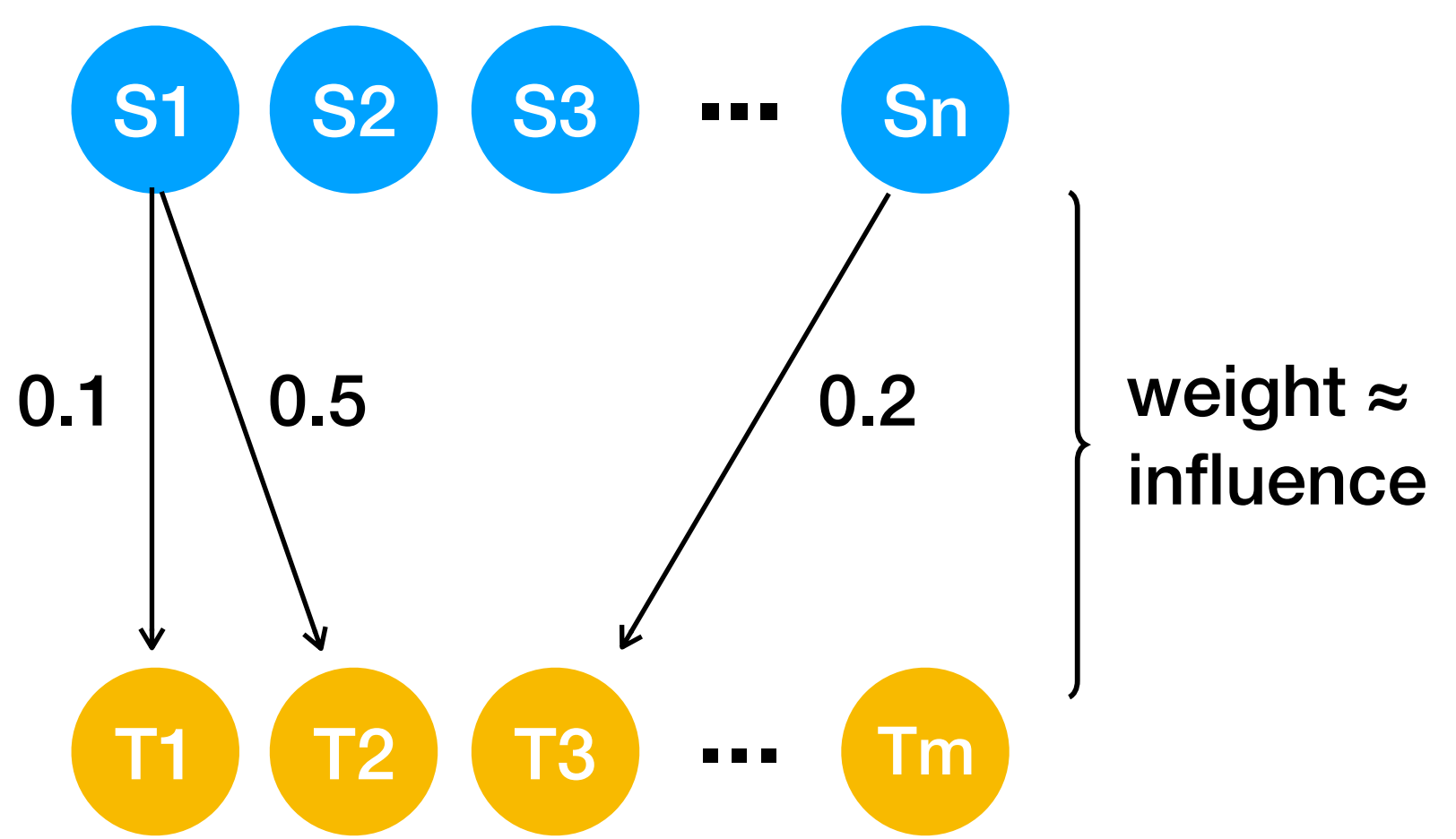
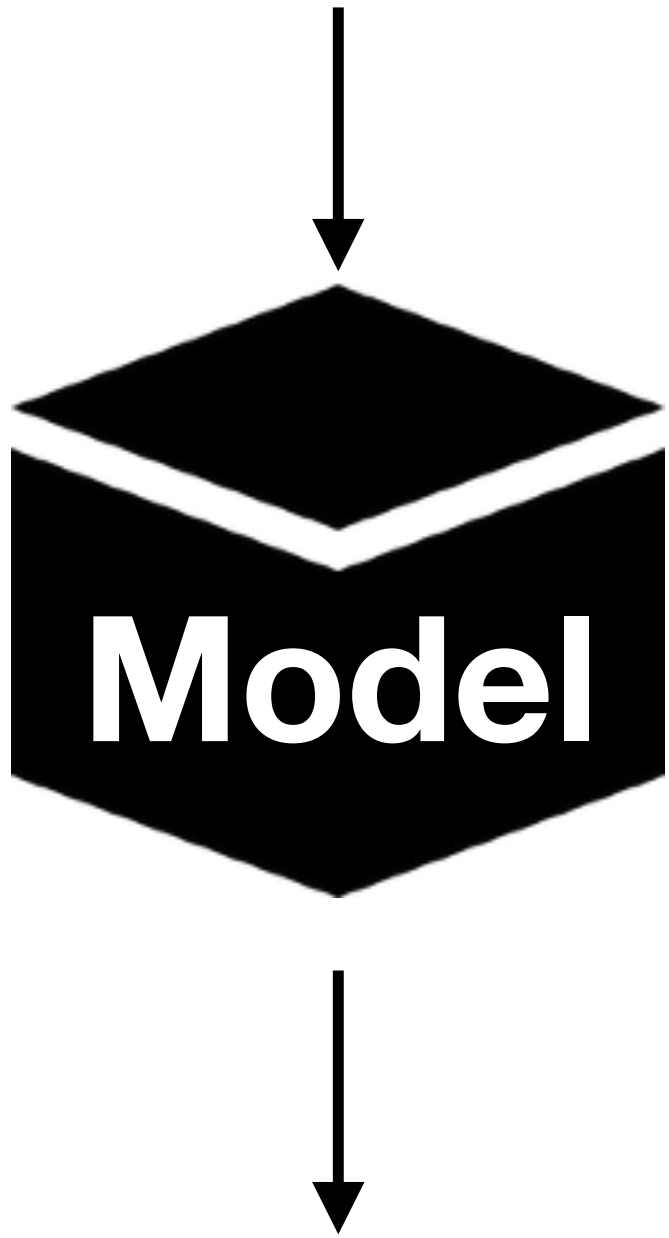
Interpretability
for Black-Box Seq2Seq Models

