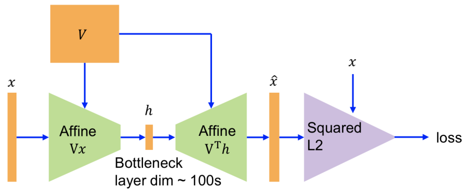
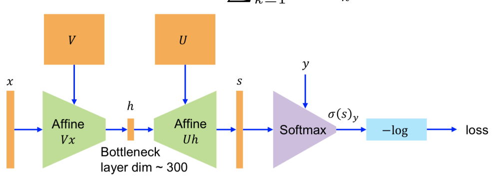
semantics **Propositional** Can attach prob; Allow logical interface; good in well-defined domain **Vector** bag-of-words (BoW) Paradigmatic Similarity→exchangable in context Embeding Use vec of context; Dimension Reduction: Latent Semantic Analysis, LSA Entire doc x: (N # doc x M # word)→V(KxM), $T = U^T V = U^T S V$ (SVD), auto-encoding, min L2



Word2Vec Local context j, y is local context words centered around word i; v: input embedding vec; u output; x: one-hot encoded input word i > **skip-grams** min cross entropy

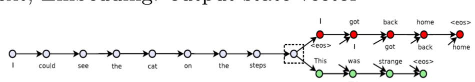
$$-\log p(j|i)\ p(j|i) = \frac{\exp(u_j^T v_i)}{\sum_{k=1}^{V} \exp(u_k^T v_i)}$$



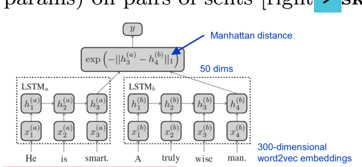
\+ relations,properties: women-man=queen-king - **expensive** to normalize softmax over all words→Approx,heuristic down-weighting of frequent words Co-occurrence Matrices C_{ij}: # doc contain both words i and j, full-doc→LSA, window-based→Word2Vec GloVe C_{ij}: # occurs of j around i, minimize: $J(\theta) = \sum_{i,j=1}^{V} f(C+ij)(u_i^T v_j + b_i + \hat{b_j} - \log C_{ij})^2, f(x) = (x/x_{\max})^\alpha$ if $x < x_{\max}$ else $1, x_{\max} = 100, \alpha = 3/4$

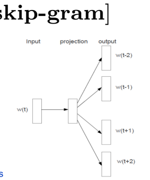
Lexical →Compositional, model text struct

Skip-Thought Vec Use seq2seq RNN to pred +1/-1 sent; can encode longer text; Embedding: output state vector



Siamese Models Train pair of networks (share params) on pairs of sents [right > **skip-gram**]

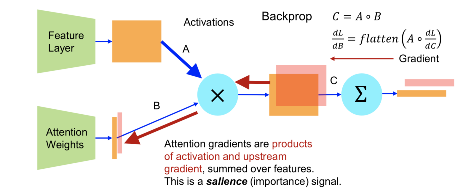




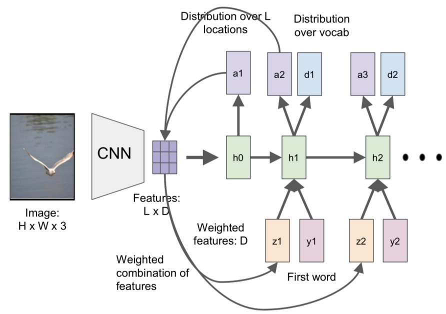
???p62/63 lec11 Train for your evaluation metric, Hidden Unit Factors 1,2, and 6

attetion + accuracy+, computation-, learn pred salience (emphasize relevant data across space or time); explanations

During training, the attention layers receives gradients which are the product of the upstream gradient and the feature layer activations (salience).

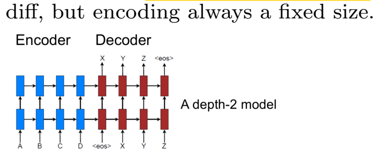


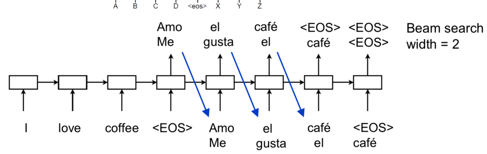
Hard vs Soft Hard Attend to a **single** input location; Cannot gradient descent; Need RL|compute a weighted comb over **some** input; Can backprop to train end2end RL vs Supervised L receive rewards from environment→not differentiable, max $\sum_t r_t$ Soft-Att Caption $z = p_a a + p_b b + ...$ Derivative dz/dp, train with GD ↔random sample p, z, dz/dp 0 almost everywhere, no GD



Att-n-LSTM i, f, o nodes receive a salience gradient, learn to weight features ???p69/73 lec 12, tend to fixed grid →pred params of mixture model; DRAW: Classify images by attending to arbitrary regions of input+Generate...Output; Transformer

Trans ?? where bi-linear interpolation Seq2Seq Encoder+Decoder Reverse the order of input sent←the head is most important and reversal eases the long-term dependencies from output to input sentence; Narrow Beam Search - Sent len diff, but encoding always a fixed size.

