4 卡路里的热量，约和一克蔗糖相等，但由于阿斯巴甜的高甜度，放入 1/200 克的阿斯巴甜就可产生相同的甜度，可大量减少能量的摄入，降低肥胖风险。

Figure 1: 无糖饮料

目前已知的人类可以感知到甜味的物质包括蔗糖 (Sucrose,C12H22O11)，糖精 (Saccharin,C7H5NO3S)，三氯蔗糖 (Sucralose,C12H19Cl3O8)，甜蜜素 (cyclamate，C6H12NNaO3S)，阿斯巴甜 (Aspartame，C14H18N2O5) 以及索马甜（thaumatin）。这些物质化学成分各异，但人都会对其产生甜感，其中三氯蔗糖、糖精、甜蜜素和阿斯巴甜都是人工合成的物质，甜度都会远高于蔗糖，但除了阿斯巴甜以外，其它所有的甜味剂都或多或少被发现对人体健康有一定的影响。从某些角度来说，制造甜味剂就是为了愚弄人类的甜度感觉系统 Q^c，使其对某些物质做出过高的估计，对含有较高添加剂的物质进行消费，即做出非最优的行为。

哺乳动物的甜度感觉系统极其复杂，人的感觉刺激由多种味觉细胞构成，主要负责感受味觉的细胞为 T1R 细胞和 T2R 细胞，鲜味主要由 T1R1 和 T1R3 两种受体负责，而甜味由 T1R1 和 T1R2 两种受体负责，T2R 则主要用来识别苦味。不同甜味剂刺激的受体也是不同的，较为自然的甜味剂一般会同时刺激 T1R1 和 T1R2，而人造的甜味剂往往只会刺激一种受体。如果这些受体不被刺激，人类是无法感受到甜味的。而缺少某些受体会让动物对甜味的感受不同，人的 T2R16 对苯基-b-D-吡喃葡萄糖苷比较敏感，所以人类会觉得其比较苦，而老鼠没有这种受体，却会觉得苯基-b-D-吡喃葡萄糖苷比较甜 [2]。虽然 T1R 最早在口腔中被发现，但是这些受体在人类的肠胃系统、内分泌系统中也广泛存在，这些受体也可能对甜度进行调整。可以说，人对甜度的感受是多处细胞受体（多源信号）信号分析的结果。

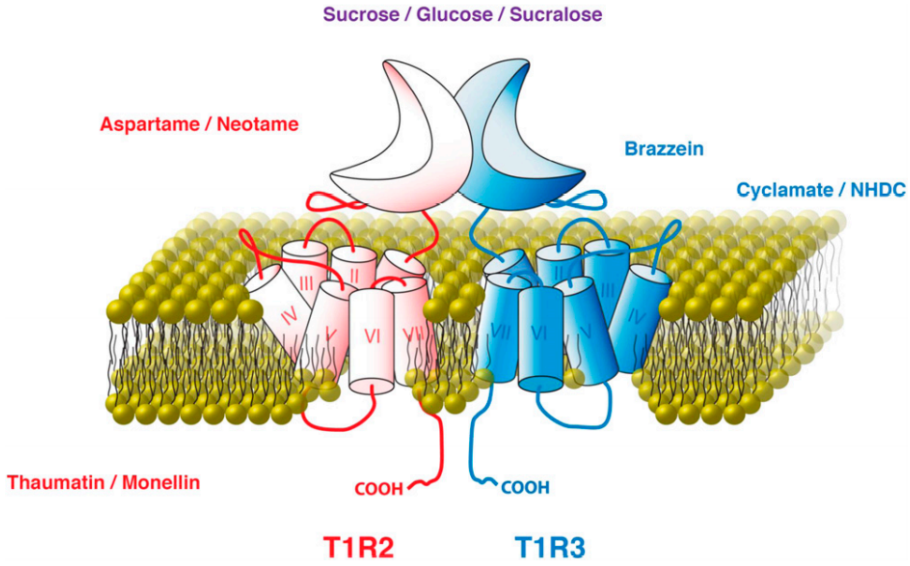

Figure 2: 不同糖刺激的受体并不相同, 阿斯巴甜/纽甜只刺激 T1R2；仙茅甜蛋白和甜蜜素则只刺激 T1R1；蔗糖，葡萄糖和三氯蔗糖则会刺激 T1R1 和 T1R2[1]

和视觉一样，味觉系统也是一个分层的系统（图3）。动物对甜度的感受可能和大脑中的奖励系统相关，也是动物进食的诱因之一。如果从数学的角度，动物对甜度的感受可表征为

$$Q^c = Q_\theta(I) \tag{2}$$

其中 I 是一个高维的向量，Q_θ 是一个深层的神经网络。这个网络有什么特点呢？第一，非常复杂，深度较深，有着一定的注意力机制（不同的通道对不同的物质激活机制不同）；第二，比较容易被愚弄，人类设计的一些化学成分会让 Q^c 输出不合理的值，进而影响动物的行为；第三，甜是一个复杂的过程，其本身是智能体和环境复杂交互的结果，甜无法用进化论（高能偏好）简单的定义，本身就是复杂神经网络的产物（图3）。

