

# 模型比較報告書

## I. 定義名詞

刪減版模型：使用flash attention代替原本attention，並且使用混合精度float32和float16交替。

未刪減模型：原始attention，沒有節省記憶體，採用小batch訓練。

兩者的所有的參數全部一樣，並且採用相同的交易策略，方便比較兩個模型(包含模型參數、optimizer參數、config)，並且使用AdamW代替SGD，因為transformer架構訓練時較不穩定，因此不用SGD。

這裡使用alphalens的方式是，把模型在T0預測T1買入T4賣出的標準化報酬率，當作是因子，拿來與實際T1買入T4賣出的真實報酬率去做分析，劃分成10個Quantile。

## II. IC值比較

從Fig. 1. 可以看出，因子預測力，原始attention表現比flash attention好，兩者標準差幾乎相同，Risk Adjusted IC也是原始attention表現較好。p-value表示，兩個模型的因子預測力，不是偶然的，而是真正有效。

	Metric	刪減版模型	未刪減模型
0	IC Mean	0.047	0.066
1	IC Std.	0.104	0.110
2	Risk Adjusted IC	0.457	0.598
3	t-stat(IC)	4.749	6.242
4	p-value(IC)	0.000	0.000
5	IC skew	0.131	-0.064
6	IC kurtosis	-0.227	-0.238

Fig. 1. IC值表格比較圖

## III. Alpha以及Beta比較

從Fig. 2.來看，對於alpha表現，原始attention表現比flash attention好，兩者的beta則是差不多，都與市場整體輕微負相關，再區分Quantile上，不論是Top Quantile Return還是Bottom Quantile Return，都是原始attention表現較好，最後的Spread(bps)也必然是原始attention表現出色。

	Metric	刪減版模型	未刪減模型
0	Ann. alpha	0.036	0.078
1	beta	-0.137	-0.132
2	Top Quantile Return (bps)	10.818	27.984
3	Bottom Quantile Return (bps)	-27.330	-39.621
4	Spread (bps)	38.148	67.605

Fig. 2. Alpha和Beta以及Quantile比較圖

## IV. Turnover比較

從Fig. 3.來看，Q1~Q10的Turnover都是原始attention較低，但是flash attention的Mean Factor Rank Autocorrelation卻比原始attention好。

指標	刪減版模型	未刪減模型
Quantile 1.0 Mean Turnover	0.432	0.341
Quantile 2.0 Mean Turnover	0.712	0.629
Quantile 3.0 Mean Turnover	0.779	0.713
Quantile 4.0 Mean Turnover	0.805	0.763
Quantile 5.0 Mean Turnover	0.812	0.779
Quantile 6.0 Mean Turnover	0.810	0.775
Quantile 7.0 Mean Turnover	0.797	0.742
Quantile 8.0 Mean Turnover	0.770	0.698
Quantile 9.0 Mean Turnover	0.699	0.628
Quantile 10.0 Mean Turnover	0.420	0.403
Mean Factor Rank Autocorrelation	0.791	0.743

Fig. 3. Turnover比較

## V. 交易策略報酬結果(單利計算)

這裡都是採用最簡單的交易策略，T0決定買入的股票，持有期間T1~T4，T4賣出獲利。

從Fig. 4. 和 Fig. 5.來看，原本Attention模型比起Flash Attention模型，整體報酬以及報酬穩定度來說，都是比較好的，而且抗跌的能力也較佳。

