1.(a) Copyuing in attenion

i.

当查询向量 q 与某个特定的键向量 k_j 的点积（dot product）显著大于 q 与其他所有键向量 k_i ($i \neq j$) 的点积时（即 $k_j^\top q \gg k_i^\top q$），分布 α 会将几乎所有的权重集中在 α_j 上。

> $\alpha_i = \frac{\exp(k_i^\top q)}{\sum \exp(k_j^\top q)}$。**Softmax 具有放大差异的特性**，当某个输入值远大于其他输入值时，其对应的输出概率会趋近于 1，而其他概率会趋近于 0。

ii.

这种情况下，输出向量 c 将近似等于与该特定键 k_j 相对应的值向量 v_j。

1.(b) An average of two

取 $q = \beta(k_a + k_b)$ ，其中 β 是一个较大的正数标量

证明：

利用题目给出的正交性（$k_i^\top k_j = 0$）和单位长度（$k_i^\top k_i = 1$）：

$k_a^\top q = k_a^\top \beta(k_a + k_b) = \beta(k_a^\top k_a + k_a^\top k_b) = \beta(1 + 0) = \beta$

$k_b^\top q = k_b^\top \beta(k_a + k_b) = \beta(k_b^\top k_a + k_b^\top k_b) = \beta(0 + 1) = \beta$

$k_i^\top q = k_i^\top \beta(k_a + k_b) = \beta(k_i^\top k_a + k_i^\top k_b) = \beta(0 + 0) = 0,\ i \neq a, b$

将上面的点积代入：

$\alpha_a = \frac{e^\beta}{e^\beta + e^\beta + \sum_{others} e^0} = \frac{e^\beta}{2e^\beta + (n-2)}$

$\alpha_b = \frac{e^\beta}{2e^\beta + (n-2)}$ （与 α_a 相同）

当 β 较大时，e^β 会远远大于常数 $(n-2)$。

$\alpha_a = \alpha_b \approx \frac{e^\beta}{2e^\beta} = \frac{1}{2}$

$\alpha_i = \frac{1}{2e^\beta + (n-2)} \approx 0,\ i \neq a, b$

所以$c \approx \frac{1}{2}(v_a + v_b)$

1.(c) Drawbacks of single-headed attention

i.

类似地，取 $q = \beta(\mu_a + \mu_b)$

因为 α 极小，随机采样的 k_i 会非常接近其均值 μ_i，即 $k_i \approx \mu_i$。

ii.

定性描述：

输出向量 c 不再稳定地等于 $\frac{1}{2}(v_a + v_b)$，而是会在 接近 v_a 和 接近 v_b 之间剧烈摆动（oscillate）。换句话说，在某些样本中 $c \approx v_a$，而在另一些样本中 $c \approx v_b$，很少会停留在中间的平均值状态。

详细分析

- k_b 是稳定的： k_b 的方差很小，所以 $k_b^\top q \approx \beta$。
- k_a 是不稳定的： k_a 在 μ_a 方向上有很大的方差。这意味着采样出来的 k_a 可以写成 $k_a \approx \gamma\mu_a$，其中 γ 是一个在 1 附近波动的系数。

  所以 $k_a^\top q \approx (\gamma\mu_a)^\top \beta(\mu_a + \mu_b) = \gamma\beta$
- Softmax 的放大效应：由于 β 很大（在第 i 问为了过滤无关项设定的）：
  - 如果 γ 稍微**大于** 1（k_a 稍微长一点），则 $e^{\gamma\beta} \gg e^{\beta}$，权重会几乎全部分配给 v_a。
  - 如果 γ 稍微**小于** 1（k_a 稍微短一点），则 $e^{\gamma\beta} \ll e^{\beta}$，权重会几乎全部分配给 v_b。
- 对 c 方差的影响：c 的方差会变得非常大。

原本我们期望 c 是一个稳定的混合向量，但现在它变成了两个极端状态（v_a 或 v_b）之间的随机切换。这说明Single-headed Attention在处理这种特定方向的噪声时非常脆弱，这也是为什么 Transformer 需要 Multi-head Attention的原因之一（多头可以通过平均化不同头的输出来抵消这种方差）。

1.(d) Benefits of multi-headed attention

i.

