Contextual Code Completion

Subhasis Das, Chinmayee Shah

My code completion so dumb

```
int n, i;
for(i = 0; ...
```

What goes here?

Me

```
int n, i;
for(i = 0; i < ???; i++) {
```

My computer

```
int n, i;
for(i = 0; ???
```

Window Based Approach

Last **W** tokens to predict the next token

Dictionary = Key + Positional

- frequently occurring
- language key words

- functions
- variables

```
int   n    ,    i    ;
 k2   p2   k10   p4   k1

for   (    i    =    0    ; …
 k7   k5   p4   k21  p10   k1
```

```
static inline void lock_fat(struct msdos_sb_info *sbi)
{
        mutex_lock(&sbi->fat_lock);
}

static inline void unlock_fat(struct msdos_sb_info *sbi)
{
        mutex_unlock(&sbi->fat_lock);
}
```

mutex_unlock
if

```
framelen = le64_to_cpu(rxfd->rfs) & IPG_RFS_RXFRAM
if (framelen > sp->rxfrag_size)
```

framelen
rxfd

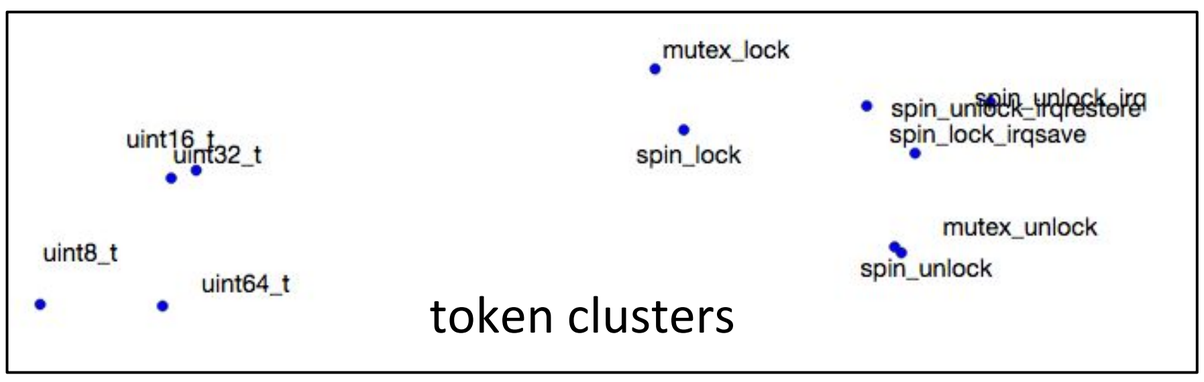
```
static const struct pci_device_id ipg_pci_tbl[] = {
        { PCI_VDEVICE(SUNDANCE, 0x1023), 0 },
        { PCI_VDEVICE(SUNDANCE, 0x2021), 1 },
        { PCI_VDEVICE(DLINK,    0x9021), 2 },
        { PCI_VDEVICE(DLINK,    0x4020), 3 },
        { 0, }
};

MODULE_DEVICE_TABLE(pci, ipg_pci_tbl);
```