2 人工神经网络攻击样本的一种理解

深度神经网络可以被视为对人类分层感知系统（视觉、味觉和听觉）的模拟，如人造甜味剂对人类感官的“愚弄”一般，深层的神经网络也极易受到所谓人造对抗样本的攻击。对于对抗样本有许多种理解的方式，MIT 的一项研究认为对抗样本“是深度神经网络对数据本身自带的通用性特征的敏感性造成的”[3]。因为我们训练神经网络使得其尽可能的做出对标签准确的预测，深度神经网络会使用任何信号来帮助其做出“正确”的分类，即使这些信号对于人来说已经是不可识别的。

为了验证这一想法，MIT 的研究对数据集的所谓鲁棒性特征和非鲁棒性特征进行了分离，对于任何给定的数据集，他们的方法可以构建出一个鲁棒的训练集，即将数据集中的非鲁棒特征去掉。同样该方法也可以构建出一个非鲁

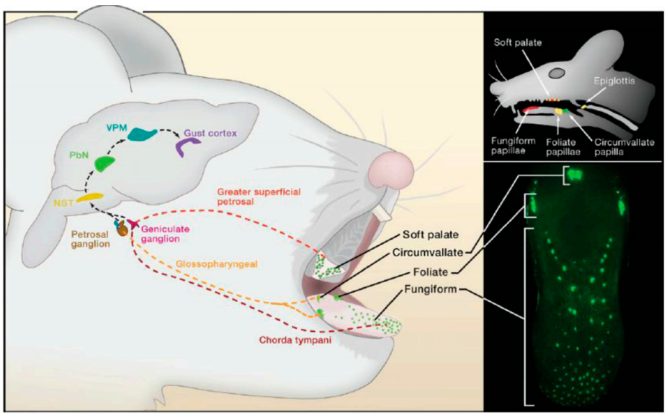

Figure 3: 老鼠的深度分层味觉感知系统

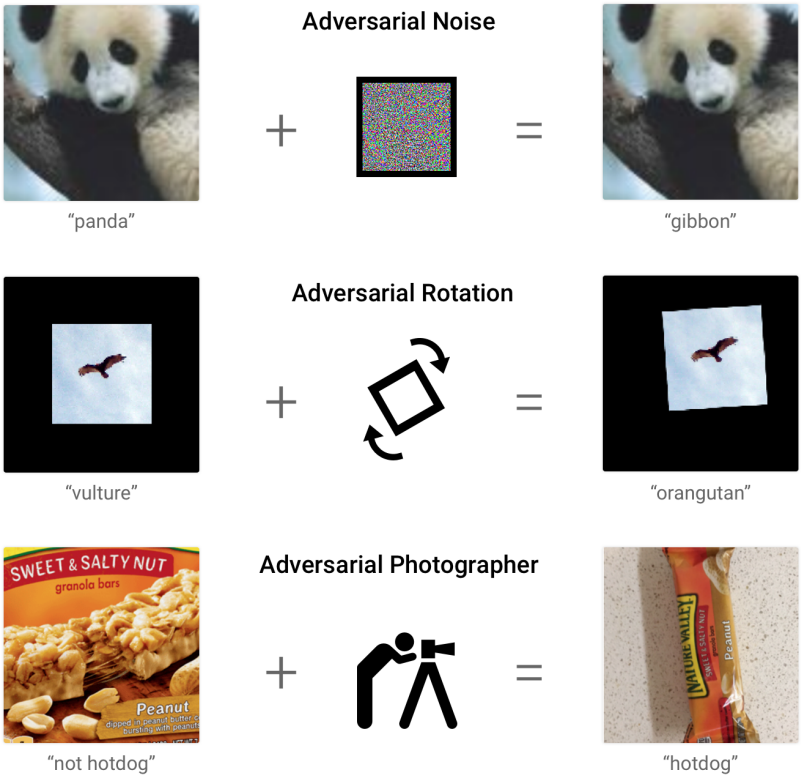

Figure 4: 对深层神经网络的对抗攻击

棒的数据集，具体的做法如图5所示。图5中的非鲁棒特征大多都是肉眼无法分辨类似于随机噪声的信号，正是这些数据集自带的对人无意义的特征容易受到攻击。而即使我们给训练集错误的标签（将狗标为猫），在测试集上深度神经网络仍可以利用狗样本中的非鲁棒特征成功识别出猫。作者所以认为，对抗样本是数据中自带的特征造成的，而并不一定是训练的方法造成的。