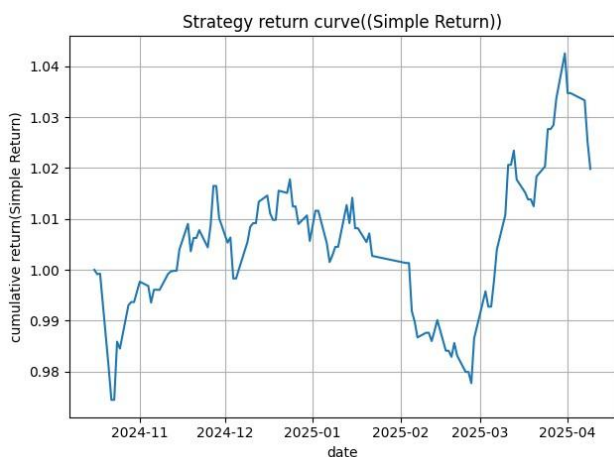


Fig. 4. Flash Attention模型



Fig. 5. 原本Attention模型

## VI. 比較報酬Quantile

以下的Quantile是計算根據算出的分數排名，劃分成Q1~Q10去買入，當日報酬是採計T1買入T4賣出的報酬率(相當於時間是5倍速流逝，這種方式只是單純看預測五日後的預測是否準確，以及模型是否能區分好壞股票，所以不是實際交易的報酬)，而且這裡採計純粹以採計T1買入T4賣出的報酬率相加，所以會出現負值。

從Fig. 6. 和 Fig. 7.可以看出，Flash Attention模型的Q1~Q10的Quantile曲線中途時常混雜在一起，然而原本Attention模型Q1~Q10的Quantile曲線分得很開，代表原本Attention模型技能區分好壞股票。

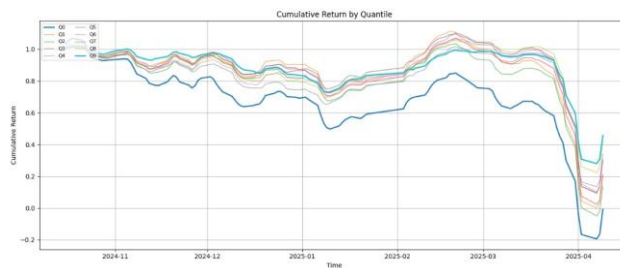


Fig. 6. Flash Attention模型 (Cumulative Quantile)

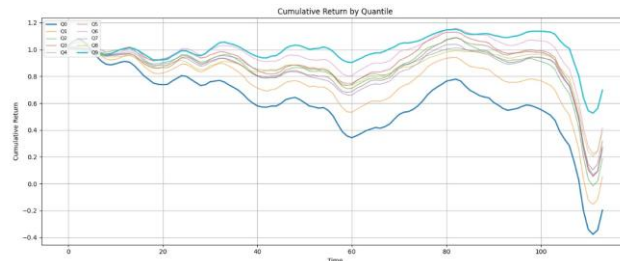


Fig. 7. 原本Attention模型 (Cumulative Quantile)

## VII. 訓練難易度

不論是用MSE還是MAE，Flash Attention模型需要額外調參數，將爆炸的梯度補救回來，而且方式會是case by case，每訓練一次，必須再調整一次，而原本Attention模型則可以一次訓練到底。並且訓練過程中，收斂速度較快較穩定，也較能重現結果，Flash Attention模型出來的結果時常不能複現。

## VIII. 數值收斂

模型預測的是標準化後報酬率，因此valid loss在不同的loss function底下有不同的標準。

MSE : valid loss < 1 才算是好模型

因為標準化後報酬率滿足mean=0, std=1，最沒有預測力的模型，什麼都不做，只輸出mean，意思是指輸出0的話。以常態分佈來說：

$$\mu = 0, \quad \sigma = 1$$

$$Z \sim \mathcal{N}(\mu, \sigma^2)$$

$$\mathbb{E}[(Z - \mu)^2] = \sigma^2 = 1$$

所以在MSE底下，valid loss < 1 才算是好。

MAE : valid loss < 0.797 才算是好模型

以常態分佈來說：

$$\mu = 0, \quad \sigma = 1$$

$$Z \sim \mathcal{N}(\mu, \sigma^2) = \mathcal{N}(0, 1)$$

$$\mathbb{E}[|Z - \mu|] = \sigma \cdot \sqrt{\frac{2}{\pi}} \approx 0.797$$

所以在MAE底下，valid loss < 0.797 才算是好。

然而，flash attention在訓練時，MSE常常在1.4以上，調很久參數才有辦法到1.02，但是原始attention非常容易就可以到0.9。

對於MAE，flash attention常常在0.9以上，而原始attention則可以到0.67以下，比0.797這個門檻還小。

## IX. 結果

使用原始attention比flash attention效果還要好，而且還更好訓練，並且收斂穩定。

