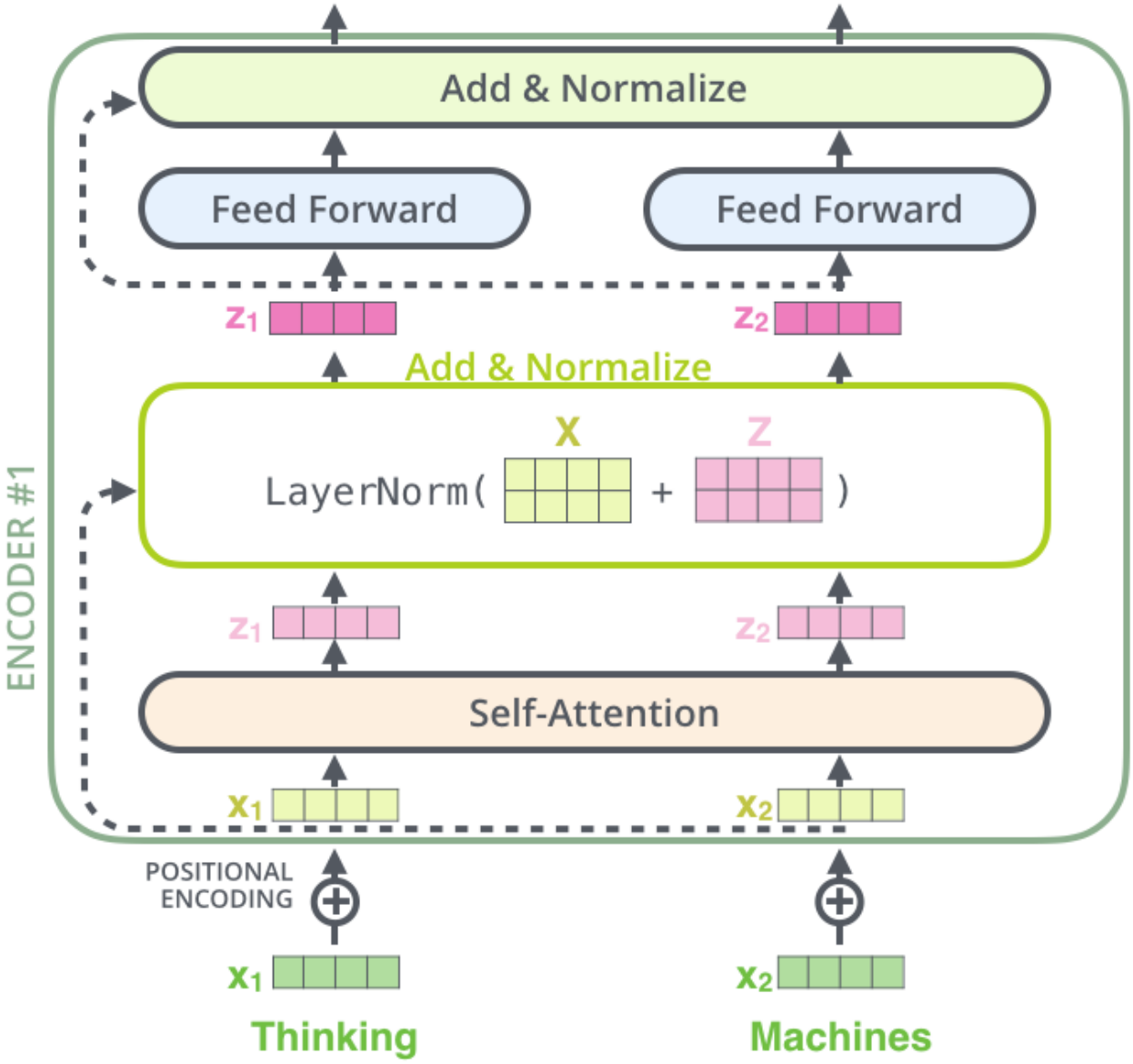
PE - позиционные эмбединги.
Архитектура трансфомера - это архитектура созданная из енкоднра и декодера, только их какое то H количество штук. Возьмем классическую схему из хабра,
И на её примере посмотрим работу трансформера. У нас на схеме вводится x1 и x2 вектора запросов, далее для каждого из них создается позиционный эмбеддинг. Что это такое? Позиционный эмбединг - вектора, который хранит положение слова в предложении, он необходим, так как без него будет проблема о незнании положения слова в предложении. $$\epsilon_{ij} = \frac{x_i W^Q (x_j W^K)^T + x_i W^Q (p_{ij}^K)^T}{\sqrt{d}}$$

ПЕ разделяются в MXCA на два типа:
1)Абсолютный
2)Относительный
В абсолютном каждый входящий токен на позиции i сопоставляется с тренировочным ебмединг вектором, который показывает строку в матрице R с размером [tokens, dims]. R - тренировочная матрица, инициализированная значениями (0,1)