具体的鲁棒和非鲁棒特征分离过程此处不赘述，鲁棒数据集的构建可见论文 [3] 的 C.4，非鲁棒数据集则在附录 C.5。论文的实验结果表明人工深度神经网络可以依赖数据中非鲁棒的特征取得较好的泛化性，泛化性较好的模型有可能利用了数据集中更多的非鲁棒特征。但非鲁棒的特征对于人类来说是缺乏可解释性的，所以作者在结论中如此说道：对抗样本实际上是一种以人为中心的现象，从分类任务表现的角度来看，模型没有理由更偏好鲁棒性特征。毕竟，鲁棒性的概念是人类指定的，因为人类希望模型具有较高的解释性，所以希望模型只依赖于鲁棒性特征（即人类

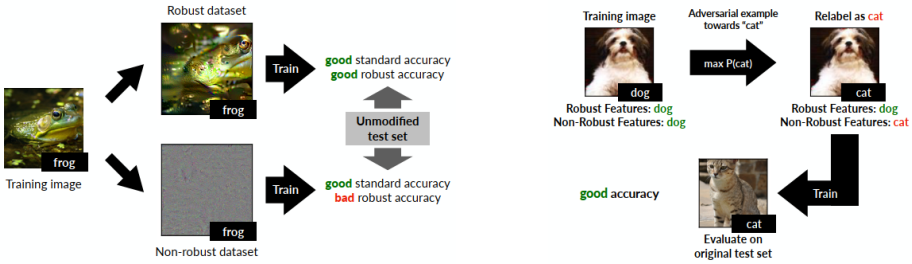

Figure 5: 鲁棒数据集和非鲁棒数据集的构建

可解释的特征)。但由于非鲁棒特征在各种数据集中的广泛存在，如果希望模型对人类更具有可解释性就需要在训练过程中加入人类的先验知识。

3 结语

在论文 [3] 中，作者定义有用的特征为

$$E_{(x,y)\sim D}[y \cdot f(x)] \geq p \tag{3}$$

D 是训练数据集，x 是输入，y 是标签，$f(x)$ 是深度神经网络生成的特征，该公式定义如果 $f(x)$ 和 y 的相关性大于 p，$f(x)$ 就是一个 p 有用的特征。

无论是人工的深度神经网络，还是我们人类的感官系统本身，都是一个极其复杂牵涉到海量参数的分层系统。$f(x)$ 拥有极强的多变性，x 中细微的，我们感官无法察觉的变化，仍然可以让 $f(x)$ 生成的特征和目标 y 产生关联性。当我们诉诸可解释性，往往是一套从语言角度来说让人可理解的陈述。当我们描述猫时" 脸圆圆的，有耳朵和胡子的哺乳动物"，“圆脸”，“耳朵”和“胡子”对于人类是可解释的，但神经网络在特定环境下被训练达到高精度目标时往往还需要其它不为人知的特征，这便是对抗样本的由来。

而回到甜度的问题，所谓的甜味剂和感官被“愚弄”的现象，其实也可以解释为一种对甜的简单解释。“喜爱甜食是因为生物需要获得快速的能量”，简单的基于进化论的解释往往会和现实中的某些观测不符。所以，世界对于人类来说有时也如同数据集对于人造的神经网络，蕴含了海量的不可解释的“非鲁棒性特征”。这也提醒我不要去寻求构建一个完美的模型，始终意识到自我和自我工作的可被愚弄性，保持开放的心态，去发现生活工作中的那些“非鲁棒性特征”，感受它们带给自己的体会和惊喜也是工作和生活的快乐源泉之一罢。

此外，我对人类本身对万事万物精确定义的能力也持怀疑态度。究竟什么是甜呢？它只是我这个智能体和环境交互过程中大脑产生的某种信号，这个交互过程牵扯着海量的“我”不知情的感官系统的参与，逐层抽象在大脑中形成的一种模式。新华字典中对甜的定义则为“像糖或蜜的滋味，喻使人感到舒服的，与“苦”相对”，这种解释本身就是只考虑到“鲁棒特征”的，如果将甜限定在简单的定义之中，往往会限制我们自身的想象力，即泛化能力。写到这里，我想起了诺兰评价他的电影《Tenet》的一句话"Don’t try to understand it. Feel it."。

4 参考文献

[1] Fernstrom, J. D., Munger, S. D., Sclafani, A., de Araujo, I. E., Roberts, A., & Molinary, S. (2012). Mechanisms for sweetness. The Journal of nutrition, 142(6), 1134S-1141S.

[2] Zhao, G. Q., Zhang, Y., Hoon, M. A., Chandrashekar, J., Erlenbach, I., Ryba, N. J., & Zuker, C. S. (2003). The receptors for mammalian sweet and umami taste. Cell, 115(3), 255-266.

[3] Ilyas, A., Santurkar, S., Engstrom, L., Tran, B., & Madry, A. (2019). Adversarial Examples Are Not Bugs, They Are Features. Advances in neural information processing systems, 32.