取 $q_1 = \beta\mu_a$， $q_2 = \beta\mu_b$

所以，$c_1 \approx v_a$， $c_2 \approx v_b$， $c = \frac{1}{2}(c_1 + c_2) \approx \frac{1}{2}(v_a + v_b)$

ii.

定性描述：

与单头注意力（1c）中输出剧烈波动的情况完全不同，在这里，输出 c 会非常稳定，在所有样本中都始终保持在理想值 $\frac{1}{2}(v_a + v_b)$ 附近。

解释：

- 对于 Head 1 ($q_1 = \beta\mu_a$)：
  - 它的任务是找出 k_a。虽然 k_a 的模长波动很大（Variance 大），但 q_1 只会将 k_a 与其他无关的键 $k_{i\neq a}$ 进行比较。
  - k_a 的得分是 $\beta \cdot \gamma$（γ 在 1 附近波动），而其他 k_i 的得分接近 0。即使 k_a 的得分在波动（比如从很大变成非常大），它依然**远大于**其他无关键的得分（0）。Softmax 只要看到一个显著的最大值，就会输出接近 1 的概率。因此，Head 1 始终稳定地锁定 v_a，不会发生目标切换。c_1 的方差很小。
- 对于 Head 2 ($q_2 = \beta\mu_b$)：
  - 它关注的是 k_b。由于 k_b 的方差很小（且 k_a 的波动方向与 q_2 正交，互不影响），Head 2 始终稳定地锁定 v_b。c_2 的方差也很小。
- 由于 c_1 稳定等于 v_a，且 c_2 稳定等于 v_b，它们的平均值 c 自然也具有极低的方差。

1.(e)

通过将复杂的混合任务分解为简单的单一任务（每个头只关注一个目标），模型避免了在不同目标之间摇摆不定（即 1.(c) 中的竞争效应），从而极大地提高了抗噪能力。

2.(a) Permuting the input

i.

1. **Self-Attention 层的推导**

首先计算置换后的 Query, Key, Value 矩阵：

- $\mathbf{Q}_{\text{perm}} = \mathbf{X}_{\text{perm}}\mathbf{W}_Q = (\mathbf{PX})\mathbf{W}_Q = \mathbf{P}(\mathbf{X}\mathbf{W}_Q) = \mathbf{PQ}$
- 同理，$\mathbf{K}_{\text{perm}} = \mathbf{PK}$
- 同理，$\mathbf{V}_{\text{perm}} = \mathbf{PV}$

接下来计算 Attention 输出 $\mathbf{H}_{\text{perm}}$：

$$\mathbf{H}_{\text{perm}} = \text{softmax}\left(\frac{\mathbf{Q}_{\text{perm}}\mathbf{K}_{\text{perm}}^\top}{\sqrt{d}}\right)\mathbf{V}_{\text{perm}}$$

代入上面的结果：

$$\mathbf{Q}_{\text{perm}}\mathbf{K}_{\text{perm}}^\top = (\mathbf{PQ})(\mathbf{PK})^\top = \mathbf{PQK}^\top\mathbf{P}^\top$$

利用题目给出的 Softmax 性质 $\text{softmax}(\mathbf{PAP}^\top) = \mathbf{P}\text{softmax}(\mathbf{A})\mathbf{P}^\top$：

$$\text{softmax}\left(\frac{\mathbf{PQK}^\top\mathbf{P}^\top}{\sqrt{d}}\right) = \mathbf{P}\text{softmax}\left(\frac{\mathbf{QK}^\top}{\sqrt{d}}\right)\mathbf{P}^\top$$

现在代入 $\mathbf{V}_{\text{perm}} = \mathbf{PV}$ 并计算 $\mathbf{H}_{\text{perm}}$。注意置换矩阵是正交的，即 $\mathbf{P}^\top\mathbf{P} = \mathbf{I}$：

$$\begin{aligned}\mathbf{H}_{\text{perm}} &= \left(\mathbf{P}\text{softmax}\left(\frac{\mathbf{QK}^\top}{\sqrt{d}}\right)\mathbf{P}^\top\right)(\mathbf{PV}) \\ &= \mathbf{P}\text{softmax}\left(\frac{\mathbf{QK}^\top}{\sqrt{d}}\right)(\mathbf{P}^\top\mathbf{P})\mathbf{V} \\ &= \mathbf{P}\text{softmax}\left(\frac{\mathbf{QK}^\top}{\sqrt{d}}\right)\mathbf{IV} \\ &= \mathbf{PH}\end{aligned}$$

