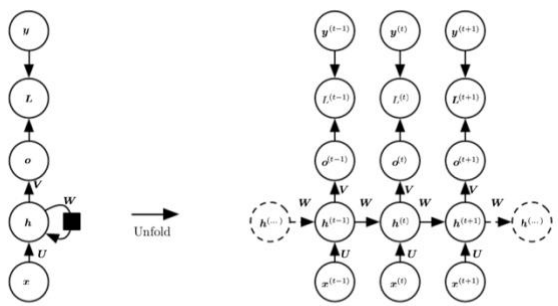
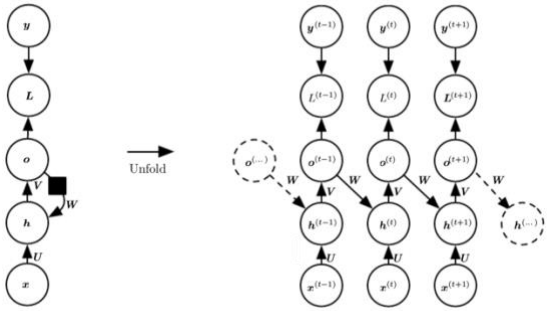
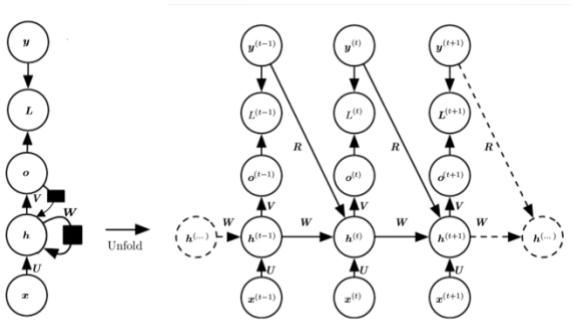
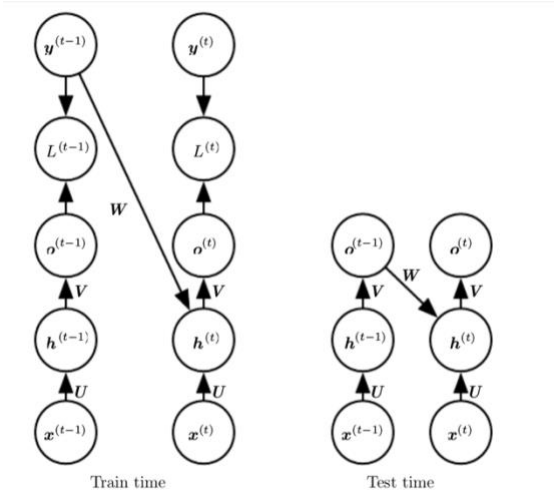
G1	G2	G3
		
Sequence-to-sequence same length: Length of both x- and y-sequences must be the same Output at each time step	Sequence-to-sequence same length: Length of both x- and y-sequences must be the same Output at each time step	Sequence-to-sequence same length: Length of both x- and y-sequences must be the same Output at each time step
Recurrent connections: • Recurrent connections between hidden units. • Optional Output-to-Hidden connection => Graph3 • Can choose to put any information it wants about the past into its hidden representation h and transmit h to the future	Recurrent connections: • Recurrence from prev. outputs to Hidden Units • Lacks H-H recurrent connections: ▪ no direct connections from **h** going forward ▪ The previous h is connected to the present only indirectly, via the predictions produced from it. • **o** is the only information it is allowed to send to the future (lack important information from the past, unless o is very high-dimensional and rich)	Recurrent connections: • Recurrent connections between hidden units *and* from previous outputs to hidden units (see G1 and G2) • Trained with BPTT and Teacher Forcing

G1	G2	G3
"Universal": can compute any function computable by a Turing Machine	Cannot simulate a universal Turing machine (expresses a smaller set of functions than G1)	Can compute any function computable by a Turing Machine Can model arbitrary distribution over sequences of y, given sequences of x
Expensive to train, not parallelizable • The forward propagation graph is inherently sequential; each time step may only be computed after the previous one → training cannot be executed in parallel • States computed in the forward pass must be stored until they are reused during the backward pass => high memory cost	Less expensive to train, training parallelizable Each time step can be trained in isolation → training can be executed in parallel	Not parallelizable (see G1)
Trainable with Teacher Forcing: No, when H-H only, Yes, with additional O-H (see graph3)	Trainable with Teacher Forcing: YES	Trainable with Teacher Forcing: YES

Teacher Forcing



- Teacher forcing is a training technique applicable to RNNs, that use model output from a prior time step as an input.

- Thought as an alternative to back-propagation through time (BPPT) for models without hidden-to-hidden connections.

- During Training:
  the target output from a prior time-step will be fed as input into the model.
  → each training step can be done in isolation
  → training parallelizable

- During Testing:
  there will be no target/ true output available.
  → an approximation of the correct output will be computed and fed into the hidden units.