Modeling Tokens

word2vec: token → dense vector

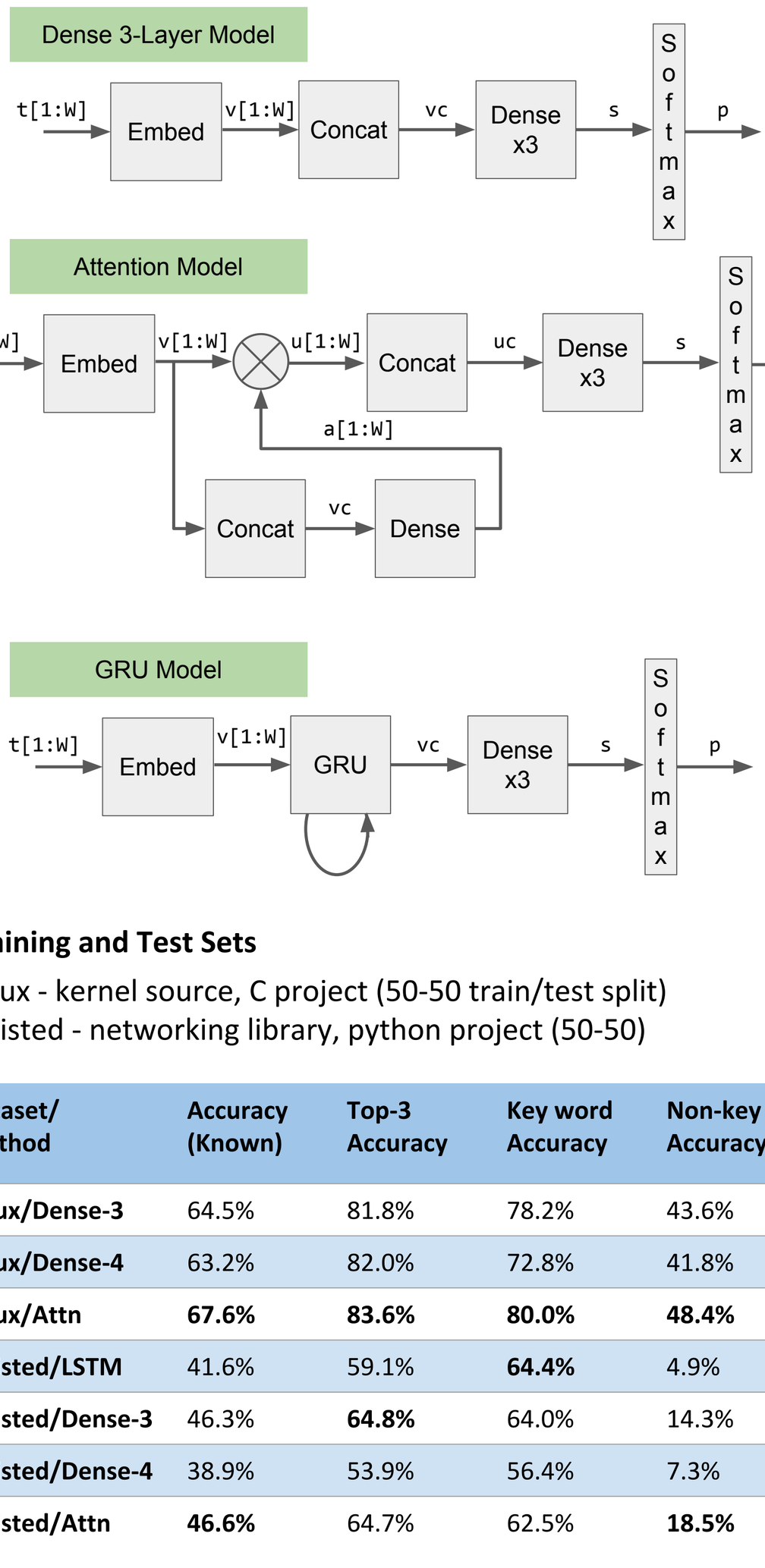
k1	{ 0.3, 0.1, -0.5, …}
p1	{-0.8, 0.9, 0.4, …}

Total # tokens = # keywords + # positions



token clusters

Learning Approaches and Results



Training and Test Sets

Linux - kernel source, C project (50-50 train/test split)
Twisted - networking library, python project (50-50)

Dataset/ Method	Accuracy (Known)	Top-3 Accuracy	Key word Accuracy	Non-key Accuracy
Linux/Dense-3	64.5%	81.8%	78.2%	43.6%
Linux/Dense-4	63.2%	82.0%	72.8%	41.8%
Linux/Attn	**67.6%**	**83.6%**	**80.0%**	**48.4%**
Twisted/LSTM	41.6%	59.1%	**64.4%**	4.9%
Twisted/Dense-3	46.3%	**64.8%**	64.0%	14.3%
Twisted/Dense-4	38.9%	53.9%	56.4%	7.3%
Twisted/Attn	**46.6%**	64.7%	62.5%	**18.5%**