[redacted]

AM+ Jaakkola, *EMNLP'17*


S1
S2
S3
...
Sn
0.1
0.5
0.2
weight ≈
influence
T1
T2
T3
...
Tm

Input: (S1,S2,...,Sn)



Output: (T1,T2,...,Tn)

(locally)

- Weighted bipartite graph summarizes model's local behavior:

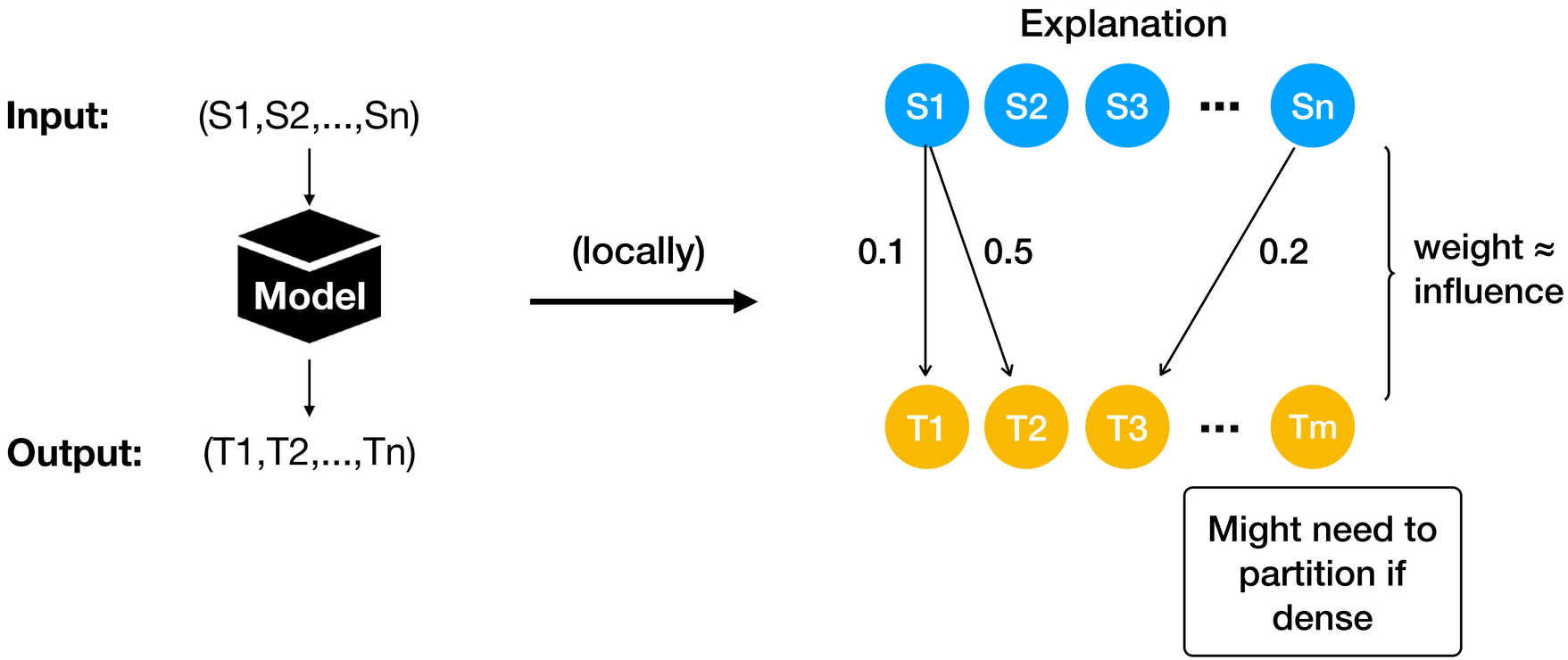
Explanation

Might need to partition if dense

Interpretability
for Black-Box Seq2Seq Models

AM+ Jaakkola, *EMNLP'17*

- Weighted bipartite graph summarizes model's local behavior:



Interpretability for Black-Box Seq2Seq Models

AM+ Jaakkola, *EMNLP'17*