2. **Feed-Forward 层的推导**

定义前馈层的公式为：$\mathbf{Z} = \text{ReLU}(\mathbf{HW}_1 + \mathbf{1}\mathbf{b}_1)\mathbf{W}_2 + \mathbf{1}\mathbf{b}_2$。

计算 $\mathbf{Z}_{\text{perm}}$：

$$\mathbf{Z}_{\text{perm}} = \text{ReLU}(\mathbf{H}_{\text{perm}}\mathbf{W}_1 + \mathbf{1}\mathbf{b}_1)\mathbf{W}_2 + \mathbf{1}\mathbf{b}_2$$

代入 $\mathbf{H}_{\text{perm}} = \mathbf{PH}$，且注意到 $\mathbf{1} = \mathbf{P1}$：

$$\mathbf{H}_{\text{perm}}\mathbf{W}_1 + \mathbf{1}\mathbf{b}_1 = \mathbf{PHW}_1 + \mathbf{P1}\mathbf{b}_1 = \mathbf{P}(\mathbf{HW}_1 + \mathbf{1}\mathbf{b}_1)$$

利用题目给出的 ReLU 性质 $\text{ReLU}(\mathbf{PA}) = \mathbf{P}\text{ReLU}(\mathbf{A})$：

$$\text{ReLU}(\mathbf{P}(\mathbf{HW}_1 + \mathbf{1}\mathbf{b}_1)) = \mathbf{P}\text{ReLU}(\mathbf{HW}_1 + \mathbf{1}\mathbf{b}_1)$$

最后计算 $\mathbf{Z}_{\text{perm}}$：

$$\begin{aligned}\mathbf{Z}_{\text{perm}} &= \mathbf{P}\text{ReLU}(\mathbf{H}\mathbf{W}_1 + \mathbf{1}\mathbf{b}_1)\mathbf{W}_2 + \mathbf{P}\mathbf{1}\mathbf{b}_2 \\ &= \mathbf{P}\left(\text{ReLU}(\mathbf{H}\mathbf{W}_1 + \mathbf{1}\mathbf{b}_1)\mathbf{W}_2 + \mathbf{1}\mathbf{b}_2\right) \\ &= \mathbf{P}\mathbf{Z}\end{aligned}$$

ii.

这一性质说明 Transformer 模型（如果不加位置编码）是顺序不敏感的（Order-invariant）。如果只是简单地打乱输入 $\mathbf{X}$，模型在 $\mathbf{Z}$ 中为每个token生成的表示 $\mathbf{Z_i}$ 也会仅仅是相互间排列顺序变了，数值完全一样，但是在真实文本中，同一个单词在语句中不同位置肯定带着不同的含义。这意味着模型无法通过结构本身区分主语和宾语，也无法捕捉序列依赖关系。

这就是为什么 Transformer 必须引入位置编码（Positional Encodings）的原因，以打破这种置换等变性，让模型知道"谁在谁前面"。

2.(b) Position embeddings

这是 CS 224N 作业 4 第 2 题 (b) 部分的解答。这部分讨论了 **位置编码（Position Embeddings）** 的作用和数学性质。

以下是针对 (i) 和 (ii) 的详细解答：

i.

Yes.

ii.

No.

为了让两个不同位置 t_1 和 t_2 ($t_1 \neq t_2$) 的位置编码向量完全相同，两个编码向量中的每一个维度都必须相同。

只看最简单的维度，即当 $i = 0$ 时（分母为 $10000^0 = 1$）：

- 向量包含 $\sin(t)$ 和 $\cos(t)$ 这两个分量。
- 如果 $\Phi_{t_1} = \Phi_{t_2}$，则必须同时满足：

  $$\sin(t_1) = \sin(t_2) \quad \text{和} \quad \cos(t_1) = \cos(t_2)$$
- 这一条件成立的前提是 t_1 和 t_2 必须相差 2π 的整数倍（即 $t_1 - t_2 = 2k\pi$）。**然而**，题目中定义的位置 t 是**整数** $(0, 1, \ldots, T-1)$。由于 π 是无理数，两个整数之差永远不可能等于 2π 的非零整数倍。因此，对于任何不同的整数位置 t，其位置编码向量 Φ 都是**唯一**的。