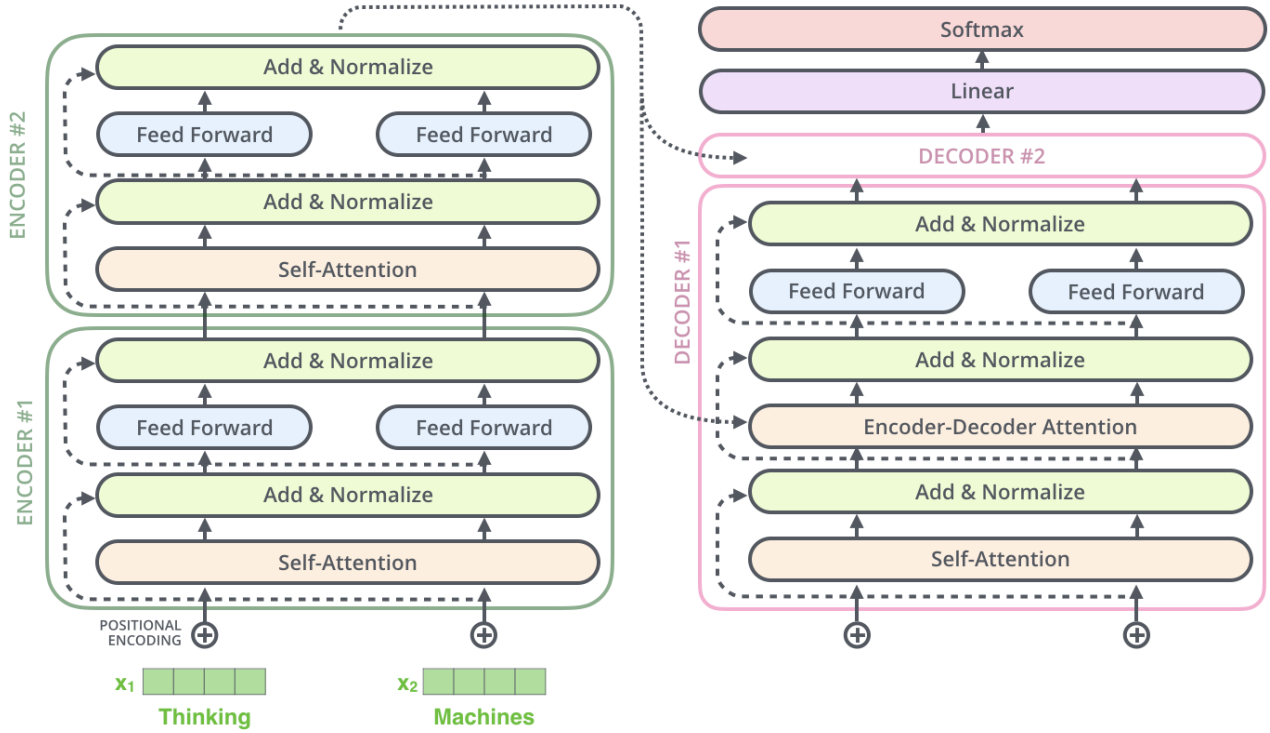
$$att = softmax(\frac{1}{\sqrt{dim}}(QK^T + QR))$$

Относительный ПЕ пресдавляет собой дистанцию между токенами. Но в них могут оказаться трудности с размерностями матрицами, там может быть маленькое несоответствие с матрицами. $$\epsilon_{ij} = \frac{x_i W^Q (x_j W^K)^T + x_i W^Q (p_{ij}^K)^T}{\sqrt{d}}$$.
Обычно используют функции косинуса и синуса в функции ПЕ, почему так углубляться не будем.



Получается мы рассматриваем енкодер, в енкодер вводятся векторы x1 и x2, далее они проганяются через слой внутренненого внимания добавляются к прошлым векторам и нормализуются, далее они проходят через обычную сеть прямого распространения и также суммируются и нормализуются.

Декодер



Энодер и декодер были расмотрены ранее, но их полносвязная работа в трансфомерах представлена на этой схеме.
Линейный слой – это простая полносвязная нейронная сеть, которая переводит вектор, созданный стеком декодеров, в значительно больший вектор, называемый логит вектором (logits vector).

Пусть наша модель знает 10 тысяч уникальных английский слов («выходной словарь» нашей модели), которые она узнала из обучающего корпуса. Это означает, что наш логит вектор будет иметь 10 000 ячеек в ширину – каждая ячейка соответствует коэффициенту одного уникального слова. Таким образом мы интерпретируем выход нашей модели с помощью линейного слоя.
Слой софтмакс переводит этот показатель в вероятности (положительные числа, сумма которых равна 1). Выбирается ячейка с наиболее высокой вероятностью и на выход данного временного отрезка подается соответствующее слово.