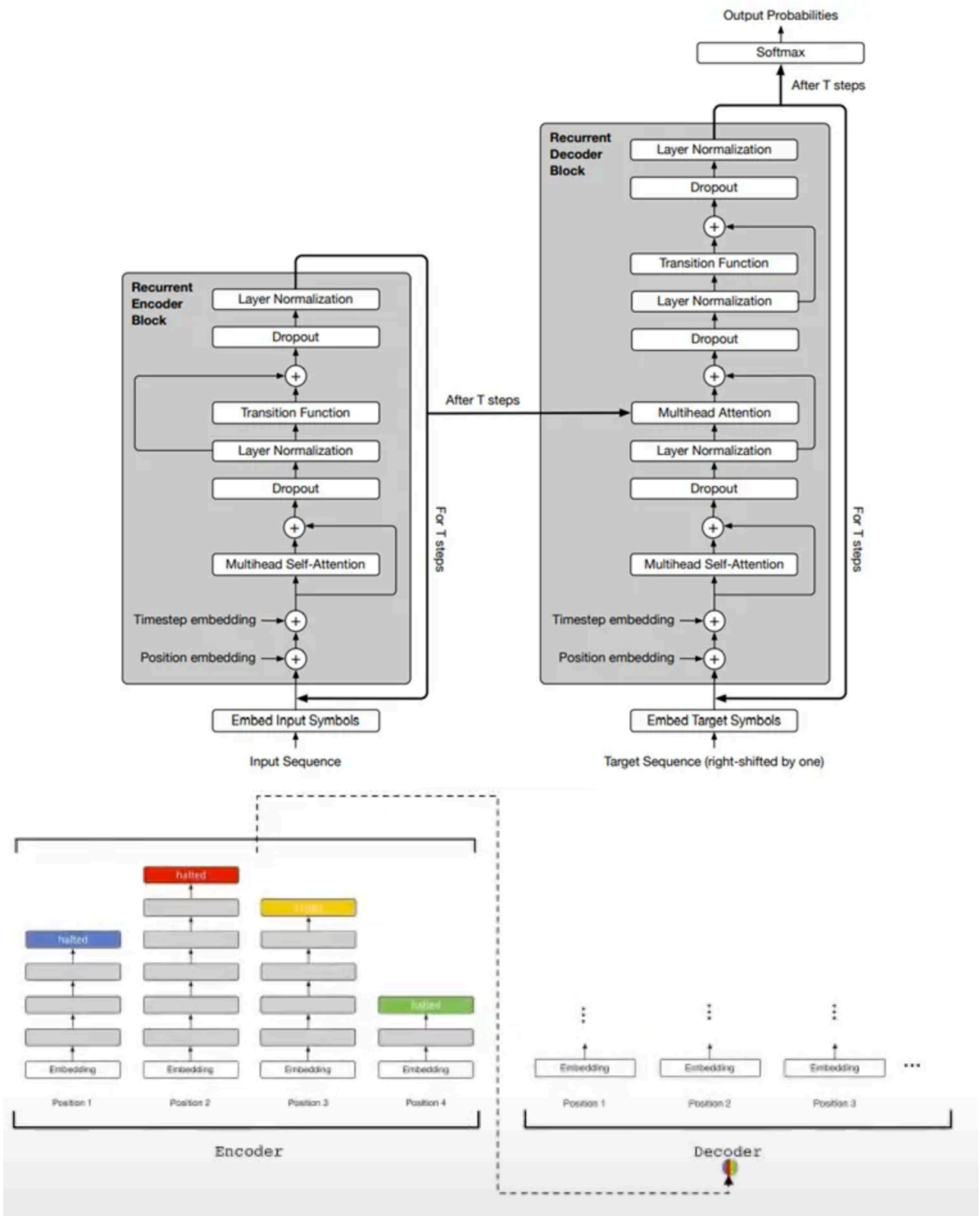
1. Universal Transformer

叫Universal Transformer是因為只要給他足夠的GPU，他就能做很多很多事情。Universal Transformer架構跟原本的Transformer差不多，額外的特點有

1. 將同一組weight應用在不同層上，以做出他想要的embedding。
2. 使用Dynamic Halting，也就是讓不同的字使用不同的深度，深度多深由model自己預測的機率決定要不要停下來。

附圖是Universal Transformer與Dynamic Halting的架構



2. Transformer的優點是translation表現好但algorithmic task表現很差，當時有的另外兩種models (Neural GPU and Neural Turing Machine)則是在algorithmic task表現好但在translation表現很差，Universal Transformer則同時在兩種面向都表現優秀，更加泛用。

Reference: 助教課
https://medium.com/@pocheng0118/a-brief-overview-vanilla-transformer-v-s-universal-transformer-無所不能的通用計算模型-d9e89dcef080