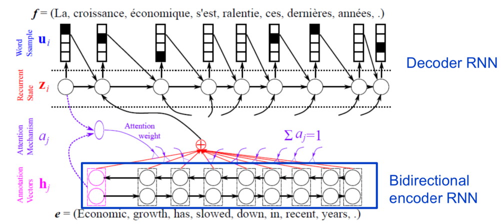




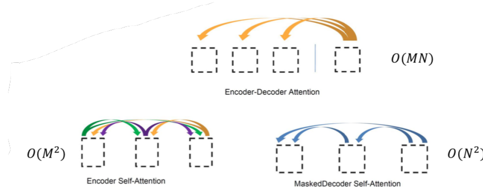
\> **BLEU** Weighted comb of n-gram precis,gram precis+sent len, $p_n = \frac{\text{max occurs in one reference}}{\text{occurs in cand}}$ $BLEU = BP \exp \sum w_n \log p_n$,

`BP=1 if c>r else exp(1-r/c)` Tends unigram→adequacy|n-gram→fluency

SoftAtt Trans Compare latent states of en/decoder (Bahdanau): Alignment scores $e_{ij} = a(s_{i-1}, h_j)$, Mixture weights $\alpha_{ij} = \frac{\exp(e_{ij})}{\sum \exp(e_{ik})}$, Context vector $c_i = \sum \alpha_{ij} h_j$

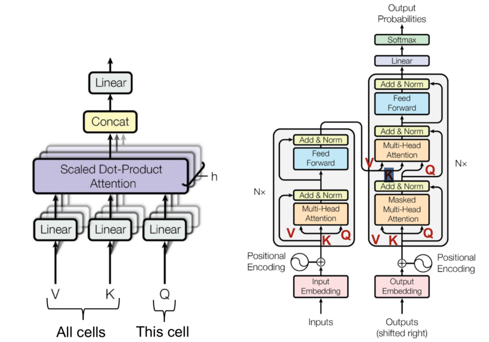


\- attention func ($a(s,h)$) complex, yet heatmap simple like word sim; att data path is another recurrent path between output states; cannot generalize to deeper nets→(Luong) Stacked LSTM with arbitrary depth:Global Att Model: Att layer sits above the en/decoder, not itself recurrent|Local Att Model: Comp best aligned position p_t first, then comp context vec centered at that pos Parse encode trees by closing parens RNN - Time in proportion to sent len; Long-range dependencies across many time steps; Tricky to learn hierarchical structures→ CNN



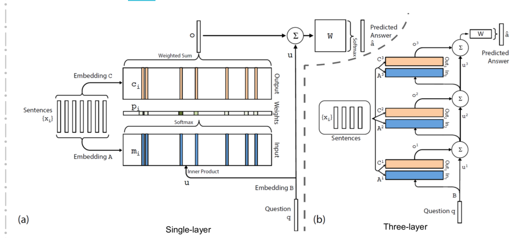
Self-Att Transformer + Can replace word-based recurrence entirely|Const path len between any two pos; Variable receptive field (the whole input seq)|Use depth to model hierarchical structure (Supports hierarchical information flow by stacking self-attention layers)|Trivial to parall|Attention weighting controls information propagation. Scaled Dot-Product Att → Multi-headed att Attention(Q,K,V) =softmax$(\frac{QK^T}{\sqrt{d_k}} mask)V$ Encoder layer: Q from prev decoder, K, V from encoder|Self-attention layer: Q, K, V all from prev decoder Att Fix Weighted average →Multiple attention layers: interpretation of inputs, heads in parallel so that each head uses different linear transformations (allow a per-input transformation, as convolution does) Positional encoding

Add loc back, Break symmetry so cells do diff, pos vec:=sinusoidal functions of position, period form a geometric series

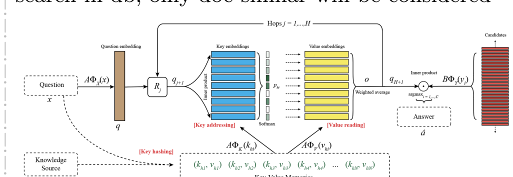


Summarization completely remove the encoder. M/N =in/output len BERT Transformered-based, Bi-dir Encoding Repres from ...: Pretrain and fine-tuning|bidirectional att|Two losses: delete a random word and pred, pred next sent from curr GPT and GPT2 Generative Pre-Training: simpler transformer-based model with more params. No fine-tuning, only adaptation of inputs.

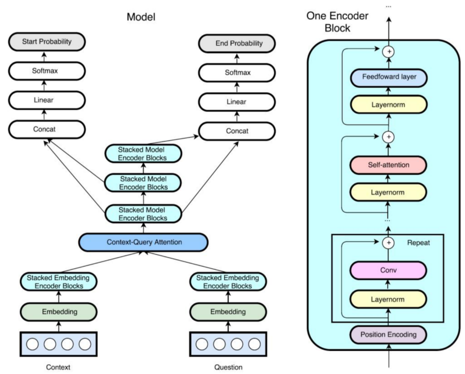
QA System Conv Activations fully predictable from inputs AttModel Agent has also dynamic attention MemNet Provide general purpose mem/pointers(via att), R/W, dynamic men for conversational agents MemNet framework (I)input2internal feature→(G)update mem→(O)produce new output→(R)convert into response Positinonal Encoding multiply input words by a linear function of position > **End2End,Multi-Hop**



KB knowledge based triple IE Information Extraction > **KeyValue** extend to Natural language text instead of structured text docs moral Use KB when possible ScaleUp full-text search in db, only doc similar will be considered



QANet ???



Dialog goal-directed tasks ssue API calls|Update API calls|Display options|Provide extra info|Conduct full dialogs ???

adversarial Traditional ML assume Training data similar to testing data Adversarial perturbations Physical Conditions (Angle,Dist,Light), Imperceptibility limitation,Fabrication/Perception Error(Color Reproduction),Background Modifications $\arg\min_\delta \lambda\|M_x\delta\| + \frac{1}{k}\sum J(f_\theta(x_i + M_x\delta), y^*) + NPS(M_x\delta)$, NPS: non-printability score optimize (Non-)Targeted Non-targeted adversarial is harder

1. LSA
2. Word2Vec
3. GloVe
x Skip-Thought